SALIENT OBJECT DETECTION VIA OBJECTNESS MEASURE

Sai Srivatsa R

Indian Institute of Technology, Kharagpur

R. Venkatesh Babu

Indian Institute of Science, Bangalore

ABSTRACT

Salient object detection has become an important task in many image processing applications. The existing approaches exploit background prior and contrast prior to attain state of the art results. In this paper, instead of using background cues, we estimate the foreground regions in an image using objectness proposals and utilize it to obtain smooth and accurate saliency maps. We propose a novel saliency measure called 'foreground connectivity' which determines how tightly a pixel or a region is connected to the estimated foreground. We use the values assigned by this measure as foreground weights and integrate these in an optimization framework to obtain the final saliency maps. We extensively evaluate the proposed approach on two benchmark databases and demonstrate that the results obtained are better than the existing state of the art approaches.

Index Terms— Image Saliency, Objectness Proposals, Image Segmentation, Superpixels

1. INTRODUCTION

The Human visual system has the ability to process parts of image which are relevant, discarding the rest. This helps us to perceive objects even before identifying them. Saliency detection, i.e. computationally detecting these relevant regions is a complex problem which takes cues from models in cognitive psychology, neurobiology and computer vision. It has gained a lot of attention in the recent years from the computer vision community owing to its use in object recognition [1], object segmentation [2], image re-targeting [3] and cropping [4], image retrieval [5] etc.

Works in saliency detection are classified into three categories : fixation prediction, Salient object detection and Objectness proposal generation.

Early models were biologically inspired and were evaluated on human eye fixation datasets. Ullman and Koch [6] define saliency at a given location as how different it is from its surrounding in color, orientation, motion, depth etc. Itti et al. [7] follow the same framework and propose a centre-surround contrast using Difference of Gaussian to obtain their Saliency maps. Ma and Zhang [8] use similar contrast analysis and extend it using fuzzy growth model. These models are evaluated on eye fixation databases. These models highlight edges and corners and are not suitable for detecting complete salient regions.

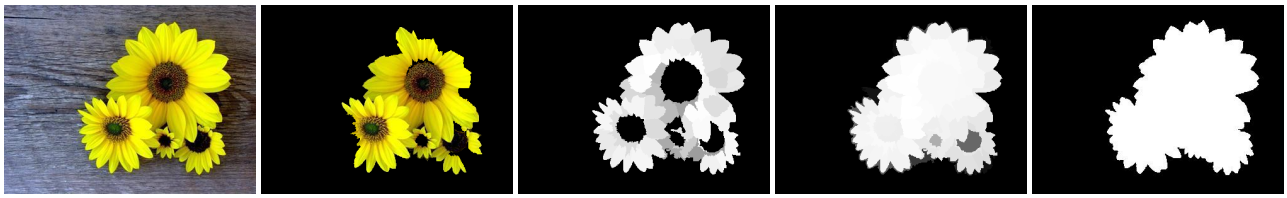
Salient object detection models aim to segment the object as a whole and are evaluated mostly on data labeled by humans such as bounding boxes or Foreground masks. These methods use low level cues such as contrast prior [9, 10, 11] or boundary prior [12, 13, 14]. Methods using contrast prior rely on uniqueness of the object and contrast between pixels or regions, center-surround differences etc. Methods based on contrast prior may be further classified into global or local methods. Local methods involve computing contrast measure in a local patch, whereas global methods use the entire image to compute the saliency. Global methods often use spatial feature as two pixels/regions which look similar but are far away need not belong to the same object. Global methods fail when the background is complex. Achanta et al. [9] computes saliency based on pixels color difference from the mean image color. Cheng. et al [10] combines global contrast with spatial differences to generate a saliency map. Perazzi et al. [15] computes saliency by decomposing the image into homgeneous elements. Contrast and spatial distribution are used to obtain pixel-accuarate saliency maps. These are estimated using high-dimensional Gaussian filter.

