SPECTRAL SALIENT OBJECT DETECTION

Keren Fu[A,B]   *Chen Gong[A]*   *Irene Y.H. Gu[B]*   *Jie Yang[A*]*   *Xiangjian He[C]*

[A] Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, China
[B] Department of Signals and Systems, Chalmers University of Technology, Gothenburg, Sweden
[C] Faculty of Engineering and Information Technology, University of Technology, Sydney, Australia

ABSTRACT

Many existing methods for salient object detection are performed by over-segmenting images into non-overlapping regions, which facilitate local/global color statistics for saliency computation. In this paper, we propose a new approach: *spectral salient object detection*, which is benefited from selected attributes of *normalized cut*, enabling better retaining of holistic salient objects as comparing to conventionally employed pre-segmentation techniques. The proposed saliency detection method recursively bi-partitions regions that render the lowest cut cost in each iteration, resulting in binary spanning tree structure. Each segmented region is then evaluated under criterion that fit Gestalt laws and statistical prior. Final result is obtained by integrating multiple intermediate saliency maps. Experimental results on three benchmark datasets demonstrate the effectiveness of the proposed method against 13 state-of-the-art approaches to salient object detection.

Index Terms— Salient object detection, Normalized cut, Pre-segmentation, Partition, Gestalt laws

1. INTRODUCTION

Saliency detection plays an important role in understanding the mechanism of human visual attention. Recently, identifying salient regions in a digital image has attracted lots of researchers. This is due to their advantages against traditional eye fixation models [1, 2, 3, 4] on applications, e.g. content-based image cropping [5], thumbnailing [6], resizing and re-targeting [7, 8]. For such applications, salient object detection has evolved to uniformly highlight an object.

Previous solutions that aim at salient region detection often model bottom-up process using local or global color statistics, e.g. color rarity or uniqueness. Zhai *et al* [9] introduce image histograms which only model luminance channel to calculate saliency. Achanta *et al* [10] provide a saliency approximation by subtracting the average color from low-pass filtered results. Goferman *et al* [8] combine local and global features to estimate the patch saliency in multi-scale fashion. Cheng *et al* [11] extend the method in [9] and incorporate histogram-based color contrast. Perazzi *et al* [12] propose *saliency filter*, which formulates complete contrast and saliency estimation using high dimensional Gaussian filters. Fu *et al* [13] integrate color contrast and color distribution for saliency estimation. Shen *et al* [14] solve saliency detection issue as a low rank matrix recovery problem. Margolin *et al* [15] define patch distinctness as L1-norm in PCA coordinates and combine with color distinctness. Cheng *et al* [16] measure saliency by hierarchical soft abstraction. Jiang *et al* [17] integrate regional features by using random forest regressor.

Since contrast priors could be difficult to distinguish among similar colors occur in foreground and background, other methods employ background smoothness priors that consider image boundary as background, e.g. the geodesic measure [18] and manifold ranking based technique [19]. Besides, some other cues are employed, e.g. the *Harris convex hull* [20, 21].

Observing the above, one recent trend for salient object detection is to over-segment an input image into non-overlapping regions. Techniques like Mean-shift [22] (used in [10, 13, 14]), graph-based segmentation [23] (used in [11]) or SLIC superpixel [24] (used in [12, 21, 20, 19]) are widely employed. These techniques are known to be useful since they are able to eliminate some background noise and reduce computation cost of saliency detection by treating each segment as a processing unit. However, these pre-segmentation methods only take local color similarity into account, and cluster pixels towards regions in a fine-to-coarse fashion. Each unit is only local and can hardly reflect any holism information. By directly computing saliency with regions generated by [22, 23, 24], it becomes difficult to uniformly highlight objects as different object parts may have different contrast property. It often happens that detection results only concentrate on several highly discriminative parts but ignore the objects' holism. This problem somewhat can be alleviated by achieving a balanced pre-segmentation, which divides a desired object into one single region. Unfortunately, it is hard to generically control just right parameters, e.g. color bandwidth in [22], k in [23] to achieve an ideal segmentation that trades off between over- and under-segmenting desired objects.

