HIERARCHICAL SALIENCY OPTIMIZATION

Hanpei Yang, Weihai Li

Key Laboratory of Electromagnetic Space Information, Chinese Academy of Sciences
School of Information Science and Technology, University of Science and Technology of China

ABSTRACT

A variety of methods have been proposed for object level saliency detection, which is useful for many content-based computer vision applications. Unlike most previous work that integrate multiple low level cues to compute the saliency map, this paper presents a novel hierarchical optimization model. First, we compute a rough saliency map using HS method, and then, boundary and foreground seeds are extracted from it, which guide the computation of the background and foreground saliency maps, respectively. Next, a combination of the two saliency maps is performed. In the end, Cellular Automata is applied to optimize it and a threshold method is taken to make the optimized saliency map closer to the ground truth. Experiments on three large datasets demonstrate that the proposed method performs favorably against the state-of-the-art methods in terms of F-measures and MAEs.

Index Terms— saliency, hierarchical, seed extraction, Cellular Automata

1. INTRODUCTION

Saliency detection is the process of identifying the location of the salient object that grabs a viewers attention, which is different from the traditional methods that predict human fixation [1]. Saliency detection are universally applied in image cropping [2], object aware image retargeting [3] [4], content-based image retrieval, and so on. Numerous methods are proposed to estimate the probability of the foreground image region, which can be divided into two kinds, either bottom-up or top-down approaches. The bottom-up approaches usually exploit the prior knowledge, such as center prior [5][6], boundary prior [7] [8], contrast prior [9] [10]. Center prior assumes the salient objects are often framed near the center of the images, but in many cases, this assumption is not true and it leads to mistake of highlighting some background region near the image center. Some methods take boundary prior to enhance the computation of saliency maps, most of which simply regard the image boundary as background. This is fragile and may fail when the salient object touch the boundary.

Considering the above-mentioned issues, we propose a hierarchical saliency optimization model. First, the salient map of the first layer is computed according to [11], based on which, we can get a more reliable location of background boundary seeds and foreground object seeds. A background-prior-based saliency map and a foreground-prior-based saliency map are computed, respectively. The pixel-wise combination of the two maps is the saliency map of the second layer. Cellular Automata [12] is utilized to enforce consistency among similar image patches and modify the saliency values of boundary cells misclassified as background seeds through interactions with neighbors. This is the final saliency map and the saliency map of the third layer. Unlike previous methods that try to integrate multiple cues and get the final map one time effort, we try to optimize the saliency map layer by layer, which is the so-called hierarchical saliency optimization model. In layer 3, we also introduce a threshold method to make the final saliency map more similar to the ground truth mask, which is more in line with the application of the saliency maps.

The rest of the paper is organized as follows. The proposed hierarchical optimization model is described in Section 2. The experiments on three popular datasets are shown in Section 3 and the conclusion is made in Section 4.

2. HIERARCHICAL OPTIMIZATION MODEL

In this section, we detail the process of the hierarchical optimization model. The model is decomposed into three layers and each layer generates a saliency map, which is the basis of the next layer. The flowchart is shown in Fig. 1.

2.1. Layer 1: hierarchical saliency detection

The first layer use the HS method [11] to generate the saliency map (HS map), which gives a fairly accurate information about the location of the background and foreground. These information contribute to the following computation of two saliency maps based on the background prior and foreground prior, respectively.

2.2. Layer 2: information about boundary and object, the combination map

To better utilize the structural information of an input image, simple linear iterative clustering (SLIC) [13] algorithm is per-

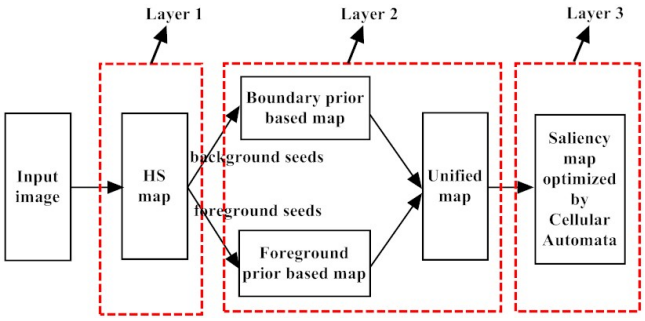

Fig. 1. Flowchart of the hierarchical saliency optimization.

formed on the input image and get the superpixels. We binarize the HS map with an adaptive threshold generated by OTSU's method [14].