However, contrast prior alone is not very effective. The other most commonly used cue is based on the assumption that most photographers do not crop the salient object along the view frame. Hence the image boundary forms the background. However boundary prior is fragile and its prone to fail even when the object is slightly touching the background. Wei et al. [14] propose a saliency measure based on shortest length between an image patch and a virtual boundary node and it overcomes the shortcomings of boundary prior by connecting the boundary regions to a virtual node through an edge with suitable boundary weight. Yang et al. [12] ranks similarity of image regions with foreground or background cues using a graph-based manifold ranking. The ranking is based on relevance of an element with respect to the given queries. Zhu et al. [13] propose a boundary connectivity measure that utilizes both contrast prior and boundary prior. Foreground and Background weights obtained are then combined using an optimization framework.

Objectness proposal generation methods propose small number of windows that are likely to contain the object in an image thereby reducing search space for classifiers. Alexe et al. [16] propose an objectness measure that combines several image cues measuring an objects' characteristics in a Bayesian framework. Zhang et al. [17] propose cascaded ranking SVM to generate an ordered set of proposals. Cheng et al. [18] proposes a binarized version of normed gradient features (BING) which can be tested using few atomic operations to generate Objectness proposals.

Jiang et al [19] integrates Objectness with Uniqueness and Focusness to obtain saliency maps. However these maps are not smooth and it is difficult to attribute these results to specific algorithm properties [15]

In this work, rather than obtaining the background from image boundary and using the boundary prior, we quickly obtain a rough estimate of foreground regions by utilizing a modified version of the recently proposed objectness proposal technique [18]. We then compute super-pixel objectness which is a measure that quantifies how likely it is for a super-pixel to be a part of the foreground. The foreground and background regions are obtained by appropriately thresholding the above measure. We propose a robust saliency measure called foreground connectivity which assigns saliency values to these super-pixels.



(a) (b) (c) (d) (e)

Fig. 1. Illustration of main phases of our algorithm. (a) Input Image. (b) Thresholded Objectness Map. (c) Foreground weights (d) Saliency Map after Optimization (e) Ground truth

Rather than combining the cues heuristically (weighted summation or multiplication), we use a principled optimization framework proposed by [13] that regards saliency as a global optimization problem. The values assigned to the super-pixels are is considered as the foreground weights in the cost function. Minimizing the cost function would result in foreground regions taking higher values and background regions taking lower values. The obtained maps are also smooth and uniform due to the smoothness constraint in the cost function.

We have extensively evaluated our method on MSRA-1000 [9] and CSSD dataset [20] and we show that the proposed method performs at par or even better than the existing methods.

The paper is organized as follows. Section 2 describes various modules of the proposed approach including the new 'Foreground connectivity' metric. Experiments and results are discussed in section 3 and sec. 4 concludes the paper.

2. METHODOLOGY

The proposed approach is as follows. We build an Objectness Map using Objectness Proposals to capture super-pixels containing the object. Next, using the foreground connectivity measure, we assign foreground weights to super-pixels. We use saliency optimization technique to combine our foreground weights with background measure as used in [13] to obtain smooth and accurate saliency maps. Figure 1 illustrates the main phases of our algorithm.

2.1. Objectness Map

Objects are stand-alone things with well-defined closed boundaries and centers [16]. When windows containing objects are resized to a smaller size, the magnitude of norm of image gradients (NG) become good discriminative features. These normed gradients are inert to change of scale/aspect ratio and translation. The fact that objects share some correlation in the NG spaces is utilized in BING [18] to detect objects. The image is resized into fixed sizes and the normed gradient values in 8×8 region is used as 64 dimensional normed gradient feature. These windows are scored with a learned linear model $\mathbf{w} \in \mathbb{R}^{64}$.

$$s_l = \langle \mathbf{w}, \mathbf{g}_l \rangle, \tag{1}$$

$$l = (i, x, y), \tag{2}$$

where s_l, $\mathbf{g}_l$, l, i and (x, y) are filter score, NG feature, location, size and position of a window respectively. Using non-maximal suppresion, top proposals are chosen and are re-ranked based on their location and size using coefficients learned using a linear SVM. BING is accurate and extremely fast as the features are binarized and only a few atomic operations are required for obtaining the objectness proposals.

In the proposed approach we have adapted BING by modifying the following modules. Instead of using a learned linear model, we use an 8×8 Laplacian of Gaussian like filter and obtained scores for the windows. We skip re-ranking them based on the learned coefficients. The proposed model has higher weights placed along the edges and it resembles the center-surround patterns [6]. With this model, we also cut down the training time.

