

Two-Stage Saliency Fusion for Object Segmentation

Guangling Sun, Jingru Ren, Zhi Liu^(✉), and Wei Shan

School of Communication and Information Engineering,
Shanghai University, Shanghai 200444, China
sunguangling@shu.edu.cn, claudiaruru@126.com,
liuzhisjtu@163.com, shanwei1993@126.com

Abstract. This paper proposes an effective two-stage saliency fusion method to generate the fusion map, which is used as a prior for object segmentation. Given multiple saliency maps generated by different saliency models, the first stage is to produce two fusion maps based on average and min-max statistics, respectively. The second stage is to perform the Fourier transform (FT) on the two fusion maps, and to combine the amplitude spectrum of average fusion map and the phase spectrum of min-max fusion map, so as to reform the spectrum, and the final fusion map is obtained by using the inverse FT on the reformed spectrum. Last, object segmentation is performed under graph cut by using the final fusion map as a prior. Extensive experiments on three public datasets demonstrate that the proposed method facilitates to achieve the better object segmentation performance compared to using individual saliency map and other fusion methods.

Keywords: Object segmentation · Saliency fusion · Statistical fusion
Fourier transform

1 Introduction

Automatic object segmentation is a key requirement in a number of applications [1]. Some object/background prior is necessary for object segmentation, and undoubtedly, saliency map is an effective option. Saliency models that generate saliency maps can be classified into two categories. One is concerned with predicting human fixation, and the pioneer work of this category originated from [2]. The other one tries to detect salient objects with well-defined boundaries to highlight the complete objects. Obviously, the latter one is a more suitable prior for object segmentation. Fortunately, plenty of saliency models with high performances are developed in the recent years [3]. Especially, integrating multiple features, multiple levels, multiple scales, multiple stages and multiple saliency maps have demonstrated the effectiveness for improving saliency detection performance, such as multiple kernel boosting on multiple features [4] and Bayesian integration of low-level and mid-level cues [5]. In [6], the stacked denoising autoencoders are used to model background and generate deep reconstruction residuals at multiple scales and directions, and then the residual maps are integrated to obtain the saliency map. In [7], four Mahalanobis distance maps based on the four spaces of background-based distribution are first integrated, and then are enhanced within Bayesian

perspective and refined with geodesic distance to generate the saliency map. In [8], the sum fusion method is evaluated and the results verify that combining several best saliency maps can actually enhance saliency detection performance. A data-driven saliency aggregation approach under the conditional random field framework is proposed in [9], which focuses on modeling the contribution of individual saliency map. Besides, the work in [10] integrates different saliency maps for fixation prediction and also demonstrates the performance improvement. A selection framework from multiple saliency maps adaptive to the input image is proposed in [11].

In this paper, we focus on fusion of multiple saliency maps, and specifically, the fused saliency map is used as an object prior for effective object segmentation. A high-quality saliency map for segmentation should effectively highlight object pixels and suppress background pixels, i.e., the contrast of saliency values between object pixels and background pixels should be as high as possible. For this purpose, the proposed saliency fusion method has the following two main contributions, which makes it different from the previous works: (1) The saliency fusion is performed in both spatial domain and frequency domain; (2) The fusions in the two domains are implemented in two stages in sequel to achieve a controllable SNR-contrast trade-off. Specifically, in the first stage, based on the statistics in the spatial domain, the average fusion map and min-max fusion map are generated. In particular, the saliency values of potential object pixels and background pixels are increased and decreased, respectively, as much as possible by using the min-max fusion. In the second stage, the above two fusion maps are further fused in the frequency domain. Specifically, the amplitude spectrum of average fusion map and the phase spectrum of min-max fusion map are integrated to reduce the errors and preserve the prominent contrast between object and background simultaneously.