Since segmentation is proved conducive to high quality saliency detection, it is crucial to choose an appropriate segmentation method. In this paper, we provide some new insights into this issue. Our work is inspired by *spectral segmentation* [25, 26]. These methods analyze the smallest eigenvectors of image's graph Laplacian matrix in order to obtain an unsupervised decomposition of an image into a collection of hard segments. We consider that graph-based normalized cut (N-cut) [25] could be better in providing good segmentation hypotheses for salient object detection. N-cut algorithm [25] aims at partitioning the graph into two parts whereas salient object detection could also be deemed as a binary classification problem (pixels are classified as "salient" and "non-salient"). We show that if N-cut segmentation [25] is moderately applied, it will drastically assist and facilitate saliency detection. We propose a salient object detection technique (namely *spectral salient object detection*, *SS* for short) that is based on a novel hierarchical spectral partitioning framework. The proposed partitioning framework benefits the attribute of N-cut while attempts to keep the holism of a salient object during segmentation.

Although there are other saliency detection approaches (such as [2]) that involve the notion of "spectral", most of them aim at modeling human eye fixation rather than detect salient regions. A main discrepancy between [2] and the proposed method is that, the former one models saliency from the spectrum of the Fourier transform, whereas the latter one is based on spectral segmentation techniques (e.g. [25, 26]).

2. NORMALIZED CUT: REVIEW

Normalized cut (or N-cut) proposed by Shi *et al* [25] normalizes the graph cut cost by using the total edge connections towards all nodes in a graph. For a given graph $G = (V, E)$ (V for vertices and E for edges), N-cut aims to find a cut that partitions V into two parts A and B such that the following can be minimized:
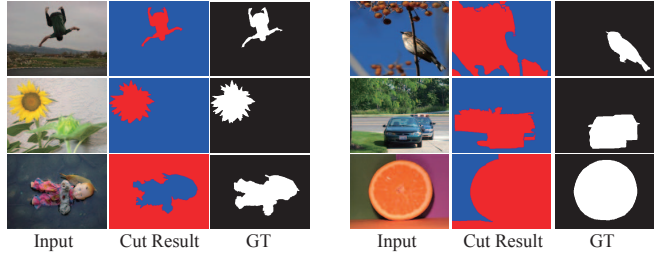
$$Ncut(A, B) = \frac{cut(A, B)}{assoc(A, V)} + \frac{cut(A, B)}{assoc(B, V)} \quad (1)$$

where $cut(A, B) = \sum_{u \in A, v \in B} e(u, v)$ is the sum of edge connection between A and B. Association $assoc(A, V) = \sum_{u \in A, t \in V} e(u, t)$ is the connection sum between A and all nodes. $assoc(B, V)$ is similarly defined. Shi *et al* [25] proved that minimizing $Ncut(A, B)$ is equivalent to minimizing energy E:

$$E = \min_y \frac{y^T (D - W) y}{y^T D y} \quad (2)$$

subject to the constraint that $y_i \in \{1, -b\}$ and $y^T D 1 = 0$ (where 1 is the vector of all ones). The continuous solution to the above problem is known as the second smallest eigenvector of the generalized eigen-system:

$$(D - W) y = \lambda D y \quad (3)$$

Fig. 1. Examples from MSRA-1000 [10] for the single N-cut segmentation. Left: successful cases. Right: failure cases. Different colors indicate different segments. GT represents ground truth mask.

In (3), W is the graph affinity matrix and D is the diagonal degree matrix s.t. $D_{ii} = \sum_j W_{ij}$. For more details, readers are referred to [25].

Since N-cut can partition a graph into two balanced parts (may be potential foreground and background) while minimize the corresponding normalized cut cost, it could be better exploited for pre-segmentation in salient object detection, which as aforementioned may be treated as a binary classification problem. Here note N-cut has been exploited for unsupervised foreground segmentation in a constrained parametric manner [27]. However in our proposed method, such graph partition is explored for a different purpose: the salient object detection. Key differences to conventional purpose in segmentation lie in:

- Fast and simple to use as a pre-processing mean for other computer vision tasks.
- The outcome is a gray-level saliency map rather than a binary hard segmentation.

