SALIENT OBJECT DETECTION VIA DOUBLE SPARSE REPRESENTATIONS UNDER VISUAL ATTENTION GUIDANCE

Xiang Wang, Yongjun Zhang, Xunwei Xie, and Yansheng Li*

School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, China

ABSTRACT

This paper introduces a novel method for salient object detection from the perspective of sparse representation under visual attention guidance. After pretreatment and regional analysis with eye fixation detection and multi scale segmentation, regions that are used to make up the foreground and background dictionaries are respectively selected by sorting the visual attraction level of all image regions. For saliency measurement, the reconstruction errors instead of common local and global contrasts are used as the saliency indicator, which is expected to improve the object integrity. In addition, the multi scale workflow is conductive to enhance the robustness for objects of different sizes. The proposed method was compared to six state-of-the-art saliency detection methods using three benchmark datasets, and it was confirmed to have more favorable performance in the detection of multiple objects as well as maintaining the integrity of the object area.

Index Terms—Salient object detection, visual attention guidance, sparse representation, reconstruction error

1. INTRODUCTION

Visual saliency is an important and fundamental research in computer vision and image interpretation that is concerned with the most visually noticeable foreground in a scene [1]. Since its first computational model in 1998 [2], a great number of saliency methods which can be generally categorized as either bottom-up methods [3, 4] or top-down methods [5, 6] have been developed.

Since Koch et al. [7] set up the foundation of visual saliency and Itti et al. [2] proposed a local color contrast method based on the contrast constraints, many related methods have been introduced, such as the graph-based visual saliency method (GBVS) [8], the Markov chain absorbed method (MC) [9], and some newer methods [10, 11]. Of late, global contrast-based methods have attracted much interest because of the common insufficient integrity of the local contrast-based methods [4]. However, as the still need for contrast comparison, the drawbacks of global contrast-based multiple objects detection methods continue to be recognized. In addition, whether for local or global contrast-based methods, an appropriate salient measure is always a crucial factor. The increasing research and experimental results gradually confirmed that the detection easily fails when salient objects touch the image boundaries if the saliency is evaluated according to the center prior or boundary prior [12]. In general, the limitations of previous salient object detection methods can be summarized as follows [13]: 1) local contrast methods tend to highlight the most distinct part of the object, while they are unable to uniformly evaluate the saliency level of the entire object area; 2) global contrast methods continue to be not effective enough in comparing different contrast values for detection of multiple objects, especially for those with large dissimilarity; and 3) boundary prior-based saliency computation may fail if the salient object touches the image boundaries, and it is unclear how to integrate the boundary prior well with other saliency measures.

In order to improve the issues, more robust boundary prior strategies have been proposed to enhance the reliability of saliency computation [14, 15]. Moreover, there are several methods based on sparse representation which make full use of the difference between the background and foreground [13, 16].

This paper proposes a new method via double sparse representation under visual attention guidance (RSRVAG), which combines the background and foreground based reconstructions and treats the reconstruction errors as the saliency indicator, aiming to avoid the integrity shortcomings of contrast-based methods and the weak robustness of boundary prior-based methods. The proposed RSRVAG method improves the DSR [13] method with the following major differences and contributions.

1) Both background and foreground based sparse representations are combined to enhance the stability of sparse representation.

2) The traditional eye fixation results [2] are introduced to extract the background and foreground dictionaries, by which, the RSRVAG method is designed to be more robust than the DSR which is based on boundary prior, especially for images with salient objects touching the boundaries.

2. METHODOLOGY

It is known that the reconstruction errors of sparse representation can effectively indicate the similarity between samples and dictionary [14]. Thus, the salient value of the regions can be determined by the representation processing based on background dictionary or foreground dictionary.

The framework of the proposed RSRVAG method is shown in Fig. 1, where simple linear iterative clustering (SLIC) algorithm [17] is used to generate superpixel regions for better capturing structural information and decreasing the whole processing units. Note that only one scale with 300 superpixels is represented in the framework. For regions at each scale, initial saliency map generated by a traditional visual attention method is used as guidance for dictionary extraction. Following the double sparse representations that are respectively based on background dictionary and foreground dictionary, all image regions' saliency value (saliency map) are calculated by combining the two groups of reconstruction errors. Then, the multi scale results are integrated to generate a final saliency map.