2 Proposed Saliency Fusion for Object Segmentation

An overview of the proposed two-stage saliency fusion for object segmentation is illustrated in Fig. 1. Given the input image in Fig. 1(a), a number of existing saliency models are exploited to generate multiple saliency maps as shown in Fig. 1(b). For example, we can use the top six high-performing saliency models as reported in [3] to generate six saliency maps for the input image. The proposed two-stage saliency fusion method sequentially performs fusion in spatial domain and frequency domain, which are described in Sects. 2.1 and 2.2, respectively. Then the fusion map is used as a prior to perform object segmentation, which is described in Sect. 2.3.

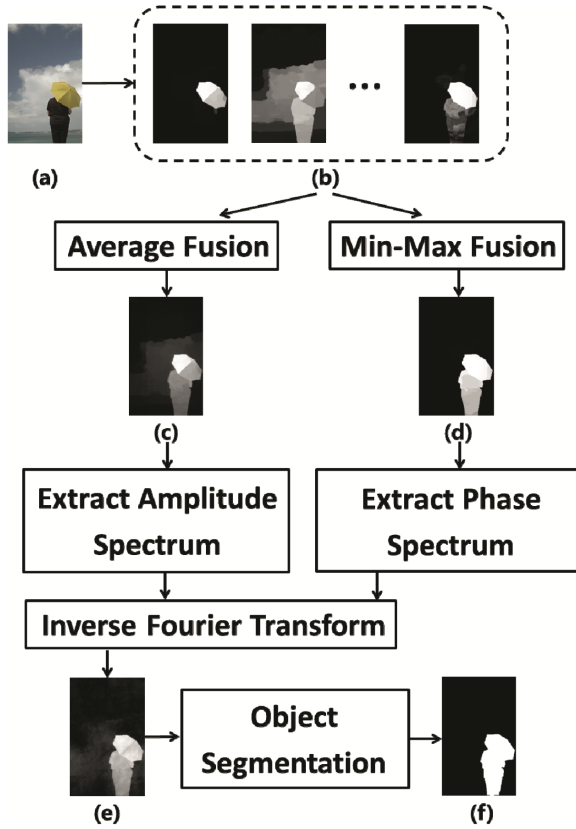


Fig. 1. Overview of the proposed method. (a) Input image; (b) multiple saliency maps generated by using different saliency models; (c) average fusion (AF) map; (d) min-max fusion (MMF) map; (e) AAPMF map by integrating the amplitude spectrum of AF map and the phase spectrum of MMF map; (f) segmentation result via graph cut.

2.1 Fusion Based on Average and Min-Max Statistics

The average statistics is one widely used statistics, and it is effective to average all saliency maps at pixel level for improving saliency detection performance [8]. Therefore we use the average operation to first generate the average fusion (AF) map as shown in Fig. 1(c). Although the AF map achieves performance improvement, it is not always a good candidate of prior for segmentation due that it cannot sufficiently highlight object pixels and suppress background pixels. In other words, the AF map generally weakens the contrast between object pixels and background pixels. Therefore, as a complement, we propose a novel fusion scheme which relies on the minimum and maximum of all saliency values. Specifically, the fusion value of each pixel is adaptive to the corresponding pixel's value in the AF map and a threshold. For each pixel, if its corresponding pixel's value in the AF map is less than a threshold, the pixel will be determined to be a potential background pixel, and the minimum among all saliency values of the pixel

will be assigned to the pixel's fusion value. Otherwise, the pixel will be determined to be a potential object pixel, and the maximum among all saliency values of the pixel will be assigned to the pixel's fusion value. We term the fusion result as the min-max fusion (MMF) map, and the min-max statistical fusion is fusion defined as follows:

$$F_{MMF}(p) = \begin{cases} \min_{i=1,2,\dots,N} [S_i(p)], & \text{if } F_{AF}(p) \leq T \\ \max_{i=1,2,\dots,N} [S_i(p)], & \text{otherwise} \end{cases}, \quad (1)$$

where $S_i(p)$ denotes the saliency value of each pixel p in the i^{th} saliency map, and N is the total number of saliency maps. F_{MMF} and F_{AF} denote the MMF map and the AF map, respectively. For each saliency map S_i , the Otsu's method [12] is applied to obtain the threshold T_i , and the average of all N thresholds is assigned to the threshold T , i.e., $T = \sum_{i=1}^N T_i/N$. For the multiple saliency maps shown in Fig. 1(b), the corresponding MMF map is shown in Fig. 1(d).