Thus saliency detection enjoys much flexibility.
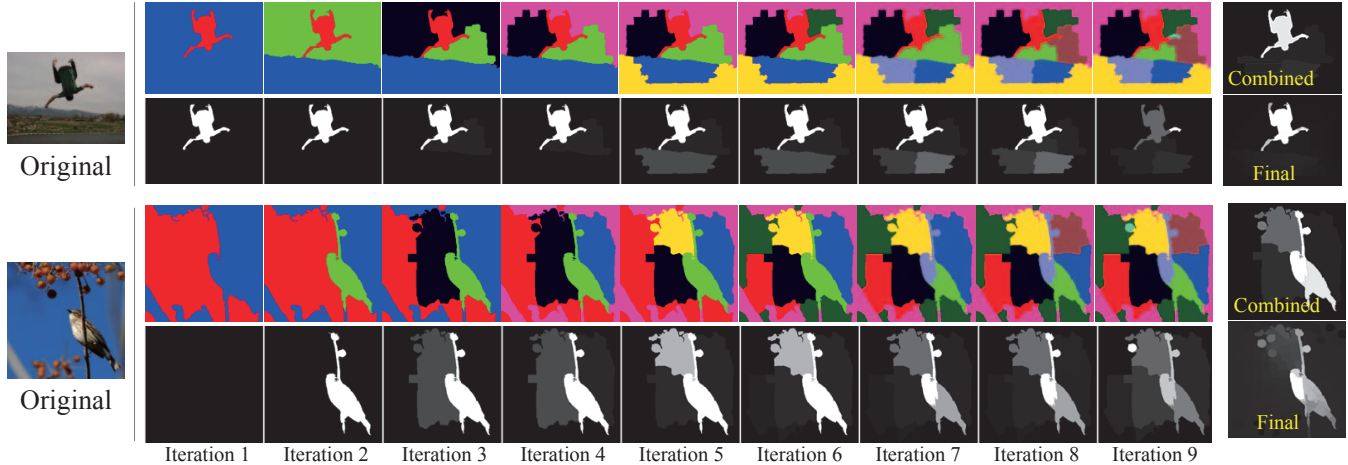
3. PROPOSED METHOD

When bring N-cut algorithm [25] into saliency detection, two issues remain to be tackled:

- Single N-cut segmentation is not always accurate for images containing multiple concepts (Fig.1). It is not reasonable to classify an image into two regions only. Hence how to recursively partition the image meanwhile pop out salient parts as early as possible by using a coarse-to-fine manner remains to be solved.
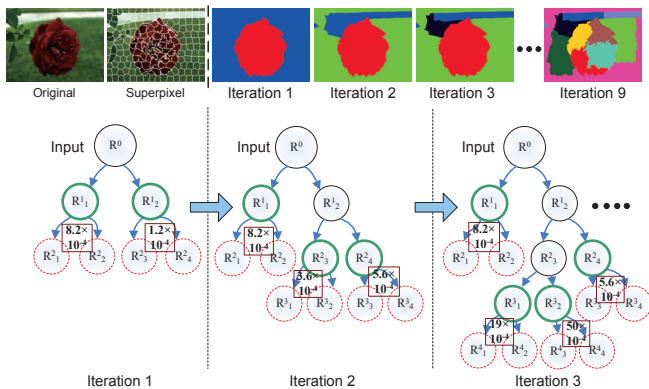- How to assign appropriate saliency values to partitioned segments remains to be considered.

We will show how we tackle the above two problems in the following subsections.

3.1. Coarse-to-Fine Spectral Partition

As a salient object is likely to exhibit unique appearance in terms of color or texture with respect to its surroundings, a low cut cost is required if one separates it from the remaining

Fig. 3. Partitioning process and intermediate saliency maps generated by our method. Different colors indicate different segments. The last column shows the integrated and refined final saliency maps. Note the saliency assignment of iteration 1 for the "bird" image is null as neither two segments pass criterion (*i*).