2.2.1. *Boundary seeds and background-based map*

The binary HS map is used to guide the selection of the boundary superpixels. 1 is for the salient object and 0 is for the background in the binary map. The selected boundary superpixels should meet two requirements. First, they should be on the border of the input image and second, the same location of the superpixels in the binary map should not contain any 1. In this situation, the boundary superpixels selected can avoid containing foreground objects with a high probability unless the HS map makes great mistakes.
When computing the boundary connectivity in [15], they regard all the superpixels near the border of the input image as boundary superpixels. We replace it with our selected boundary superpixels, which is more accurate. Then we can get a robust background-based saliency map using method in [15].

2.2.2. *Object seeds and foreground-based map*

We check the binary HS map and the superpixels at the same time, and label those superpixels as foreground seeds, in which over 90% of the corresponding positions in the binary HS map are 1. The foreground-based map is calculated as shown in Eq. (1).

$$\mathrm{S}_i^f = \sum_{n \neq i, n \in FG} \frac{\lambda}{d\left(\mathrm{c}_i, \mathrm{c}_n\right) + \partial d\left(\mathrm{L}_i, \mathrm{L}_n\right)} \tag{1}$$

Where the centroid location vector and mean CIELab color vector of the ith superpixel are denoted by L_i and c_i, respectively. $d(\mathrm{c}_i, \mathrm{c}_n)$ and $d(\mathrm{L}_i, \mathrm{L}_n)$ are respectively the Eiuclidean color and spatial distance between the ith superpixel and the nth superpixel which belongs to the foreground seed set FG. λ and ∂ are set to 1 in our experiment as in [16].

Type	AUC	MeanF	MAE
$FG\_map + BG\_map$	**0.894**	**0.659**	0.171
$FG\_map * BG\_map$	0.875	0.656	0.171
$BG\_map * (1 - e^{-6 * FG\_map})$	0.893	0.648	0.176
$BG\_map * e^{(FG\_map)}$	0.883	0.658	**0.170**
$BG\_map + 0.5 * FG\_map$	0.891	0.658	0.171
$e^{(FG\_map)} + e^{(BG\_map)}$	0.886	0.657	0.171

Table 1. Comparison between 6 combination methods (Dataset: ECSSD, best in bold)

2.2.3. *The combination map*

We denote the saliency map generated by Section 2.2.1 as BG_map, and the saliency map generated by Section 2.2.2 as FG_map. AUC, MeanF, MAE are short for area under curve, mean F-measure and mean absolute error respectively (details in Section 3). As shown in Table 1, six combination methods are tested on ECSSD dataset with three evaluation indexes: AUC, MeanF and MAE. In general, adding the background-based map and the foreground-based map directly is the best of all.

2.3. Layer 3: saliency map optimization

Cellular Automata [17] is a dynamic system with a complex self-organizing behavior and it is widely used to simulate the evolution process of many complicated systems. As done in [12], we first compute the impact factor matrix: $\mathrm{F} = [\mathrm{f_{ij}}]_{\mathrm{N*N}}$.

$$f_{ij} = \begin{cases} \exp\left(\frac{-\|c_i, c_j\|}{\sigma^2}\right) & j \in NB(i) \\ 0 & i = j \ or \ otherwise \end{cases} \tag{2}$$