2.2 Fusion Based on Fourier Transform of AF Map and MMF Map

Depending on the min-max operation, it is understandable that the MMF map owns the expected property of highlighting potential object pixels and suppressing potential background pixels, respectively, as intensively as possible. However, the MMF map will inevitably falsely highlight some background pixels and/or suppress some object pixels. Therefore, the AF map is exploited to alleviate such errors via the use of Fourier Transform (FT).

As we know, saliency map is in nature a grey-scale map, so its signal-noise ratio (SNR) is reflected in the amplitude spectrum and the boundaries between object regions and background regions are reflected in the phase spectrum. The two spectrums are obtained via FT on AF map and MMF map. The SNR of AF map is higher than that of MMF map since the AF map is the average of multiple saliency maps, while the MMF map contains more noises introduced by the min-max operation including wrongly judged object pixels and background pixels. On the other hand, the contrast between object pixels and background pixels in the MMF map are higher than that in the AF map, also due to the min-max operation for generating MMF map. Obviously, the higher the contrast between object pixels and background pixels is, the more complete boundaries between object regions and background regions will be preserved. Therefore, we choose the amplitude spectrum of AF map and the phase spectrum of MMF map to reform a new spectrum so as to obtain a better SNR-contrast trade-off than both AF map and MMF map.

The inverse FT (IFT) is performed on the reformed spectrum to obtain the final fusion map, which is abbreviated to AAPMF (Amplitude spectrum of AF map and Phase spectrum of MMF map based Fusion) map, as follows:

$$F_{AAPMF} = \text{IFT}[\mathbf{A}(F_{AF}), \mathbf{P}(F_{MMF})], \quad (2)$$

where \mathbf{A} refers to extracting the amplitude spectrum and \mathbf{P} refers to extracting the phase spectrum. The AAPMF map as shown in Fig. 1(e) is used as a prior for object segmentation.

To further verify the reasonableness and effectiveness of the proposed spectrum integration strategy, we also generate the fusion map by reforming the new spectrum with the amplitude spectrum of MMF map and the phase spectrum of AF map, and similarly, we denote it as AMPAF map. Some examples of AF, MMF, AAPMF and AMPAF maps are shown in Fig. 2. It can be seen from Fig. 2 that AAPMF maps can highlight object regions and suppress background regions better than the other maps, and the corresponding segmentation results using AAPMF maps also achieve the better quality than the other segmentation results. Therefore, the AAPMF map can serve as the better prior for object segmentation.

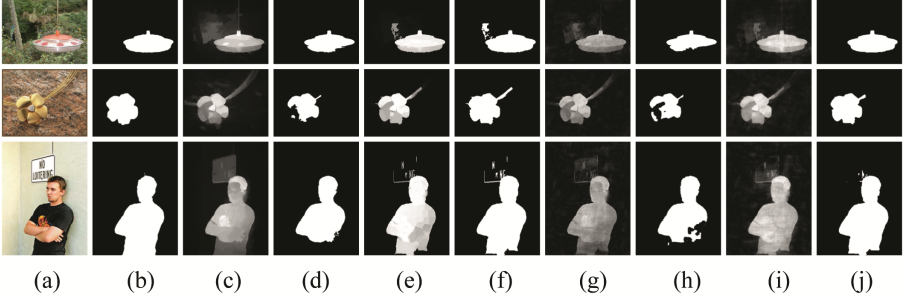


Fig. 2. The impact of amplitude spectrum and phase spectrum of saliency map. (a) Input image; (b) ground truth; (c) AF map; (d) segmentation result using (c); (e) MMF map; (f) segmentation result using (e); (g) AMPAF map; (h) segmentation result using (g); (i) AAPMF map; (j) segmentation result using (i).

