Optik 125 (2014) 569–574

Contents lists available at ScienceDirect

Optik

journal homepage: www.elsevier.de/ijleo

Visual saliency detection via multiple background estimation and spatial distribution

Shuhan Chen*, Weiren Shi, Wenjie Zhang

College of Automation, Chongqing University, Chongqing, China

ARTICLE INFO

Article history:
Received 2 March 2013
Accepted 3 July 2013

Keywords:
Saliency Detection
Background estimation
Multiple background maps
Spatial distribution

ABSTRACT

Generating saliency maps with full resolution or pixel-level in linear computation time and low probability of falsely marking background as salient regions are still challenging in visual saliency. Concentrating on it, we propose a simple and efficient saliency detection approach. In detail, this is mainly contributed by estimating multiple robust background maps for each image. Once background maps are generated, salient objects can be easily obtained only by measuring the differences between the original images and their corresponding background maps, which is similar with background subtraction in video applications. To further improve performance, spatial distribution is incorporated as a high-level factor. Evaluation on public dataset indicates the proposed scheme achieves superior results in both accuracy and efficiency, which suggest the potential application such as image segmentation in large image collections.

1. Introduction

Visual saliency has become a very active topic in computer vision and attracted more and more scholars researching on it due to its helpfulness to image segmentation [1], object detection [2], image retargeting [3], object recognition [2]. Saliency detection aims to extract salient regions or objects in an image. However, various applications lead to different requirements for the extracted saliency maps. But in general, it is agreed that for good saliency detection, it should meet the following criteria: (1) Good detection. There should be a low probability of failing to mark real salient regions, and low probability of falsely marking background as salient regions. It corresponds to uniformly highlighted whole salient regions in the previous literatures [4,5]. (2) High resolution. Saliency maps should have high or full resolution to accurately locate salient objects and retain original image information as much as possible. (3) Computational efficiency. In order to work in large image collections and facilitate efficient subsequent processing, saliency maps should be fast and easy to generate [5] (Fig. 1).

Most existing visual saliency approaches are contrast-based due to the observation that human cortical cells have high response on contrast stimulus in receptive fields [7]. These contrast-based methods measure saliency though investigating the rarity of image regions with respect to its surroundings. In [9], Itti et al. proposed a saliency model by computing feature maps for luminance, color and orientation and using center-surrounding operator across multiscales. Inspired by Itti's model, Ma and Zhang [10] and Frintrop et al. [6] use center-surrounded feature distances to estimate saliency, while Frintrop et al. computed it with square filters and further use integral images to speed up the calculations. Different from the above models, Hou and Zhang [12] proposed the spectral residual approach processed in the frequency domain. Liu et al. [19] located the salient object by learning a conditional random field to combine multi-sale contrast, center-surround histogram, and color spatial distribution. By contrasting with both local and global surroundings and combing high-level features, Goferman et al. [3] proposed a context-aware saliency detection method which can highlight salient objects along with their contexts. However, such methods usually produce higher saliency values near edges and cannot uniformly highlight the whole salient objects which do not meet our first performance criterion.

To overcome the above shortcomings, Zhai and Shah [13] proposed a pixel-level saliency detection method by comparing each pixel to all others in the image, but color information is ignored for efficiency. Achanta et al. [4] thought it is caused by failing to exploit all the spatial frequency content of the original image and then proposed a simple yet efficient frequency tuned method by measuring color differences from the average image color. Although generated full resolution saliency maps, it does not work well when the background is complex or the salient objects comprise more than half the pixels of the image. In [14], Achanta further improved it by varying the bandwidth of the center surround-filtering. Recently, Cheng et al. [5] proposed global contrast-based approaches including a histogram-based contrast (HC) method and a region-based (RC) method. However, HC method generates full resolution saliency

* Corresponding author. Tel.: +86 023 65112760; fax: +86 023 65112760.
E-mail addresses: c.shuhan@gmail.com (S. Chen), wrs@cqu.edu.cn (W. Shi), daaiyiyejian@cqu.edu.cn (W. Zhang).