Where $NB(i)$ denotes the neighbors of cell i, σ^2 equals to 0.1 and controls the strength of how likely the two superpixels are . In order to normalize the impact matrix, $\mathrm{D} = diag\{d_1, d_2, ..., d_N\}$ is generated, where $d_i = \sum_j f_{ij}$, and then F^*=$\mathrm{D}^{-1} \cdot \mathrm{F}$.
Coherence matrix is proposed to balance the fact that each cells next state is determined by itself and its neighbors, denoted as $C^* = diag\{c_1^*, c_2^*, ..., c_N^*\}$.

$$c_i = \frac{1}{\max(f_{ij})} \tag{3}$$

$$c_i^* = a \cdot \frac{c_i - \min(c_j)}{\max(c_j) - min(c_j)} + b \tag{4}$$

Where a and b are set to 0.6 and 0.2, empirically. $j = 1, 2, ..., N$. N is the number of superpixels or number of cells. The updating rule is shown in Eq. (5) and the initial S^{t} when t=0 is the combination map in Section 2.2.3. After T times iterations, S^T is regarded as the optimized saliency map.

$$\mathrm{S}^{\mathrm{t+1}} = \mathrm{C}^* \cdot \mathrm{S}^{\mathrm{t}} + (\mathrm{I} - \mathrm{C}^*) \cdot \mathrm{F}^* \cdot \mathrm{S}^{\mathrm{t}} \tag{5}$$