After obtaining the Objectness proposals, we generate the Objectness Map. Objectness score of a window tells us how likely it is to contain an object. We use these objectness proposals to obtain pixel-wise objectness ($PixObj$) score which tells us how likely it is for a pixel to be a part of an object. Pixel-wise objectness score is given by

$$PixObj(p) = \sum_{i=1}^{k} s_i G_i(x, y) \tag{3}$$

where $s_1, s_2 ... s_k$ are the objectness scores of the proposals containing pixel p and G_i is a Gaussian window having same dimensions as that of the given proposal , x and y are relative x and y coordinate of pixel p with respect to the given proposal.

Sum of pixel-wise object probability in a super pixel region gives us Objectness score of that super pixel region (which is used to construct the Objectness Map).

$$Objectness(R) = \sum_{i \in R} PixObj(p_i) \tag{4}$$

where p_i is a pixel belonging to super pixel region R. To obtain super pixel regions, we use SLIC [21] as it is fast and it preserves boundaries.

We use adaptive thresholding to obtain the Objectness Map. We observe that these Objectness Maps in few cases are able to segment out the salient object completely, however most objectness maps either miss out parts of object or include parts of background in it. This happens as the Objectness proposals are rectangular regions containing both foreground and background.

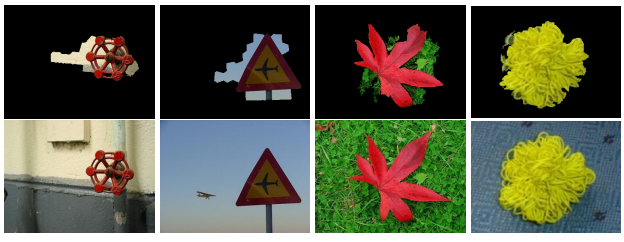


Fig. 2. Rough estimate of foreground obtained using Objectness Maps

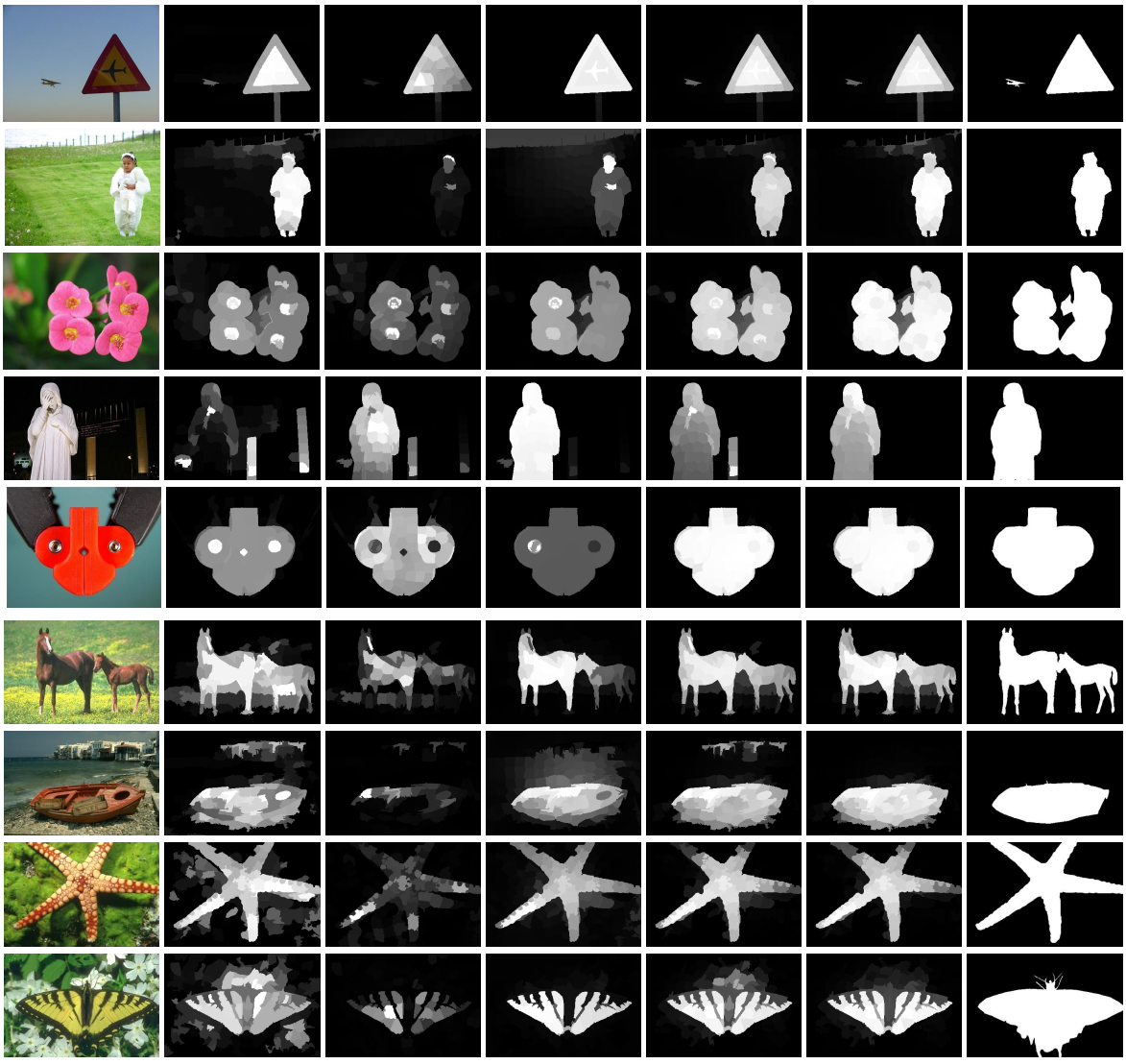


Fig. 3. Visual Comparison of Saliency Maps. (a) is the original image. Saliency map obtained using (b) GS [14] (c) SF [15] (d) MR [12] (e) SO [13] (f) Proposedd and (g) Ground Truth . The proposed approach generates Saliency maps that are accurate, smooth and uniform

2.2. Foreground Connectivity

Thresholded Objectness Maps roughly capture the super-pixels which are a part of the foreground. This is exemplified in Figure 2. We propose a novel saliency measure called 'foreground connectivity' that assigns saliency values based on a super-pixels connectivity to the estimated foreground. We construct a graph with super-pixels as nodes. Super-pixels that are adjacent in the images are connected by an edge with a weight equivalent to Euclidean distance of their mean LAB values. We now define foreground connectivity of a super-pixel R as :

$$FG(R) = \frac{\sum_{k=1}^{N} d(R, R_k).\delta(R_k)}{\sum_{k=1}^{N} d(R, R_k).(1 - \delta(R_k))} \quad (5)$$

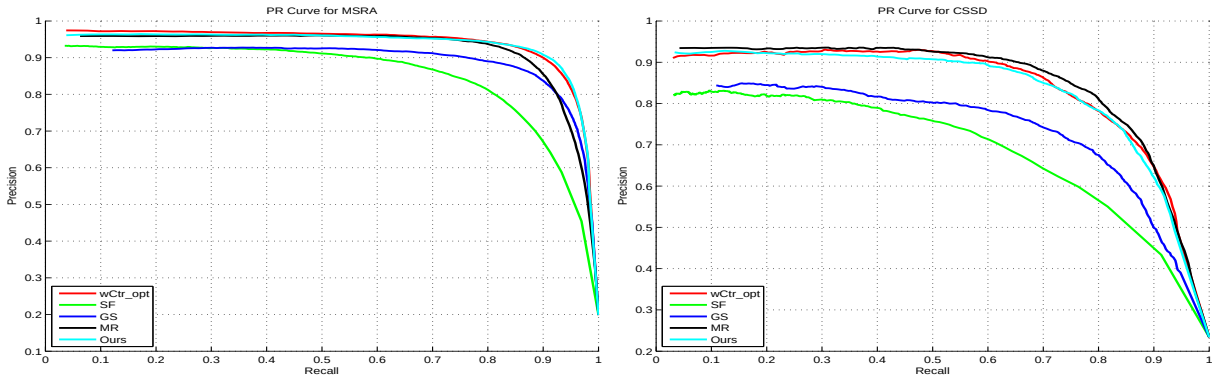
where $d(R, R_k)$ denotes the shortest distance between R to R_k and $\delta(.)$ is 1 for a super-pixel if it is estimated as foreground by the Objectness Map, and N is the total number of super-pixels.