0030-4026/$ – see front matter © 2013 Elsevier GmbH. All rights reserved.
http://dx.doi.org/10.1016/j.ijleo.2013.07.014

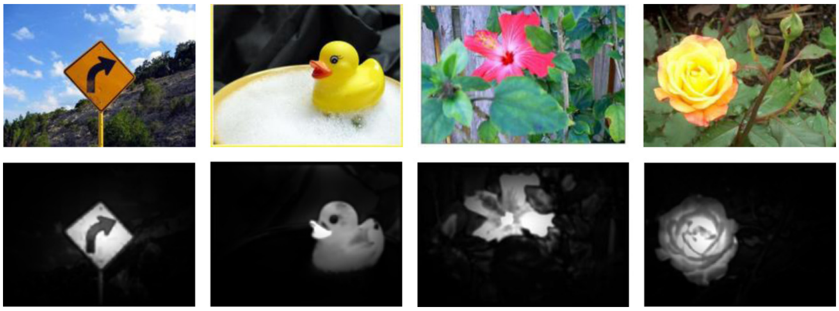

Fig. 1. Original images (top) and their saliency maps using our algorithm (bottom).

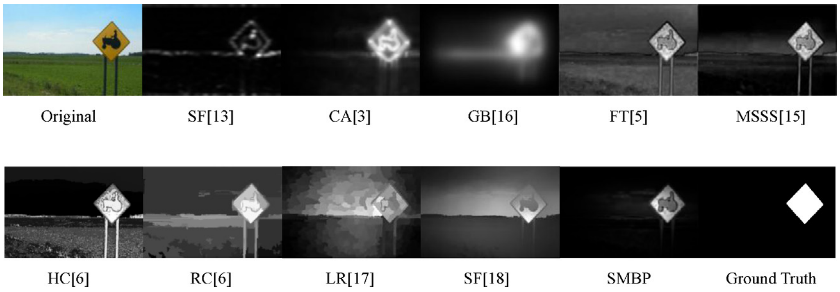

Fig. 2. Saliency maps computed by different state-of-the-art methods with our proposed method (SMBP). Most results highlight edges, falsely mark background as salient regions, or are of low resolution.

map efficiently but cannot perform well in textured scenes, and RC-maps achieve better results than HC-maps by incorporating spatial relations but are not full resolution and efficient.

More recently, there are also some novel and encouraging approaches proposed. By representing an image as a low-rank matrix plus sparse noises in a learned feature space, Shen and Wu [16] proposed a unified model to incorporate low-level features with higher-level guidance for saliency detection. Perazzi et al. [17] proposed a contrast-based method combining global contrast and spatial relations together to detect salient objects, which can generate full resolution saliency maps and simultaneously be implemented efficiently with linear complexity with the help of permutohedral lattice embedding [18]. Although gets satisfactory results in most images, they still falsely marks background as salient regions in some cases, which can be seen in Fig. 2.

Aim to generate saliency maps with full resolution in linear computation time and low probability of falsely marking background as salient regions, a multiple background maps based saliency detection approach is proposed in this paper. Furthermore, spatial distribution is also incorporated as a high-level factor to improve performance. It is noted that it can be simply combined into previous methods to improve both. Experiments on publicly available dataset show our approach outperforms the state-of-art works.

2. The proposed method

As we know, background subtraction is a simple and efficient method for foreground detection in video applications [11]. In this work, we tackle visual saliency as a foreground detection problem. Some similar works based on this idea have been explored in literatures [4,14]. In such works, mean value is computed as background, where such averaging operation is implemented in whole image or local region. However, such single value based background map is not robust because background regions are always complex in natural scenes. In such case, background regions will be falsely marked as salient regions as can be seen from Fig. 2. Thus, multiple background maps are necessary. Once background maps are generated, candidate saliency maps can be easily produced only by measuring the differences (Euclidean distance metric in CIELAB space is used in this paper) between the background maps and the original images. Finally, combination operation is needed to fuse these candidate saliency maps. Furthermore, we also incorporate spatial distribution to suppress background simultaneously highlight salient objects. The workflow of our method is shown in Fig. 3. Detailed discussion will be presented in the following subsections.