2.3 Object Segmentation

Given the saliency prior such as AAPMF map, the object segmentation is formulated as assigning labels to each pixel by solving an energy minimization problem under the framework of graph cut [13]. As a result, each pixel p gets its label $L_p \in \{0, 1\}$, where $L_p = 1$ denotes object and $L_p = 0$ denotes background. The energy function is defined as follows:

$$\begin{aligned}
 E &= \sum_p D(L_p) + \lambda \sum_{\substack{(p,q) \in \Omega \\ L_p \neq L_q}} \theta(L_p, L_q) \\
 D(L_p) &= \begin{cases} S_F(p), & L_p = 0 \\ 1 - S_F(p), & L_p = 1 \end{cases} \\
 \theta(L_p, L_q) &= \exp \left[-\frac{(I_p - I_q)^2}{2\sigma^2} \right] \cdot \frac{1}{\text{dist}(p, q)},
 \end{aligned} \tag{3}$$

where $D(\cdot)$ is the data term, $\theta(\cdot)$ is the smoothness term, and Ω is the set of pairs of neighboring pixels. The parameter λ is used to balance the two terms, and is set to 0.1 for a moderate effect of smoothness. S_F denotes the saliency prior, which uses the AAPMF map here. In the smoothness term, I_p denotes the color feature of the pixel p , the parameter σ^2 is set to 2.5, and $disp(p, q)$ is the Euclidean distance between a pair of pixels, p and q . The max-flow algorithm [14] is adopted to minimize the energy function and obtain the labels of pixels, which represent the object segmentation result. For example, Fig. 1(f) is the object segmentation result by using AAPMF map as the saliency prior.

3 Experimental Results

3.1 Experimental Setting

To verify the effectiveness of the proposed saliency fusion method for object segmentation, we evaluated its performance on the three public benchmark datasets including MSRA10 K [15], ECSSD [16] and PASCAL-S [17], with 10000, 1000 and 850 images, respectively. For each image in the three datasets, the manually annotated pixel-level binary ground truth of objects is provided. According to the benchmark [3], we selected the top six saliency models with the highest performances, i.e., DRFI [18], QCUT [19], RBD [20], ST [21], DSR [22] and MC [23], to generate saliency maps. In addition to AAPMF map, we also tested AF map, MMF map and AMPAF map, and compared with the other two fusion methods, i.e. maximum and multiplication, which generate MaxF map using the pixel-wise maximum of all saliency values as the fusion value and PF map using the pixel-wise multiplication of all saliency values as the fusion value, respectively. We specified S_F in Eq. (3) with each of the above mentioned saliency maps and fusion maps to obtain the corresponding object segmentation results.

3.2 Quantitative Comparison

We evaluated all segmentation results using the conventional F-measure defined as follows:

$$F_\beta = \frac{(1 + \beta^2) Precision \times Recall}{\beta^2 Precision + Recall}, \quad (4)$$

and the weighted F-measure, which is recently introduced in [24], as follows:

$$F_\beta^\omega = \frac{(1 + \beta^2) Precision^\omega \times Recall^\omega}{\beta^2 Precision^\omega + Recall^\omega}, \quad (5)$$

where $Precision^\omega$ and $Recall^\omega$ (namely weighted *Precision* and weighted *Recall*) are computed by the extended basic quantities including true positive, true negative, false positive and false negative, which are weighted according to the pixels' location and

neighborhood. The coefficient β^2 is set to 0.3 indicating more importance of precision than recall as suggested in [3]. Here we compute all measures for each image and then obtain the average on all images in a given dataset for performance comparison. All results are listed in Table 1. In each row of Table 1, the 1st, 2nd and 3rd place of performance are marked with red, green and blue, respectively. It can be observed that in terms of F-measure and weighted F-measure, AAPMF map consistently performs best on all datasets. This objectively reveals the overall better performance of AAPMF map as a prior for object segmentation. Particularly, the advantage of AAPMF map over AF map, MMF map and AMPAF map further demonstrates the reasonableness of reforming the spectrum by combining the amplitude spectrum of AF map and the phase spectrum of MMF map.