A higher similarity of a super-pixel with the estimated foreground ensures lower value in the numerator and a higher value in the denominator leading to smaller value of FG (implying higher connectivity). We take the reciprocal of FG and use it as the foreground weights (w^{fg}).

2.3. Saliency Optimization

Generally, several saliency cues are combined heuristically using weighted summation or multiplication. Instead we use an existing optimization framework to combine our foreground weights with background weights as used in [13]. The cost function to be minimized is defined as

$$\sum_{i=1}^{N} w_i^{fg}(t_i - 1)^2 + \sum_{i=1}^{N} w_i^{bg}(t_i)^2 + \sum_{i,j} w_{ij}(t_i - t_j)^2 \quad (6)$$



(a) (b)

Fig. 4. Comparison of PR curves on (a)MSRA-1000 database and (b) CSSD-200 database

where t_i denotes the final value of saliency assigned to p_i after minimizing the cost, w_i^{fg} denotes foreground weights, w_i^{bg} denotes background weights associated with super-pixel p_i. High w_i^{fg} encourages p_i to take values close to 1 and high w_i^{bg} encourages p_i to take values close to 0. w_{ij} is the smoothness coefficient.We use the same parameter settings for w_{ij} and w_i^{bg} as used in [13].

3. RESULTS

We evaluate the proposed approach on two benchmark datasets. The first one is the MSRA-1000 [9] dataset which is one of the most extensively used databases. The second one is the CSSD dataset [20]. The MSRA-1000 dataset has large varieties in content and background but are simple and smooth. The CSSD dataset on the other hand has structurally complex images for evaluation. Results obtained on both the datasets are evaluated on ground-truth masks labeled by humans. We compare the results of our algorithm with recent four state of the art methods : Saliency Filter (SF) [15] , geodesic Saliency (GS) [14] , Manifold Ranking (MR) [12] and Saliency Optimization (SO)[13]. Results on MSRA-1000 [9] and CSSD [20] are shown in Figure 3. We evaluate our method using Precision-Recall curves and Mean Absolute Error (MAE).

3.1. Precision and Recall

Precision is the fraction of pixels assigned correctly against the total number of pixels assigned salient. Whereas recall is the fraction of pixels labeled correctly in relation to the number of ground truth pixels. Precision and recall vary inversely and hence it is essential to evaluate them simultaneously. Hence, we use a Precision-Recall curve similar to previous works . For various value of threshold between $[0 \ldots 255]$, we obtain binary maps and using the ground truth mask,we compute precision and recall values. Figure 4 shows that the proposed approach performs at par with the other state of the art algorithms. However precision recall curves have some serious limitations. Precision-recall curves do not consider the fraction of pixels correctly assigned as not salient. Presence of pixels incorrectly assigned as salient brings down the performance of saliency map despite it being smooth and having higher values assigned to pixels that are salient.

3.2. Mean Absolute Error

To overcome this limitation, we use Mean Absolute Error (MAE) as suggested by [15]. It measures how similar a saliency map is to the ground truth. For a saliency map S, ground truth mask G, then MAE is defined as

$$MAE = \frac{1}{W \times H} \sum_{x=1}^{H} \sum_{y=1}^{W} |S(x,y) - G(x,y)| \qquad (7)$$

where H and W denote the height and width of the image. Results have been averaged out over all the images in the database. The proposed approach performs better than the state of the art methods in terms of MAE (see Table 1).

	MSRA	CSSD
GS [14]	0.109	0.178
SF [15]	0.129	0.204
MR [12]	0.085	0.150
SO [13]	0.068	0.136
Proposed	**0.064**	**0.132**

Table 1. Comparison of MAE values of different saliency methods

3.3. Running Time

The average running time of the proposed approach on an Intel Core i5-4200U CPU @ 1.60 GHz with 6GB RAM is 0.27s excluding preprocessing and superpixel segmentation using SLIC [21].