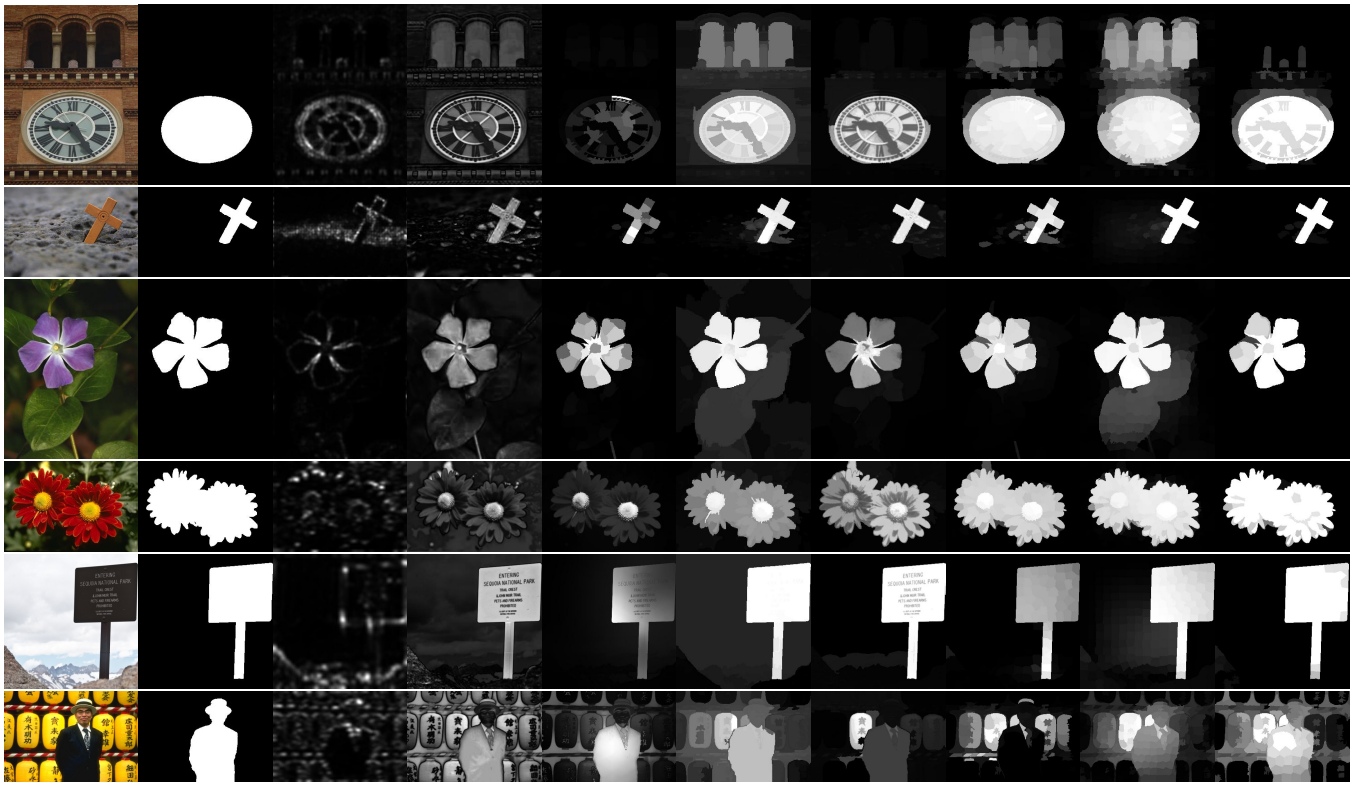

Fig. 2. Visual comparison of our saliency maps with 7 state-of-the-art methods.

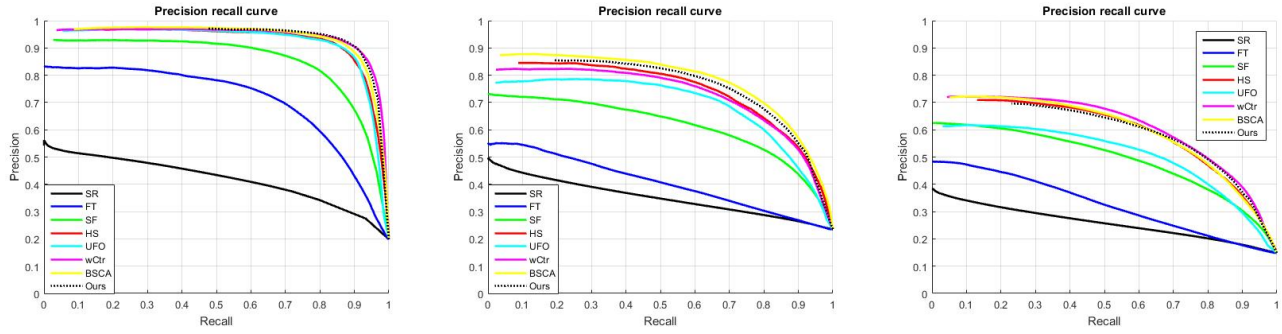

Fig. 3. P-R curves. (a) on ASD; (b) on ECSSD; (c) on DUT-OMRON

Actually the optimized map has some small values which makes it not much like the ground truth mask. In actual application such as object segmentation or cropping, the closer the saliency map is to the ground truth mask, the better result it can achieve. So empirically, we set the salient values that are larger than θ_2 to 1, and set the values that are less than θ_1 to 0. $f(x, y)$ means the value in position (x, y).

$$f(x,y) = \begin{cases} 0 & f(x,y) \le \theta_1 \\ f(x,y) & \theta_1 < f(x,y) < \theta_2 \\ 1 & f(x,y) \ge \theta_2 \end{cases} \quad (6)$$

Where θ_1 and θ_2 are set to 0.05 and 0.9 empilically for the consideration of the possible wrongly detected saliency maps as well as the properly detected saliency maps.

3. EXPERIMENTS

We evaluate our algorithm on three public available datasets: ASD [18], ECSSD [11], and DUT-OMROM [9]. ASD contains 1000 images, where the salient objects are manually labeled with pixel-wise ground truth. ECSSD is short for Com-

Table 2. Quantitative comparison of MAE and MeanF on ASD (best in bold)

Evaluation index	SR07	FT09	SF12	UFO13	HS13	wCtr14	BSCA15	Our method
MeanF	0.127	0.483	0.639	0.760	0.801	0.838	0.822	**0.876**
MAE	0.212	0.206	0.129	0.109	0.111	0.065	0.086	**0.059**

Table 3. Quantitative comparison of MAE and MeanF on ECSSD (best in bold)