Table 1. Average F-measure (F) and average F_β^ω -measure (F_β^ω) of object segmentation results on three public benchmark datasets (D) consisting of MSRA10 K (M), ECSSD (E) and PASCAL-S (P).

D	Metric	Top Six Saliency Models						Saliency Fusion Methods					AAPMF (proposed)
		DRFI	DSR	MC	QCUT	RBD	ST	MaxF	PF	AF	MMF	AMPAF	
M	F	.856	.784	.784	.728	.816	.844	.836	.505	.861	.874	.845	.880
	F_β^ω	.839	.777	.774	.735	.808	.820	.783	.525	.856	.846	.840	.861
E	F	.740	.651	.647	.604	.645	.708	.715	.356	.734	.756	.709	.760
	F_β^ω	.715	.638	.621	.604	.619	.667	.642	.377	.717	.711	.694	.723
P	F	.617	.541	.553	.511	.580	.600	.628	.279	.629	.651	.597	.652
	F_β^ω	.583	.517	.525	.503	.544	.555	.552	.292	.601	.598	.575	.606

*For Table 1, in the first column, D denotes Datasets, M, E and P denote MSRA10K, ECSSD and PASCAL-S dataset, respectively; in the second column, F and F_β^ω denotes F-measure and F_β^ω -measure, respectively.

3.3 Qualitative Comparison

Some object segmentation results are shown in Fig. 3 for a qualitative comparison. Overall, the segmentation results with AAPMF maps show the best visual quality compared to others. Besides, it can be seen from Fig. 3(c)–(h) that the segmentation results with the saliency maps may miss some object regions and/or contain some background regions. The results with MaxF maps shown in Fig. 3(i) usually introduce some irrelevant regions, while the results with PF maps shown in Fig. 3(j) usually miss some portions of object regions. Compared to the results with AF, MMF and AMPAF maps shown in Fig. 3(k)–(m), the results with AAPMF maps shown in Fig. 3(n) can generally segment more complete objects with more accurate boundaries.

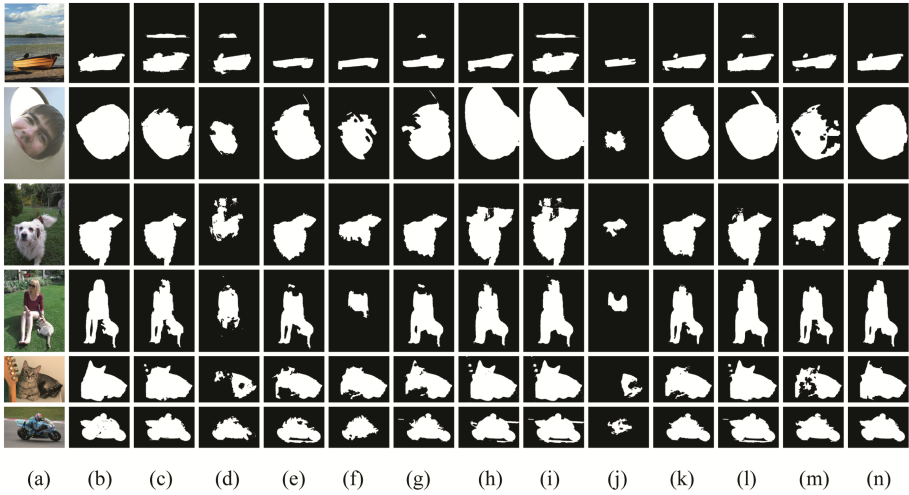


Fig. 3. Object segmentation results of sample images from MSRA10 K (the first two rows), ECSSD (the middle two rows) and PASCAL-S (the bottom two rows). (a) original images; (b) ground truths; (c)–(h) segmentation results with DRFI, DSR, MC, QCUT, RBD and ST; (i)–(j) segmentation results with MaxF and PF; (k)–(m) segmentation results with AF, MMF and AMPAF; (n) segmentation results with AAPMF (our results).