Fig. 2. Spanning binary tree structure. In each iteration, we choose the segment with the lowest cut cost to continue partition. The green solid circles indicate leaves (segments) generated in the current iteration. The superscript denotes depth whereas the subscript denotes index. The red dot circles indicate the pre-partitioned results. Cut cost values are shown in the boxes.

image. Motivated by this, we propose a hierarchical coarse-to-fine segmentation framework, which attempts to perform a good saliency cut at each iteration.
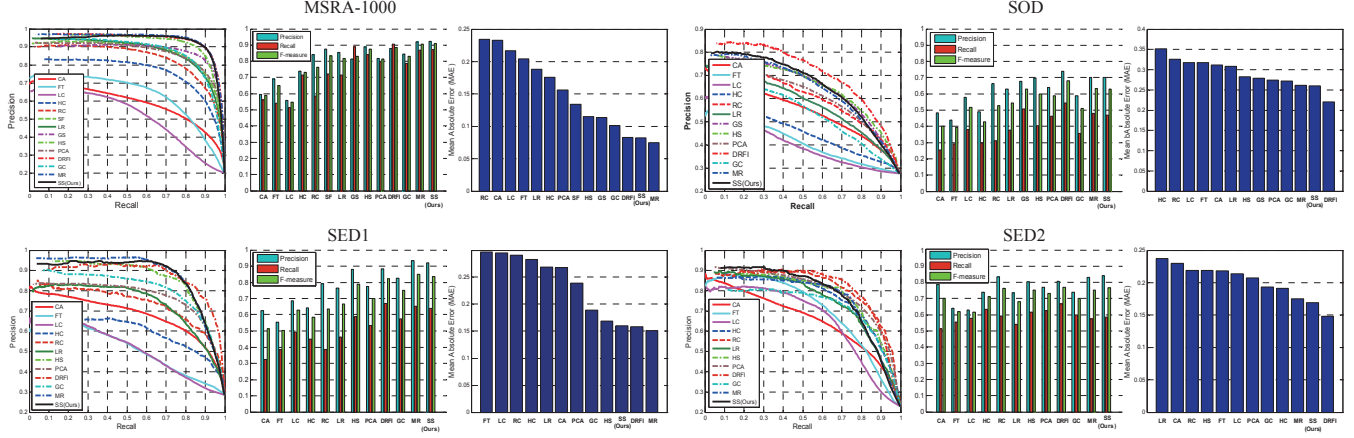
We first segment an input image using SLIC superpixels [24] (200 superpixels were used) and set out to construct an initial graph $G = (V, E)$ by treating superpixels as vertices V and color similarity of adjacent superpixels (defined using $\exp(-\gamma x)$, where x is color distance) as edges. We also construct edges between two arbitrary superpixels that touch image borders since they may belong to the same potential background (the effect of this operation is shown in the "bird" image in Fig.1). Regarding edges as affinity, we compute for the second smallest generalized eigenvector of system (3). After this step, one may choose to obtain binary indication values by thresholding the eigenvector using zero, median or mean eigenvector value, or finding the splitting point such that (2) has the best N-cut value [25]. Here we

choose the mean value since it provides sufficient accuracy and efficiency. Fig.1 shows several examples. Observing the successful cases in Fig.1, one can see that N-cut may provide guidance for saliency detection. On the other side, observing the failure cases in Fig.1, it could be found N-cut is not omnipotent. It is worth noting that if bottom up segmentation methods like [24, 23, 22] are employed to achieve the equally successful performance in Fig.1, their parameters need to be carefully and specially tuned for different images. In contrast, N-cut explicitly separates an image into two discriminative parts nearly without any parameter tuning, which reveals the reason why we recommend N-cut.

In the following procedure, attempt is made to cut out a salient object as early as possible, so that it can gain more enhancement in the final integration stage (will be shown in 3.3). In each iteration, we choose *one region from those generated in the last iteration* to implement N-cut (inspired by the 2-way cut in [25]), where the corresponding sub-graph is derived from G. As illustrated before, cutting out a salient object requires a low cut cost (2). Thus a reasonable way is to partition the part that requires the lowest cut cost, generating a potential cut which may separate different concepts.

In case of cutting out unwanted small discriminative segments which are likely to belong to parts of a salient target (e.g. arms of the jumping boy in Fig.3), we further penalize a cut cost value (2) by dividing the corresponding superpixel number (deemed proportional to region's area) in a whole region. Such bi-partition procedure leads to binary spanning tree structure (Fig.2) whose hierarchy stops until a manually determined number of leaves (regions) are generated.