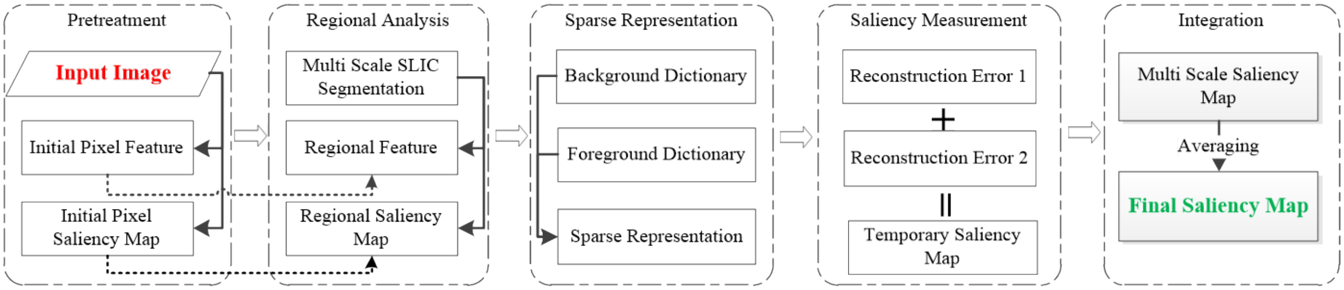


Fig. 1. Framework of the proposed method.

2.1. Dictionary Selection

As illustrated in Fig. 2, the saliency of image pixels can be roughly obtained by traditional visual attention methods like IT [2] and GBVS [8]. Although only soft edges exist in the results, the area of salient objects still can be highlighted to some extent. Under the guidance of visual attention, regions with low visual attraction are selected to fulfill the background dictionary, while the opposites build the foreground dictionary.

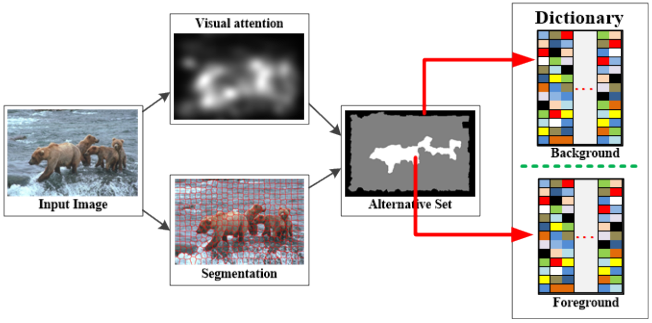


Fig. 2. Illustration of dictionary selection.

The dictionary selection process is described as follows.

1) Calculating the initial regional saliency map ISM by averaging the saliency of pixels within the regions. Giving the number of regions as N.

2) Setting a proportionality p_f, and taking the first $p \times N$ larger elements as foreground dictionary D_f.

3) Setting a proportionality p_b, and taking the first $p \times N$ smaller elements as background dictionary D_b.

2.2. Sparse Representation and Saliency Measurement

Each region is represented by a feature vector as $F = \{mR, mG, mB, mL, ma, mb, mx, my, mf_x, mf_y, mf_{xx}, mf_{yy}, mf_{xy}\}$ consisting of color, spatial, and geometry information that are widely used in saliency detection, where mX is the mean X feature value of pixels in the region, including RGB and Lab colors R, G, B, L, a, b, pixel coordinates x, y, and the first and second gradients $f_x, f_y, f_{xx}, f_{yy}, f_{xy}$.

The entire segmented image is represented as $I = \{F_1, F_2, \dots, F_N\} \in R^{D\times N}$, where N is the number of regions, D is the feature dimension, and all the feature values are normalized to (0, 1). Given foreground dictionary D_f and background dictionary D_b, region i is encoded by Eq. 1 and 2, and the reconstruction errors are calculated by Eq. 3 and 4.

$$\alpha_{bi} = \arg\min \left\| F_i - D_b \alpha_{bi} \right\|_2^2 + \lambda_b \left\| \alpha_{bi} \right\|_1 \tag{1}$$

$$\alpha_{fi} = \arg\min \left\| F_i - D_f \alpha_{fi} \right\|_2^2 + \lambda_f \left\| \alpha_{fi} \right\|_1 \tag{2}$$

$$\varepsilon_{bi} = \left\| F_i - D_b \alpha_{bi} \right\|_2^2 \tag{3}$$

$$\varepsilon_{fi} = \left\| F_i - D_f \alpha_{fi} \right\|_2^2 \tag{4}$$

where α_{bi}, α_{fi} is the sparse code vector obtained by dictionaries D_b and D_f ; λ_b, λ_f are the regularization parameters that are empirically set to 0.01 in the experiment; $\varepsilon_{bi}, \varepsilon_{fi}$ are the reconstruction errors as a result of the background and foreground sparse representation.