4. CONCLUSION

In this paper, we present a simple and efficient method that utilizes Objectness Proposals and foreground connectivity measure to detect salient objects in an image. Unlike recent methods, we obtain our saliency maps by estimating foreground regions in an image instead of using boundary priors. Our method combined with the optimization framework produces accurate and smooth saliency maps that perform better than other methods in terms of MAE when tested on two widely used datasets. In future, we plan to investigate better cues rather than depending on contrast or boundary prior alone, better connectivity measures and better objectness proposal techniques that can perform well with backgrounds that are even more complex.

5. REFERENCES

[1] U. Rutishauser, D. Walther, C. Koch, and P. Perona, "Is bottom-up attention useful for object recognition?," 2004, pp. 37–44.

[2] M. Donoser, M. Urschler, M. Hirzer, and H. Bischof, "Saliency driven total variation segmentation," 2009, pp. 817–824.

[3] Sorkine O. Shamir A Rubinstein M., Guterrez D., "A comparative study of image retargeting," in *SIGGRAPH Asia*, 2010.

[4] Csurka G. Marchesotti L., Cifarelli C., "A framework for visual saliency detection with applications to image thumbnailing," in *ICCV*, 2009.

[5] Tao Chen, Ming-Ming Cheng, Ping Tan, Ariel Shamir, and Shi-Min Hu, "Sketch2photo: Internet image montage," *ACM Transactions on Graphics*, vol. 28, no. 5, pp. 124:1–10, 2009.

[6] C. Koch and S. Ullman, "Shifts in selective visual attention: towards the underlying neural circuitry," *Human Neurbiology*, vol. 4, pp. 219–227, 1985.

[7] Laurent Itti, Christof Koch, and Ernst Niebur, "A model of saliency-based visual attention for rapid scene analysis," vol. 20, no. 11, pp. 1254–1259, 1998.

[8] Yu-Fei Ma and Hong-Jiang Zhang, "Contrast-based image attention analysis by using fuzzy growing," 2003.

[9] Radhakrishna Achanta, Sheila Hemami, Francisco Estrada, and Sabine Süsstrunk, "Frequency-tuned salient region detection," 2009, pp. 1597–1604.

[10] Ming-Ming Cheng, Guo-Xin Zhang, Niloy J. Mitra, Xiaolei Huang, and Shi-Min Hu, "Global contrast based salient region detection," 2011, pp. 409–416.

[11] S. Goferman, L. Zelnik-Manor, and A. Tal, "Context-aware saliency detection," 2010, pp. 2376–2383.

[12] Huchuan Lu Xiang Ruan Chuan Yang, Lihe Zhang and Ming-Hsuan Yang, "Saliency detection via graph-based manifold ranking," 2013.

[13] Yichen Wei Wangjiang Zhu, Shuang Liangy and Jian Sun, "Saliency optimization from robust background detection," 2014.

[14] Wangjiang Zhu Yichen Wei, Fang Wen and Jian Sun, "Geodesic saliency using background priors," 2012.

[15] Federico Perazzi, Philipp Krähenbühl, Yael Pritch, and Alexander Hornung, "Saliency filters: Contrast based filtering for salient region detection," 2012, pp. 733–740.

[16] B. Alexe, T. Deselaers, and V. Ferrari, "Measuring the objectness of image windows," vol. 34, no. 11, 2012.

[17] Ziming Zhang, Jonathan Warrell, and Philip HS Torr, "Proposal generation for object detection using cascaded ranking svms," 2011, pp. 1497–1504.

[18] Ming-Ming Cheng, Ziming Zhang, Wen-Yan Lin, and Philip H. S. Torr, "BING: Binarized normed gradients for objectness estimation at 300fps," in *IEEE CVPR*, 2014.

[19] Peng Jiang, Haibin Ling, Jingyi Yu, and Jingliang Peng, "Salient region detection by ufo: Uniqueness, focusness and objectness," in *ICCV'13*, 2013, pp. 1976–1983.

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

[21] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and Sabine Susstrunk, "SLIC superpixels compared to state-of-the-art superpixel methods," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 11, pp. 2274–2282, 2012.