To elaborate on the function of such tree structure, it provides a coarse-to-fine hierarchical segmentation. A salient object is more likely to appear in a coarser level, e.g. the flower in Fig.2. Hence the accumulation effect of the hierarchy can contribute to the enhancement of salient objects, as will be

Fig. 4. Quantitative comparisons (Precision-Recall, F-measure, Mean Absolute Error) between our method (SS) and the state-of-the-art solutions on MSRA-1000 [10], SOD [28] and SED [29]. The abbreviations have been listed at the beginning of Section 5.

described in 3.3.

3.2. Saliency Assignment to Each Segment

As the tree structure above only provides coarse-to-fine segmentation, it still remains unclear how salient each segment is. In this section, we propose to use several criterion to assign saliency value to each segment, e.g. $\{R_1^1, R_2^1\}$ of iteration 1 and $\{R_1^2, R_3^2, R_4^2\}$ of iteration 2 in Fig.2.

Following the figure-ground segregation in Gestalt laws [30], segments with the following properties are likely to be perceived as figures:

(*i*) segments with closed boundaries (surroundedness).

(*ii*) segments highly differentiating from the "ground".

Additionally, according to the statistical prior [31], human occurs center bias when watching an image. Thus:

(*iii*) segments near image center are likely to be salient.

Jointly considering (*i*)-(*iii*), saliency level of R_n^m is defined:

$$S_n^m = \begin{cases} 0 & \text{if } l_n^m > 1 \\ C(R_n^m) \exp(\alpha \sum_{k=1}^K ||c_n^m - c_k^{ground}||_2) & \text{otherwise} \end{cases} \quad (4)$$

According to (*i*), closed boundary indicates a segment should not be cropped by any image borders (an image gets four boarders). Let l_n^m be the number of borders a segment touch. Considering special cases such as the half-length portrait in photography, we discard segments that touch more than one image borders by setting S_n^m to zero.

According to (*ii*), we follow the observation that the border parts of an image are likely to be background [18, 19] and regard color appearance of initial superpixels which touch image borders as "ground unit" c_i^{ground} ($i = 1, 2, ...$). Color discrepancy between segment's mean color c_n^m and its K most

similar "ground unit" $\{c_k^{ground}\}_{k=1}^K$ (K is set as one quarter of the total boarder superpixel number) is computed. Function $\exp(\alpha x)$ is used to further enhance the contrast.

According to (*iii*), we further incorporate in (4) the segment's center bias as:

$$C(R_n^m) = \frac{1}{|R_n^m|} \sum_{I_i \in R_n^m} \exp(||p_i - p_c||^2/\sigma^2) \quad (5)$$
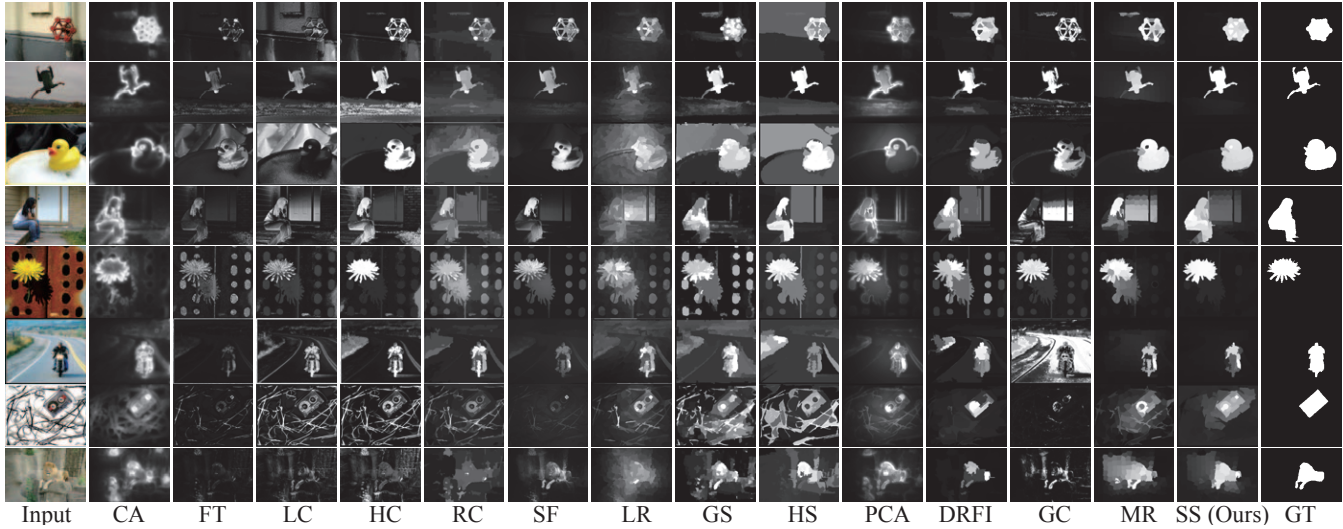
where p_i is the location of pixel I_i and p_c represents the image center. The standard deviation σ is set to one third of the longest dimension of the input image. In reality, computing exponent in (5) pixel-wisely for each input is less efficient. Instead we first load a Gaussian distribution saved off-line and then resize it via interpolation to adapt to the input size. Mean value in each segment is then computed by summing up the pixel Gaussian value and dividing the region's area $|R_n^m|$.