As described in DSR, the errors caused by background-based sparse representation can directly measure the saliency, so the foreground-based reconstruction errors can work in the opposite way, based on which, the region saliency value is simply measured according to Eq. 5 in this paper.

$$Sal_i = \varepsilon_{bi} / (\varepsilon_{fi} + \sigma^2) \tag{5}$$

where Sal_i is the saliency value of region i, σ^2 is a regulatory factor that is set to 0.1 in the experiment.

2.3. Saliency Map Integration

To enhance the robustness to objects of different sizes, superpixels at different scales were generated by setting multiple number parameters of the SLIC algorithm empirically as 100, 200, 300, and 400 in the proposed RSRVAG method, which were commonly used in many previous studies. After the multiple scale saliency measurements, FSM_i $(i = 1, 2, 3, 4)$ was obtained and the final saliency map was integrated by averaging by averaging the multiscale results. The integration is illustrated in Fig. 3.

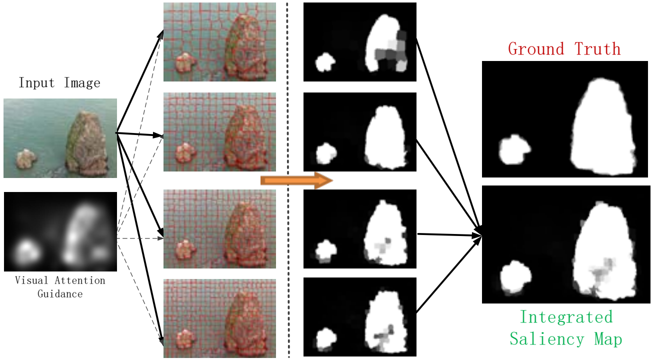


Fig. 3. Integration of multiscale results.

3. EXPERIMENTS AND ANALYSIS

With the parameters in Table 1, the performance of the proposed RSRVAG method was evaluated on three benchmark datasets (described in Table 2) with comparison to six state-of-the-art saliency detection algorithms: IT [2], HC [4], MC [9], RBD [14], RRFC [15], and DSR [13]. The evaluation measures were selected as $Precision\text{-}Recall$ (PR) ↑ curve, $F\text{-}Measure$ ↑ curve, and the mean absolute error (MAE ↓) which were fully analyzed in benchmark [12]. The up-arrow ↑after a measure indicates that the larger the value achieved, the better the performance; while the down-arrow ↓ means the smaller, the better.

Table 1. Key parameters of the proposed RSRVAG method

Parameter	Value
MultiScales	100,200,300, 400
P_b, P_f	0.2
λ_b, λ_f	0.01
σ^2	0.1
β^2	0.3

Table 2. Datasets list

Dataset	Description
MRSA-ASD [4]	Single object images
SED2 [3]	Images contain two objects
ECSSD [6]	Structurally complex images

As visually displayed in Fig. 4, it is apparent that for the specific datasets in the experiments, the proposed RSRVAG successfully extracted accurate entire salient objects, regardless of whether they were single objects, multiple objects, or images with complex structures.

For single objects, the eye fixation (IT) and the traditional contrast-based method (HC) only produced fuzzy contours, but the improved methods (MC, RBD, RRFC, and DSR) clearly obtained better results. However, the RSRVAG realized some obviously further improvements. As far as multiple objects, the RSRVAG worked well while only a few of the state-of-the-art methods could deal with objects with large dissimilarities and saliency differences. In addition, for images with complex structure or similarity between the background and the foreground, the results show that the RSRVAG exhibited greater detection capability.

From the quantitative point of view, as shown in Fig.5, the proposed RSRVAG method performed quite competitively in terms of the PR curve and F^{β} with the latest improved methods, such as RBD, RRFC, and DSR, which are considered to be the outstanding performers in the six state-of-the-art methods utilized in the experiments. At the same time, the three measures of the proposed RSRVAG method were more balanced and stable, while the state-of-the-art methods always had at least one measure value that was relatively low. In the case of MAE, the proposed RSRVAG method performed obviously better in the comparisons regardless of whether it was a single object, multiple objects, or images that were structurally complex.