2.1. Multiple background estimation

Natural scenes usually include large dissimilar parts that are homogeneous by themselves, such as sky, lawn, or ground. Take Fig. 4 for example, most part of the top region in the image is sky, while the bottom is lawn. Thus, the bottom part of the image is

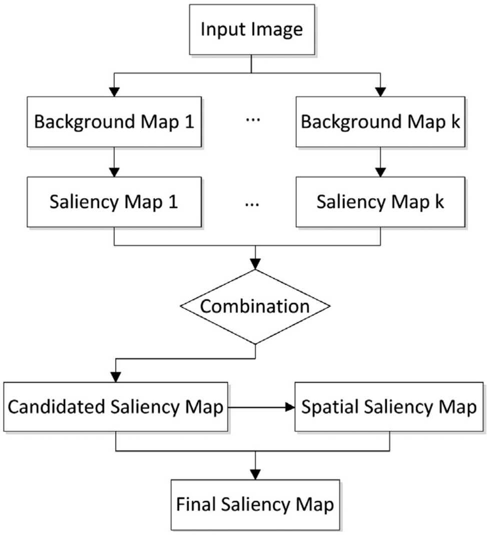

Fig. 3. Workflow of our method.

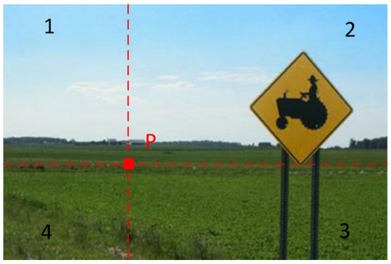

Fig. 4. Illustration of our method.

more suitable to estimate background for pixel P. Therefore, proper selection of local region is essential for background estimation. In this paper, we propose a proper local region based background estimation scheme.

As illustrated in Fig. 4, an image is first separated into four parts for each pixel which are clockwise named parts 1, 2, 3 and 4, respectively. By averaging in each part, four candidate background values are produced for each pixel. It can be further explained by the following formulation:

$$BG(i,j) = \left\{BG_1(i,j), \ldots, BG_k(i,j)\right\} \tag{1}$$

where $BG_k\,(i,j)$ is the average CIELAB vector of each part in image I at position (i,j) as given by:

$$BG_k(i,j) = \frac{1}{N_k}\sum_{I(i,j)\in \text{part } k} I(i,j) \tag{2}$$

and in which N_k is the number of pixels in part k, and k = 1–4.

Once candidate background maps are generated, the corresponding saliency maps can be easily produced by measuring the Euclidean distance between background maps and original images.

$$D(i,j) = \left\{\left\|I_f(i,j) - BG_1(i,j)\right\|, \ldots, \left\|I_f(i,j) - BG_k(i,j)\right\|\right\} \tag{3}$$

where $I_f\,(i,j)$ is the image pixel vector value after Gaussian blurring with a 5 × 5 separable binomial kernel and $||\ ||$ is L_2 norm.

Then, candidate saliency maps are fused as below:

$$S_B(i,j) = \min\left\{D(i,j)\right\} \tag{4}$$

For efficient computation, we also propose an approximate implementation. Images are separated into 5 × 5 patches with non-overlap and background maps are generated in patch-level, and then a smoothing operation (Gaussian blurring) is needed to reduce the artifacts introduced in the patch-based background estimation. Integral images are also used to speed up in our implementation as done by [6,14]. We call it SMB Patch (SMBP) which meets the applications that require high speed, such as image thumbnail generation/cropping for batch image browsing [20], and bounding box based object extraction [21]. Experiment in Section 3 demonstrates computation cost is significant saved with almost no loss in performance. For applications that require high accuracy such as image segmentation [22], we propose to use SMB Original (SMBO) implemented in pixel-level.

2.2. Spatial distribution

For textured images, candidate saliency maps generated by Eq. (4) are not satisfactory enough. As can be seen from Fig. 5, some background regions are falsely marked as salient regions and some salient regions are not fully highlighted. To solve it, spatial distribution information is further introduced to improve performance.

Most of the previous methods [4,5,14] neglect the spatial relation which is proven to be an important clue for saliency detection [17]. Ideally salient objects are generally grouped together while background colors are mostly distributed over the whole image [3,17]. Based on this observation, we incorporate spatial distribution as a high-level factor to suppress background simultaneously highlights salient objects.