3.3. Final Saliency Map Formulation

An integrated saliency map (vectorization form denoted by s) is obtained by superpixel-wisely summing the intermediate saliency maps whose values are assigned to superpixels. For this reference please see "Combined" in the last column of Fig.3. Finally, the resulting map is further smoothed by a graph based regularization procedure which minimizes the following to pop up details and make the final result more visually consistent:

$$\min_x (x - s)^T(x - s) + \mu x^T L x \quad (6)$$

where x is the vectorized form of the final saliency map, in which each element indicates saliency of the corresponding superpixel. Laplacian matrix $L = D - W$ is derived from the initial graph G. μ is a tradeoff between the data term and the smoothing term (set to a large number, e.g. 1000, to emphasize smoothing). In reality, we only need to compute the closed-form solution obtained by differentiating (6):

Fig. 5. Visual comparisons on MSRA-1000 [10]. The abbreviations have been listed at the beginning of Section 5.

$$x = (I + \mu L)^{-1}s \qquad (7)$$

where I is the identity matrix. Fig.3 shows the intermediate partitioning results and the corresponding saliency assignment. Our rationale for integration is that a salient object, due to its global unique appearance, could pop out as early as possible with the help of our coarse-to-fine partitioning framework. Besides, it is less reliable to adopt only one hierarchy. In our system, we construct a tree structure with ten leaves (9 iterations) to perform saliency detection. This is found enough because usually an image may not contain too many visual concepts.

4. EXPERIMENTS AND RESULTS

In this section, we compared the proposed scheme with several state-of-the-art methods. **Metrics**: We compare in terms of three metrics which are widely adopted: (a) *Precision-Recall (PR)*, (b) *F-measure* [10, 11, 19, 16, 12], (c) *Mean Absolute Error (MAE)* [16, 12]. **Datasets**: Comparisons are made on three datasets as suggested in survey [32], including MSRA-1000 [10], SOD [28], SED [29] (contains two subset: SED1 for 1 object and SED2 for 2 objects). **Competitors**: We compare our technique (SS) with the state-of-the-art salient region detection methods: CA (Context Aware) [8], FT (Frequency Tuned) [10], LC (Luminance Contrast) [9], HC (Histogram Contrast) [11], RC (Region Contrast) [11], SF (Saliency Filter) [12], LR (Low Rank) [14], GS (Geodesic Saliency) [18], HS (Hierarchical Saliency) [33], PCA [15], DRFI (Discriminative Regional Feature Integration) [17], GC (Global Cue) [16], MR (Manifold Ranking) [19]. Either code or results of these techniques are publicly available.
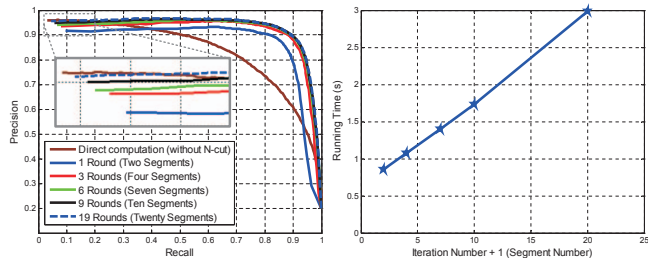
Comparisons: The Precision-Recall curves generated by using fixed threshold from 0 to 255 are shown in Fig.4. The performance of our method is comparable to the most re-