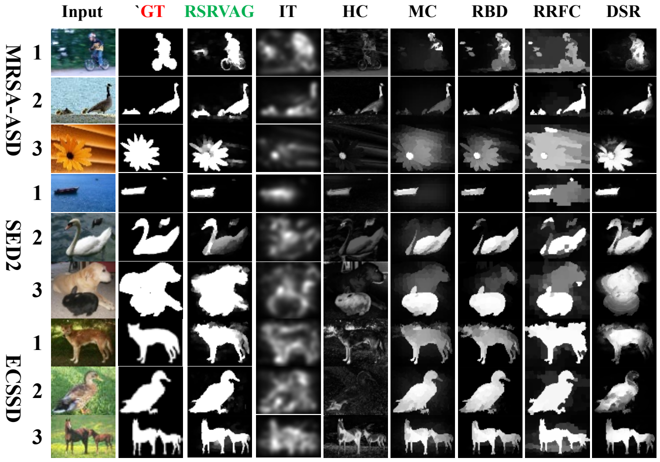
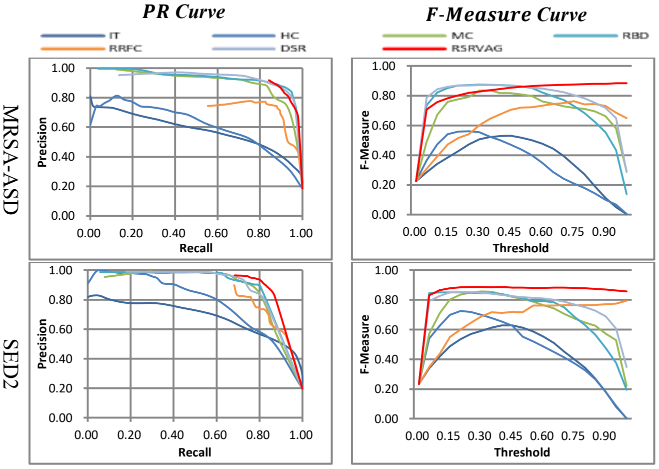


Fig. 4. Visual comparison on MRSA-ASD, SED2, and ECSSD datasets.



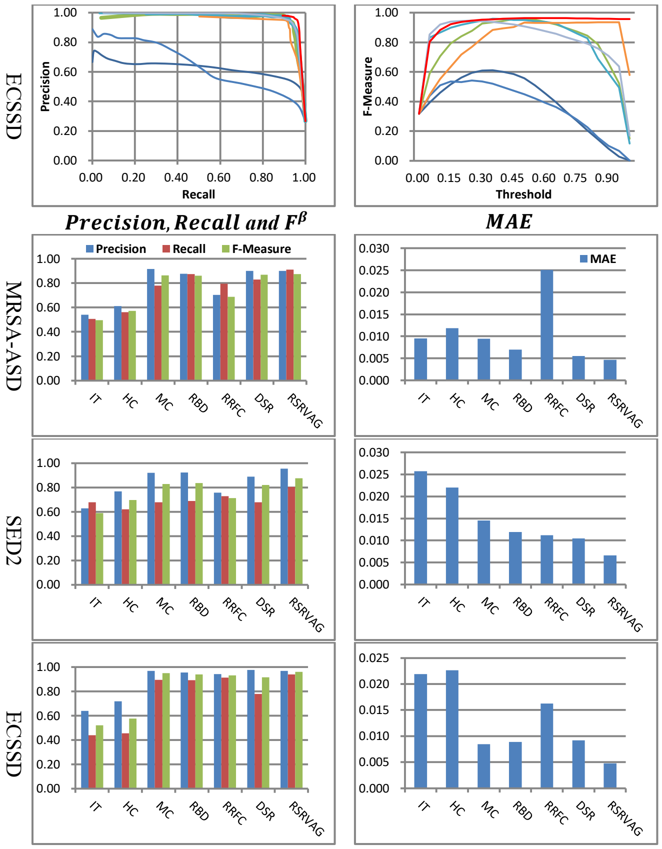


Fig. 5. Quantitative comparison results on MSRA-ASD, SED2, and ECSSD datasets.