Spatial variance is used in our work to measure the spatial distribution of the salient pixels. As motivated before, high variance indicates a spatial widely distributed pixel which should be considered as a background pixel and suppressed. Hence we define it as:

$$S_S(i,j) = \exp\left(-\frac{1}{2\sigma^2}\left\|p(i,j) - \bar{p}\right\|\right) \tag{5}$$

where $p\,(i,j)$ is the position of color (i,j) and $\bar{p}$ is the mean position of the salient object defined as:

$$\bar{p} = \frac{1}{N_S}\sum_{(i,j)\in I} S_B(i,j)p(i,j) \tag{6}$$

where N_S is the sum of $S_B\,(i,j)$. σ controls the strength of spatial weighting and larger values reduce the effect of spatial weighting. In our implementation, we set σ as 0.3 of image width and height.

As the final step, we incorporate the spatial saliency as a high-level factor into the saliency map proposed above.

$$\text{Saliency }(i,j) = S_B(i,j)S_S(i,j) \tag{7}$$

Background regions falsely marked as salient regions can be suppressed with the help of the product. The final resulting saliency map is normalized to [0, 1]. Visual examples in Fig. 5 show the effectiveness of the proposed scheme.

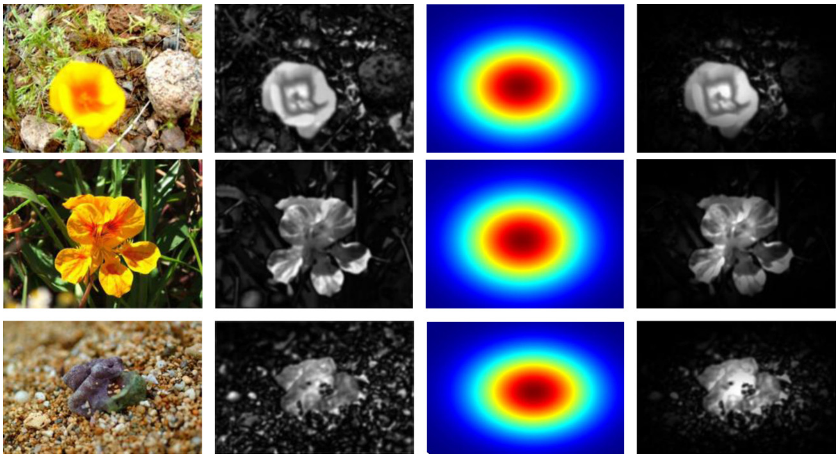

Fig. 5. From left to right: input image, our saliency map without spatial information, spatial saliency, and our final saliency map combining spatial saliency.

In summary, our algorithm first generates multiple background maps through local averaging operation. It then measures the color differences between background maps and original image to compute candidate saliency maps. Saliency map can be easily generated by fusing them. With the help of spatial distribution information, we further reduce the probability of falsely marking background as salient regions and simultaneously highlight the inside region of salient objects. Encouragingly, a combination of our spatial saliency with previous methods lacking spatial distribution information can usually improve both.

3. Experiments

3.1. Dataset and evaluation criterion

We evaluate our approach quantitatively on public dataset MSRA-1000 with binary ground truth, which is provided by Achanta et al. [4] and has been widely used in saliency evaluation. Precision, Recall, *F*-measure and mean absolute error (MAE) are evaluated in our experiments.

Given corresponding masks the precision and recall rate for each image are quantified as follows:

$$\text{precision} = \frac{\sum_{i=1}^{W}\sum_{j=1}^{H} B(i,j)G(i,j)}{\sum_{i=1}^{W}\sum_{j=1}^{H} B(i,j)} \tag{8}$$

$$\text{recall} = \frac{\sum_{i=1}^{W}\sum_{j=1}^{H} B(i,j)G(i,j)}{\sum_{i=1}^{W}\sum_{j=1}^{H} G(i,j)} \tag{9}$$

where *B* is the binary object mask generated by thresholding corresponding saliency map and *G* is the corresponding ground truth.

Fixed thresholding and adaptive thresholding are performed, respectively, in the process of generating binary object masks. Based on this, two different experiments are performed. In our first evaluation we segment saliency maps using a fixed threshold ranged from 0 to 255 and then calculate precision and recall rates by comparing with the corresponding ground truth masks. In the second evaluation, adaptive threshold is performed to binarize the saliency maps, which is defined as twice the mean saliency values [4]:

$$T_a = \frac{2}{W \times H}\sum_{i=1}^{W}\sum_{j=1}^{H} \text{Saliency }(i,j) \tag{10}$$

where *W* and *H* are the width and the height of the saliency map, respectively.