3.4 Computation Cost

Our method is implemented using Matlab on a PC with an Intel Core i7 4.0 GHz CPU and 16 GB RAM. The average processing time for an image with a resolution of 400×300 is 0.88 s excluding the generation of saliency map. The first-stage fusion takes 0.51 s, the second-stage fusion takes 0.06 s, and the object segmentation takes 0.31 s. It can be seen that the two-stage saliency fusion is computationally efficient. The other saliency fusion methods are also computationally efficient. Specifically, MaxF, PF, AF, MMF and AMPAF take 0.02 s, 0.03 s, 0.03 s, 0.48 s and 0.58 s, respectively, to generate saliency fusion maps. The object segmentation based on these saliency fusion maps also takes about 0.31 s.

3.5 Discussion

Since the saliency fusion heavily depends on the individual saliency maps involved in the fusion, the proposed method will fail once all the individual saliency maps are insufficient to highlight salient objects. Some failure examples from the PASCAL-S dataset, which is a more challenging dataset for object segmentation, are shown in Fig. 4. Nonetheless, the results reported in Table 1 indicate that even in the PASCAL-S dataset with more challenging images, the segmentation performance using AAPMF map as a prior is still better than the results using individual saliency maps, MaxF, PF, AF, MMF and AMPAF maps as a prior.

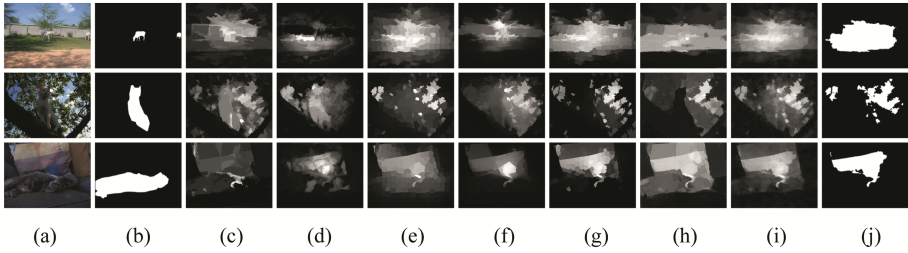


Fig. 4. Failure examples of the proposed method. (a) Input images; (b) ground truths; (c)–(h) saliency maps generated by the top six models: DRFI, DSR, MC, QCUT, RBD and ST; (i) AAPMF maps generated by the proposed fusion method; (j) our segmentation results by using AAPMF map as a prior.

4 Conclusion

This paper proposes a novel approach to fuse multiple saliency maps in two stages for object segmentation. In the first stage, the AF map and MMF map are obtained based on the average and min-max statistics, respectively. In the second stage, the amplitude spectrum of AF map and the phase spectrum of MMF map are integrated to generate the AAPMF map via the use of FT and IFT. Experimental results demonstrate that the two-stage saliency fusion as a prior actually boosts the performance of object segmentation.

Acknowledgements. This work was supported by the National Natural Science Foundation of China under Grant No. 61471230, Shanghai Municipal Natural Science Foundation under Grant No. 16ZR1411100, and the Program for Professor of Special Appointment (Eastern Scholar) at Shanghai Institutions of Higher Learning.

References

1. Hu, S.M., Chen, T., Xu, K., Cheng, M.M., Martin, R.R.: Internet visual media processing: a survey with graphics and vision applications. *Vis. Comput.* **29**(5), 393–405 (2013)
2. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* **20**(11), 1254–1259 (1998)
3. Borji, A., Cheng, M.M., Jiang, H., Li, J.: Salient object detection: a benchmark. *IEEE Trans. Image Process.* **24**(12), 5706–5722 (2015)
4. Zhou, X., Liu, Z., Sun, G., Ye, L., Wang, X.: Improving saliency detection via multiple kernel boosting and adaptive fusion. *IEEE Sig. Process. Lett.* **23**(4), 517–521 (2016)
5. Xie, Y., Lu, H., Yang, M.H.: Bayesian saliency via low and mid-level cues. *IEEE Trans. Image Process.* **22**(5), 1689–1698 (2013)
6. Han, J., Zhang, D., Hu, X., Guo, L., Ren, J., Wu, F.: Background prior based salient object detection via deep reconstruction residual. *IEEE Trans. Circuits Syst. Video Technol.* **25**(8), 1309–1321 (2015)
7. Zhao, T., Li, L., Ding, X., Huang, Y., Zeng, D.: Saliency detection with spaces of background-based distribution. *IEEE Sig. Process. Lett.* **23**(5), 683–687 (2016)