cent state-of-the-art techniques HS [33], DRFI [17] and MR [19]. Note our method achieves the best precision and F-measure on MSRA-1000, the second best precision and F-measure on SED1, and the highest precision on SED2. Besides, ours achieves nearly equal performance to MR [19] on SOD dataset, where DRFI [17] performs best due to its sophisticated learning-based feature integration strategy. The MAE of our method is also comparable, i.e. second best on MSRA-1000, SOD and SED2 and the third best on SED1.

Fig.5 shows the visual comparisons. Results generated by our method are not only closer to the ground truth but also more consistent with human perceptions. Notably, with the assistance of N-cut [25], our method can handle images that make the state-of-the-art methods less satisfactory, such as the 4th, 7th and 8th row in Fig.5. Note for the 8th row, the lion presents very similar color with the rocky background.

Effect of Number of Segments: Fig.6 shows the performance of SS on MSRA-1000 [10] when changing the segment number (i.e. iteration number+1). As exhibited in Fig.6, only cutting an image into two pieces (using 1 iteration) does not perform well enough, due to failure cases that resemble those in Fig.1. The precision improves as segment number increases. However, there is only marginal precision improvement under low recall by increasing 10 segments to 20 segments. Note more iterations lead to more computation of eigenvectors, which means more time cost (see Fig.6). As a tradeoff between accuracy and efficiency, we choose 9 iterations (10 segments) as default.

To further illustrate the benefit of the N-Cut based partition, we also show the PR curve (Fig.6) of directly computing saliency according to (4) and (5) on the initial superpixels (i.e. without N-cut partition and integration). The performance degrades quickly when recall increases, reflecting the N-cut is conducive to discovering object holism.

Fig. 6. The performance of our method on MSRA-1000 dataset [10] when changing the segment number (iteration number+1). Left: Precision-Recall curves under varied segment numbers. Right: The corresponding running time (for qualitative comparison) evaluated on Dual Core 2.6GHz PC with 3GB RAM using unoptimized Matlab code.

Limitation: As we use N-cut (a kind of balanced cut) to perform pre-segmentation, we indeed assume that desired salient contents can be split out after certain hierarchy. Though our method works well on highlighting big objects, we find it less satisfactory on detecting very small objects in complex scene (because they often can not be cut out effectively by N-cut). This most happens in SED2 where each image contains one large and one small object. That is why our method performs less better on SED2 in Fig.4. The underlying reason is that it is still ill-posed to identify whether a small high-contrast region is a background distracter or a single salient object by using only bottom up cues.

5. CONCLUSION

We have proposed to use N-cut [25] as a pre-segmentation technique for salient object detection, which helps effectively discover the object holism. A coarse-to-fine partition framework is introduced and accompanied by incorporating saliency measurement for segments based on Gestalt laws and statistical prior. Our method achieves comparable results on three widely used datasets with 13 state-of-the-art solutions.

6. REFERENCES

[1] L. Itti and E. NieBur et al, "A model of saliency-based visual attention for rapid scene analysis," *TPAMI*, vol. 20, no. 11, pp. 1254–1259, 1998. 1

[2] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," in *CVPR*, 2007. 1, 2

[3] N. Bruce and J. Tsotsos, "Saliency based on information maximization," in *NIPS*, 2005. 1

[4] A. Borji and L. Itti, "State-of-the-art in visual attention modeling," *TPAMI*, vol. 35, no. 1, pp. 185–207, 2013. 1

[5] F. Stentiford, "Attention based auto image cropping," in *Workshop on Computational Attention and Applications, ICVS*, 2007. 1

[6] L. Marchesotti et al, "A framework for visual saliency detection with applications to image thumbnailing," in *ICCV*, 2009. 1

[7] Y. Ding, X. Jing, and J. Yu, "Importance filtering for image retargeting," in *CVPR*, 2011. 1

[8] S. Goferman, L. Zelnik-Manor, and A. Tal, "Context-aware saliency detection," in *CVPR*, 2010. 1, 5