In addition to precision and recall, *F*-measure is also calculated to obtain an overall performance, and β^2 is set to 0.3 the same as [4,5,16,17]:

$$F_\beta = \frac{(1+\beta^2)\,\text{Precision} \times \text{Recall}}{\beta^2 \times \text{Precision} + \text{Recall}} \tag{11}$$

The MAE is calculated between the continuous saliency map and the binary ground truth, which is defined as:

$$\text{MAE} = \frac{1}{W \times H}\sum_{i=1}^{W}\sum_{j=1}^{H} \left|\text{Saliency }(i,j) - G(i,j)\right| \tag{12}$$

Lower MAE value indicates better performance, which provides a better estimate of dissimilarity between the saliency map and ground truth [17].

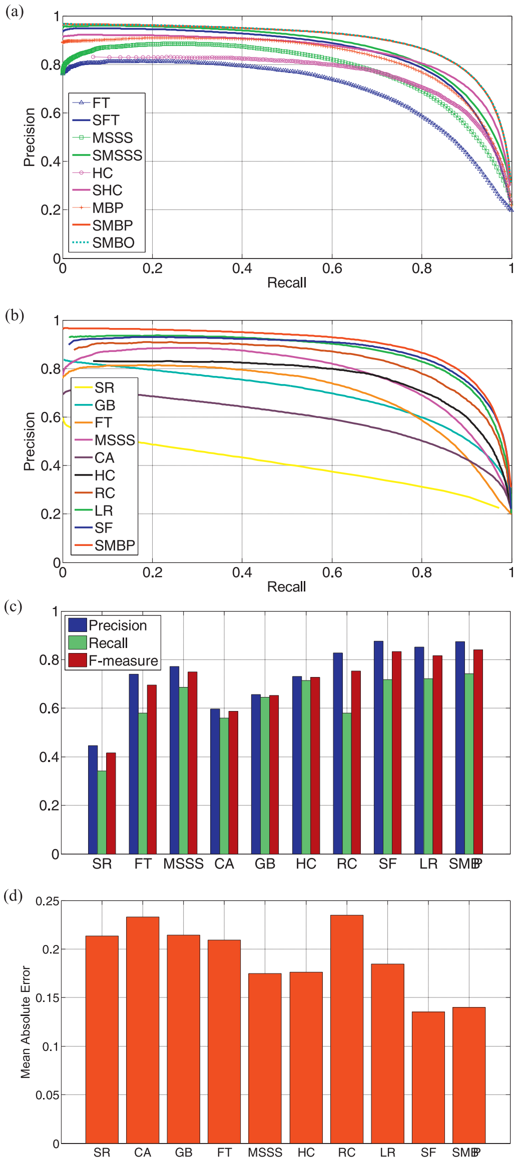

Fig. 6. Experiment results: (a) Precision–recall curves of different methods combining our spatial saliency. (b) Precision–recall curve for all algorithms. (c) Average precision, recall and *F*-measure for adaptive thresholds. (d) Mean absolute error of the different saliency methods to ground truth.

3.2. Validation of spatial distribution

In this experiment, we first compare our approach with different components produced by Eqs. (4) and (7), shortly called MBP and SMBP, respectively. To further test whether our spatial saliency is complementary to previous methods, we also incorporate it to some previous methods: FT [4], MSSS [14], and HC [5]. This incorporation can be simply implemented by replacing S_B with corresponding saliency map in Eq. (7). As illustrated in Fig. 6(a), all these methods perform over a wide range after combining our spatial saliency due to the complementary property of spatial distribution. It also found that SMBP and SMBO almost have the same precision–recall curve. Thus only SMBP is compared in the following experiments.