4. CONCLUSION

This paper proposes a novel visual attention guided salient object detection method via double sparse representations based on background and foreground dictionaries. The reconstruction errors, which can effectively reflect the similarity between the target samples and the dictionaries, are used as the salient indicator; and a multi scale operation acts as an improvement for object details in different sizes.

The experimental results show that the proposed RSRVAG method performed better in the comparison with some state-of-the-art methods. RSRVAG also was confirmed to be capable of working more effectively and efficiently on images with complex structures and detections of multiple objects, which was reflected by its ability to extract an integrated and uniform salient object area. In terms of the limitations, the salient results obtained by sparse representation strictly rely on the dictionaries that were built simply based on the sorting of eye fixation results, which is to say that the detection may fail when the visual attention guidance is poor; and while the integration of double sparse representation and multiscales is slightly simple and rough that may cause some under-detections or over-detections, thus future work should consider developing it by trying to weaken the dependence on initial visual attention guidance and enhance the integration strategies.

ACKNOWLEDGMENTS

This research was partially supported by the National Natural Science Foundation of China under Grants 41322010 and 41571434.

REFERENCES

[1] A. Borji and L. Itti, “State-of-the-art in visual attention modeling,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 185-207, Jan. 2013.

[2] L. Itti, C. Koch, and E. Niebur, “A model of saliency-based visual attention for rapid scene analysis,” *IEEE Trans. Patt. Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254-1259, 1998.

[3] S. Alpert, M. Galun, R. Basri, A. Brant, “Image segmentation by probabilistic bottom-up aggregation and cue integration,” in *Proc. IEEE Conf. CVPR*, pp. 1-8, Jun. 2007.

[4] M. M. Cheng, N. J. Mitra, X. Huang, et al., “Global contrast based salient region detection,” *IEEE Trans. Patt. Anal. Mach. Intell.*, vol. 37, no. 3, pp. 569-582, 2015.

[5] A. Borji, “Boosting bottom-up and top-down visual features for saliency estimation,” in *Proc. IEEE Conf. CVPR*, pp. 438-445, Jun. 2012.

[6] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, and H. Shum, “Learning to detect a salient object,” *IEEE Trans. Patt. Anal. Mach. Intell.*, vol. 33, no. 2, pp. 353-367, 2011.

[7] C. Koch and S. Ullman, “Shifts in selective visual attention: towards the underlying neural circuitry,” *Human Neurobiol.*, vol. 4, no. 4, pp. 219-227, 1985.

[8] J. Harel, C. Koch, and P. Perona, “Graph-based visual saliency,” in *Proc. Conf. Adv. Neural Inf. Process. Syst. NIPS*, pp. 545-552, 2006.

[9] B. Jiang, L. Zhang, H. Lu, et al., “Saliency detection via absorbing markov chain”, in *Proc. IEEE ICCV*, pp. 1665-1672, Dec. 2013.

[10] J. Chen, B. Ma, H. Cao, et al., “Updating initial labels from spectral graph by manifold regularization for saliency Detection,” *Neurocomputing*, 2017.

[11] J. Zhang, K. A. Ehinger, H. Wei, et al., “A novel graph-based optimization framework for salient object detection,” *Pattern Recognition*, vol. 64, pp. 39-50, 2017.

[12] A. Borji, M. M. Cheng, H. Jiang, et al. “Salient Object Detection: A Benchmark,” *IEEE Trans. Image. Proc.*, vol. 24, no. 12, pp. 5706-5722, 2015.

[13] X. Li, H. Lu, L. Zhang, et al., “Saliency detection via dense and sparse reconstruction,” in *Proc. IEEE ICCV*, pp. 2976-2983, Dec. 2013.

[14] W. Zhu, S. Liang, Y. Wei, et al., “Saliency optimization from robust background detection,” in *Proc. IEEE Conf. CVPR*, Jun. 2014, pp. 2814-2821.

[15] K.Ohm M. Lee, and Y. Lee, “Salient object detection using recursive regional feature clustering,” *Information Sciences*, vol. 387, pp. 1-18, 2017

[16] L. Zhang, X. Lv, and X. Liang, “Saliency analysis via hyperparameter sparse representation and energy distribution optimization for remote sensing images,” *Remote Sensing*, vol. 9, no. 6, pp. 636, 2017.

[17] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk, “SLIC superpixels,” Tech. Rep., 2010.