Evaluation index	SR07	FT09	SF12	UFO13	HS13	wCtr14	BSCA15	Our method
MeanF	0.142	0.284	0.411	0.533	0.620	0.620	0.648	**0.665**
MAE	0.264	0.289	0.228	0.205	0.228	0.171	0.182	**0.170**

Table 4. Quantitative comparison of MAE and MeanF on DUT-OMRON (best in bold)

Evaluation index	SR07	FT09	SF12	UFO13	HS13	wCtr14	BSCA15	Our method
MeanF	0.137	0.268	0.401	0.451	0.520	0.550	0.522	**0.558**
MAE	0.181	0.250	0.183	0.174	0.227	**0.144**	0.191	0.162

plex Scene Saliency Dataset and it is more challenging than ASD because of its semantically meaningful but structurally complex scenes. The last DUT-OMRON contains 5168 challenging images. The target is of various sizes and the background is quite complicated. 7 state-of-the-art methods are selected as baselines, including SR07 [19], FT09[18], SF12 [20], UFO13[21], HS13 [11], wCtr14 [15], BSCA15 [12].

Figure 2 gives some examples to show the visual comparison between various methods. The saliency maps generated by the proposed algorithms highlight the salient objects well with fewer noisy results. The last row shows an example of failed detection of salient object resulting from the colorful and confusing background.

We compare our algorithms with the state-of-the-art methods using precision and recall (P-R) curve, F-measure and mean absolute error (MAE). The saliency map is segmented with the threshold ranging from 0 to 255 to get the binary map, which is used to be compared with the ground truth mask to compute the precision and recall later. The P-R curve is composed of the mean precision and recall of all the saliency maps on different thresholds. F-measure is computed as followed:

$$\mathrm{F}_\beta = \frac{\left(1+\beta^2\right)\mathrm{Precision}\times\mathrm{Recall}}{\beta^2\times\mathrm{Precision}+\mathrm{Recall}} \tag{7}$$

Where β^2 is set to 0.3 as done in [9]. MeanF is the average of all the F-measures. As showed in Fig. 3, when recall achieves the largest value from 0.9 to 1, wCtr14, BSCA15 and our method have similar results.

Although P-R curves are commonly used, they limit in that they only concentrate on whether the object saliency is higher than the background saliency. But in applications, the difference between the saliency map and the ground truth is of much importance, so mean absolute error (MAE) is introduced into evaluation, which shows the average per-pixel difference between the binary ground truth and the saliency map, normalized to [0,1].

$$MAE = \frac{1}{H}\sum\nolimits_{h=1}^{H}|S(h)-GT(h)| \tag{8}$$

Table 2-4 show the MeanF and MAE of different methods on ASD, ECSSD and DUT-OMRON, respectively. Obviously, our method has the best MeanF on all three datasets and does the best in MAE on ASD and ECSSD. Comparing to the BCSA15 [12], our method at least successfully reduces the MAE on ASD and ECSSD by 31.39% and 6.6% respectively, which is much better than the result declared in [22]. In general, our method has a good performance in MAE and MeanF, which are important evaluation indexes in applications.

4. CONCLUSION

In this paper, we propose a novel hierarchical saliency optimization model. A saliency map is obtained by HS method firstly, from which background seeds and foreground seeds are extracted. In this way, we can avoid taking the false boundary pixels as background seeds in a large probability. Background-based saliency map and foreground-based map are computed by background and foreground seeds, respectively. We integrate the two maps into the unified map in order to take both advantages of them, and use the Cellular Automata to optimize it. Threshold method is adopted to further optimize the saliency map in order to make it closer to the ground truth mask. Experiments show that our methods do well in evaluation index of MeanF and MAE, which means that our final saliency map is much more similar to the ground truth mask and is much more suitable for applications like object segmentation and cropping.

5. REFERENCES

[1] Laurent Itti, Christof Koch, and Ernst Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254–1259, 1998.