3.3. Performance evaluation

We also compare our approach with 9 typical state-of-art methods, including GB [15], SR [12], FT [4], MSSS [14], CA [3], HC [5], RC

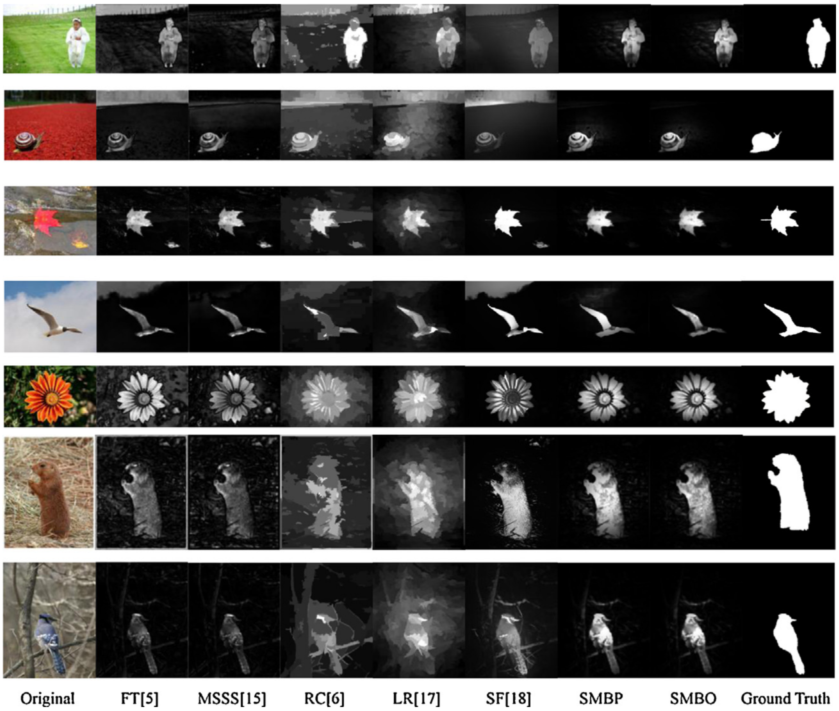

Fig. 7. Visual comparison of different methods on the MSRA-1000 database.

Table 1
Comparison of averaging running times.

Method	FT	FT	MSSS	HC	RC	SMBP	SMBO
Time (s)	0.023	0.103	1.447	0.017	0.245	0.128	3.017
Code	C++	Matlab	Matlab	C++	C++	Matlab	Matlab

[5], LR [16] and SF [17]. Such methods are selected for comparison due to their large varieties. Among them, FT, MSSS, HC and SF output full resolution saliency maps and perform efficiently, SR works in the frequency domain, RC is regional contrast based, LR introduces low rank matrix recovery theory and incorporates high-level priors, and LR, SF are proposed recently. Saliency maps of SR, FT, HC, and RC are generated in C++ provide by [5]. For the other methods, we used the authors' implementations or results.

Performance results are shown in Fig. 6. It is clear that our approach not only gets better precision–recall curve but also achieves the best precision, recall and *F*-measure. For MAE measure, our approach also outperforms the others except SF due to its operation of containing at least 10% saliency pixels in the final step. Limited by space, we present only some examples of visual results of our scheme compared with some recent methods in Fig. 7. Visually, the SMBP method also exceeds the other methods. We can find that our saliency maps are full resolution. Although the RC and LR methods successfully highlight the salient objects too, their saliency maps are not pixel-level. It is also interesting to observe that the previous methods cannot work well when background of image contains two or more large dissimilar parts such as sky and lawn which are very common in natural scenes. Encouragingly, such parts will not be marked as saliency regions in our approach which can be clearly seen in Fig. 7.

Running times of different methods are shown in Table 1. Timing tests have been taken on the same hardware devices (Intel Core 2 Duo CPU 2.26 GHz with 4GB RAM). For better comparison, we also test the Matlab version of FT which is one of the most efficient saliency detection methods, and our approach is only slightly slower than FT, which demonstrates it is efficient and suitable for real-time applications.

4. Conclusion and future works

In this paper, we presented a multiple background maps based saliency detection approach which can be implemented in linear computation time and generate full resolution saliency maps. This is mainly contributed by our robust multiple background estimation which is implemented by averaging in proper local regions. Once generated background maps, saliency maps are easily produced only by measuring the differences between the original images and their corresponding background maps, which is similar to foreground detection in video applications. To further reduce the probability of falsely marking background as salient regions, we also incorporate the spatial distribution as a high-level factor which can be easily combined into the previous methods. Experimental results on public dataset indicate the proposed scheme achieves better results in comparison with various state-of-art works.