[9] Y. Zhai et al, "Visual attention detection in video sequences using spatiotemporal cues," *ACM Multimedia*, pp. 815–824, 2006. 1, 5

[10] R. Achanta, S. Hemami, F. Estrada, and S. Süsstrunk, "Frequency-tuned salient region detection," in *CVPR*, 2009. 1, 2, 4, 5, 6

[11] M. Cheng, G. Zhang, N. Mitra, X. Huang, and S. Hu, "Global contrast based salient region detection," in *CVPR*, 2011. 1, 5

[12] F. Perazzi, P. Krahenbul, Y. Pritch, and A. Hornung, "Saliency filters: Contrast based filtering for salient region detection," in *CVPR*, 2012. 1, 5

[13] K. Fu, C. Gong, J. Yang, and Y. Zhou., "Salient object detection via color contrast and color distribution," in *ACCV*, 2012. 1

[14] X. Shen and Y. Wu, "A unified approach to salient object detection via low rank matrix recovery," in *CVPR*, 2012. 1, 5

[15] R. Margolin, A. Tal, and L. Zelnik-Manor, "What makes a patch distinct," in *CVPR*, 2013. 1, 5

[16] M. Cheng, J. Warrell, W. Lin, S. Zheng, V. Vineet, and N. Crook, "Efficient salient region detection with soft image abstraction," in *ICCV*, 2013. 1, 5

[17] H. Jiang, J. Wang, Z. Yuan, Y. Wu, N. Zheng, and S. Li, "Salient object detection: A discriminative regional feature integration approach," in *CVPR*, 2013. 1, 5

[18] Y. Wei, F. Wen, W. Zhu, and J. Sun, "Geodesic saliency using background priors," in *ECCV*, 2012. 1, 4, 5

[19] C. Yang, L. Zhang, H. Lu, X. Ruan, and M. Yang, "Saliency detection via graph-based manifold ranking," in *CVPR*, 2013. 1, 4, 5

[20] Y. Xie and H. Lu, "Visual saliency detection based on bayesian model," in *ICIP*, 2011. 1

[21] K. Fu, C. Gong, I. Gu, and J. Yang, "Geodesic saliency propagation for image salient region detection," in *ICIP*, 2013. 1

[22] D. Comaniciu and P. Meer, "Mean shift: a robust approach toward feature space analysis," *TPAMI*, vol. 24, no. 5, pp. 603–619, 2002. 1, 3

[23] P. Felzenszwalb and D. Huttenlocher, "Efficient graph-based image segmentation," *IJCV*, vol. 59, no. 2, pp. 167–181, 2004. 1, 3

[24] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Ssstrunk, "Slic superpixels compared to state-of-the-art superpixel methods," *TPAMI*, vol. 34, no. 11, pp. 2274–2282, 2012. 1, 3

[25] J. Shi and J. Malik, "Normalized cuts and image segmentation," *TPAMI*, vol. 22, no. 8, pp. 888–905, 2000. 2, 3, 5, 6

[26] Y. Weiss, "Segmentation using eigenvectors: A unifying view," in *ICCV*, 1999. 2

[27] J. Carreira and C. Sminchisescu, "Constrained parametric min-cuts for automatic object segmentation," in *CVPR*, 2010. 2

[28] V. Movahedi and J. Elder, "Design and perceptual validation of performance measures for salient object segmentation," in *IEEE Computer Society Workshop on Perceptual Organization in Computer Vision*, 2010. 4, 5

[29] S. Alpert, M. Galun, R. Basri, and A. Brandt, "Image segmentation by probabilistic bottom-up aggregation and cue integration," in *CVPR*, 2007. 4, 5

[30] S. Palmer, "Vision science: Photons to phenomenology," *The MIT press*, 1999. 4

[31] B. Tatler, "The central fixation bias in scene viewing: selecting an optimal viewing position independently of motor bases and image feature distributions," *JoV*, vol. 14, no. 7, 2007. 4

[32] A. Borji, D. Sihite, and L. Itti, "Salient object detection: A benchmark," in *ECCV*, 2012. 5

[33] Q. Yan, L. Xu, J. Shi, and J. Jia, "Hierarchical saliency detection," in *CVPR*, 2013. 5