8. Borji, A., Sihite, D.N., Itti, L.: Salient object detection: a benchmark. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012. LNCS, pp. 414–429. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33709-3_30
9. Mai, L., Niu, Y., Liu, F.: Saliency aggregation: a data-driven approach. In: 26th IEEE Conference on Computer Vision and Pattern Recognition, pp. 1131–1138. IEEE Press, Portland (2013)
10. Le Meur, O., Liu, Z.: Saliency aggregation: does unity make strength? In: Cremers, D., Reid, I., Saito, H., Yang, M.-H. (eds.) ACCV 2014. LNCS, vol. 9006, pp. 18–32. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-16817-3_2
11. Zhang, C.Q., Tao, Z.Q., Wei, X.X., Cao, X.C.: A flexible framework of adaptive method selection for image saliency. *Pattern Recognit. Lett.* **63**(1), 66–70 (2015)
12. Otsu, N.: A threshold selection method from gray-level histograms. *IEEE Trans. Syst. Man Cybern.* **9**(1), 62–66 (1979)
13. Boykov, Y., Jolly, M.: Interactive graph cuts for optimal boundary and region segmentation of objects in n-d images. In: 8th IEEE International Conference on Computer Vision, pp. 105–112. IEEE Press, Vancouver (2001)
14. Boykov, Y., Kolmogorov, V.: An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *IEEE Trans. Pattern Anal. Mach. Intell.* **26**(9), 1124–1137 (2004)
15. THUR15000. <http://mmcheng.net/gsal/>
16. Yan, Q., Xu, L., Shi, J., Jia, J.: Hierarchical image saliency detection on extended CSSD. *IEEE Trans. Pattern Anal. Mach. Intell.* **38**(4), 717–729 (2015)
17. Li, Y., Hou, X., Koch, C., Rehg, J.M., Yuille, A.L.: The secrets of salient object segmentation. In: 27th IEEE Conference on Computer Vision and Pattern Recognition, pp. 280–287. IEEE Press, Columbus (2014)
18. Jiang, H., Wang, J., Yuan, Z., Wu, Y., Zheng, N., Li, S.: Salient object detection: a discriminative regional feature integration approach. In: 26th IEEE Conference on Computer Vision and Pattern Recognition, pp. 2083–2090. IEEE Press, Portland (2013)
19. Aytekin, C., Kiranyaz, S., Gabbouj, M.: Automatic object segmentation by quantum cuts. In: 22th IEEE International Conference on Pattern Recognition, pp. 112–117. IEEE Press, Stockholm (2014)
20. Zhu, W., Liang, S., Wei, Y., Sun, J.: Saliency optimization from robust background detection. In: 27th IEEE Conference on Computer Vision and Pattern Recognition, pp. 2814–2821. IEEE Press, Columbus (2014)
21. Liu, Z., Zou, W., Le Meur, O.: Saliency tree: a novel saliency detection framework. *IEEE Trans. Image Process.* **23**(5), 1937–1952 (2014)
22. Li, X., Lu, H., Zhang, L., Ruan, X., Yang, M.H.: Saliency detection via dense and sparse reconstruction. In: 14th IEEE International Conference on Computer Vision, pp. 1937–1952. IEEE Press, Sydney (2013)
23. Jiang, B., Zhang, L., Lu, H., Yang, C., Yang, M.H.: Saliency detection via absorbing Markov chain. In: 14th IEEE International Conference on Computer Vision, pp. 2979–2983. IEEE Press, Sydney (2013)
24. Margolin, R., Zelink-Manor, L., Tal, A.: How to evaluate foreground maps? In: 27th IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255. IEEE Press, Columbus (2014)