Although our multiple background estimation based saliency approach holds for most images and performs well in general, inevitable it fails when our background estimation are invalid. For future work we plan to improve it from the following directions: (1) incorporate the spatial distribution in a better way to make it more suitable for multiple objects far from each other; (2) combine candidate saliency maps in a way better than our straightforward combination by minimizing; (3) fully highlight the salient objects especially these regions having similar color with background.

Acknowledgments

The authors would like to express their sincere thanks to the anonymous reviewers for their invaluable suggestions and comments to improve this paper. This work is supported in part by the National Key Technology R&D Program under Grant 2011BAK07B03.

References

[1] J. Han, K.N. Ngan, M. Li, H. Zhang, Unsupervised extraction of visual attention objects in color images, IEEE Trans. Circuits Syst. Video Techn. 16 (1) (2006) 141–145.

[2] U. Rutishauser, D. Walther, C. Koch, P. Perona, Is bottom-up attention useful for object recognition? in: Proc. IEEE CVPR, 2004, pp. 37–44.

[3] S. Goferman, L. Zelnik Manor, A. Tal, Context aware saliency detection, in: Proc. IEEE CVPR, 2010, pp. 2376–2383.

[4] R. Achanta, S. Hemami, F. Estrada, S. Susstrunk, Frequency tuned salient region detection, in: Proc. IEEE CVPR, 2009, pp. 1597–1604.

[5] M.M. Cheng, G.X. Zhang, N.J. Mitra, X. Huang, S.M. Hu, Global contrast based salient region detection, in: Proc. IEEE CVPR, 2011, pp. 409–416.

[6] S. Frintrop, M. Klodt, E. Rome, A real-time visual attention system using integral images, in: International Conference on Computer Vision Systems, 2007.

[7] J. Reynolds, R. Desimone, Interacting roles of attention and visual salience in V4 Neuron 37 (5) (2003) 853–863.

[9] L. Itti, C. Koch, E. Niebur, A model of saliency-based visual attention for rapid scene analysis, IEEE Trans. Pattern Anal. Mach. Intell. 20 (11) (1998) 1254–1259.

[10] Y.F. Ma, H.J. Zhang, Contrast-based image attention analysis by using fuzzy growing, in: ACM Multimedia, 2003, pp. 374–381.

[11] Thanarat Horprasert, David Harwood, Larry S. Davis, A statistical approach for real-time robust background subtraction and shadow detection, in: Proc. IEEE ICCV, 1999.

[12] X. Hou, L. Zhang, Saliency detection: a spectral residual approach, in: Proc. IEEE CVPR, 2007, pp. 1–8.

[13] Y. Zhai, M. Shah, Visual attention detection in video sequences using spatiotemporal cues, in: ACM Multimedia, 2006, pp. 815–824.

[14] R. Achanta, S. Susstrunk, Saliency detection using maximum symmetric surround, in: Proc. ICIP, 2010.

[15] J. Harel, C. Koch, P. Perona, Graph-based visual saliency, Adv. Neural Inf. Process. Syst. (2007) 19–545.

[16] X.H. Shen, Y. Wu, A unified approach to salient object detection via low rank matrix recovery, in: Proc. IEEE CVPR, 2012.

[17] F. Perazzi, P. Krahenbuhl, Y. Pritch, A. Hornung, Saliency filters: contrast based filtering for salient region detection, in: Proc. IEEE CVPR, 2012.

[18] A. Adams, J. Baek, M.A. Davis, Fast high-dimensional filtering using the permutohedral lattice, Comput. Graph. Forum 29 (2) (2012) 753–762.

[19] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, H. Shum, Learning to detect a salient object, in: Proc. IEEE CVPR, 2007, pp. 1–8.

[20] L. Marchesotti, C. Cifarelli, G. Csurka, A framework for visual saliency detection with applications to image thumbnailing, in: Proc. IEEE ICCV, 2009.

[21] N.J. Butko, J.R. Movellan, Optimal scanning for faster object detection, in: Proc. IEEE CVPR, 2009.

[22] V. Lempitsky, P. Kohli, C. Rother, T. Sharp, Image segmentation with a bounding box prior, in: Proc. IEEE ICCV, 2009.