[2] L Marchesotti, C Cifarelli, and G Csurka, "A framework for visual saliency detection with applications to image thumbnailing," *2009 IEEE 12th International Conference on Computer Vision*, , no. Iccv, pp. 2232–2239, 2009.

[3] Yuanyuan Ding, Jing Xiao, and Jingyi Yu, "Importance filtering for image retargeting," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2011, pp. 89–96.

[4] Jin Sun and Haibin Ling, "Scale and object aware image thumbnailing," *International Journal of Computer Vision*, vol. 104, no. 2, pp. 135–153, 2013.

[5] Stas Goferman, Lihi Zelnik-Manor, and Ayellet Tal, "Context-aware saliency detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 10, pp. 1915–1926, 2012.

[6] Huaizu Jiang, Jingdong Wang, Zejian Yuan, Tie Liu, Nanning Zheng, and Shipeng Li, "Automatic Salient Object Segmentation Based on Context and Shape Prior," *British Machine Vision Conference*, pp. 1–12, 2011.

[7] Yichen Wei, Fang Wen, Wangjiang Zhu, and Jian Sun, "Geodesic saliency using background priors," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2012, vol. 7574 LNCS, pp. 29–42.

[8] Huaizu Jiang, Jingdong Wang, Zejian Yuan, Yang Wu, Nanning Zheng, and Shipeng Li, "Salient object detection: A discriminative regional feature integration approach," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2013, pp. 2083–2090.

[9] Chuan Yang, Lihe Zhang, Huchuan Lu, Xiang Ruan, and Ming Hsuan Yang, "Saliency detection via graph-based manifold ranking," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3166–3173.

[10] Ming Ming Cheng, Jonathan Warrell, Wen Yan Lin, Shuai Zheng, Vibhav Vineet, and Nigel Crook, "Efficient salient region detection with soft image abstraction," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 1529–1536.

[11] Qiong Yan, Li Xu, Jianping Shi, and Jiaya Jia, "Hierarchical saliency detection," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2013, pp. 1155–1162.

[12] Yao Qin, Huchuan Lu, Yiqun Xu, and He Wang, "Saliency detection via Cellular Automata," *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 110–119, 2015.

[13] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, P Fua, and S Susstrunk, "SLIC Superpixels," *EPFL Technical Report 149300*, , no. June, pp. 15, 2010.

[14] Nobuyuki Otsu, "A threshold selection method from gray-level histograms," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 9, no. 1, pp. 62–66, 1979.

[15] Wangjiang Zhu, Shuang Liang, Yichen Wei, and Jian Sun, "Saliency optimization from robust background detection," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 2814–2821, 2014.

[16] Jianpeng Wang, Huchuan Lu, Xiaohui Li, Na Tong, and Wei Liu, "Saliency detection via background and foreground seed selection," *Neurocomputing*, vol. 152, pp. 359–368, 2015.

[17] John Neumann, "The general and logical theory of automata.," *Cerebral mechanisms in behavior; the Hixon Symposium*, vol. 10, no. 3, pp. 1–41, 1951.

[18] Radhakrishna Achantay, Sheila Hemamiz, Francisco Estraday, and Sabine Süsstrunky, "Frequency-tuned salient region detection," *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2009*, , no. Ic, pp. 1597–1604, 2009.

[19] Xiaodi Hou and Liqing Zhang, "Saliency detection: A spectral residual approach," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2007.

[20] Federico Perazzi, Philipp Krahenbuhl, Yael Pritch, and Alexander Hornung, "Saliency filters: Contrast based filtering for salient region detection," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 733–740, 2012.

[21] Peng Jiang, Haibin Ling, Jingyi Yu, and Jingliang Peng, "Salient region detection by UFO: Uniqueness, focusness and objectness," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 1976–1983.

[22] Hong Liu, Shuning Tao, and Zheyuan Li, "Saliency detection via global-object-seed-guided cellular automata," in *2016 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2016, pp. 2772–2776.