Saliency Detection Using Min-cut Proposals Framework

Meiling Sun[1,a], Fengxia Li[1,a], Sanyuan Zhao[1, a], Da Huo[1, a], Chenguang Yang[2, b]

[1]Beijing Institute of Technology, Beijing, China

[2]Chinese Electronic Equipment System Corporation Institute, Beijing, China

[a]sunmeiling625@163.com, [b]297756723@qq.com

Keywords: Saliency Detection, Objectness Proposals, Min-cut, SVM.

Abstract. In saliency detection, almost all the approaches map an image into a graph and assign the saliency value to each element, e.g. pixel, region or superpixel. In this paper, we first utilize a series of image features among superpixels in the support vector machine (SVM) to train linear predicted models. For a well-performance model we take cross validation in the supervised learning. Then, we take the SVM regression models to predict initial saliency maps, while using SVM classifier to get the foreground and background seeds. Besides, we employ an objectness min-cut algorithm to obtain the segments of different proposals. Finally, after ranking these proposals, we select the top one integrating with the initial maps to achieve the final saliency maps. The proposed approach is tested extensively on four different databases and then compared with existing algorithms.

Introduction

Human get among 80% of the external information via human visual system (HSV), with the capability of capturing and processing information selectively, we call it the selective visual attention mechanism. This kind of mechanism has two common processing stages in computer vision: detecting the salient object and segmenting its boundary. Salient object is defined as the specialty or the uniqueness of the scene [1].

Images can be depicted as pixel-level superpixel-level and patch-level in image processing. The main purpose of image segmentation is to assign each pixel a label whether it belongs to the object or the background. While the salient object detection only focuses on most notable objects, which treats the segmenting as binary marking. In practice, to make a balance between accuracy and computational speed, recent salient objects detection methods commonly use the concept of superpixel.

Along with the prevalence of the superpixel algorithm, (e.g. [8]), proposed region-based models get increasingly popular. Saliency known as the specialty or uniqueness in the terms of global region contrast has been researched widely on superpixel-level. In [14], the algorithm decomposes the image into basic superpixel, computes the uniqueness and distribution of superpixel, assign the saliency to each pixel, finally get a pixel-accurate saliency map. A two-stage saliency detection superpixel-level algorithm is used on an undirected weighted graph based on manifold ranking in [19].

Except for the similarity of processing elments, the algotithms above all get the information from the input image, we can also get cues from the extrinsic information, such as ground-truth annotation [9], similar images [6], video sequence diagram [4], or depth map [13].

In addition to the spatial information given by a single image, video sequences can provide additional temporary cues of time, such as moving objects. In [10], they introduce an unsupervised method, incorporating the geodesic distance, using both spatial edges and temporal motion boundaries as features for salient video object segmentation.

Our paper is organized as shown in Fig.1. To get a better performance in detection procedures, we combine the appearance features and regional properties as the superpixel feature. In addition, we take the geodesic distance, scale and rotation features into consideration. We adopt a supervised machine learning approach to coarsely predict the initial saliency maps and select the foreground and background seeds. Finally, we cut and rank objectness proposals. We fuse the proposals with the initial saliency maps. The frame of our algorithm is shown in Fig.1. Experiments are carried out on MSRA-1000 [7], ECSSD dataset [18], HKU-IS [21] and Pascal-s [20] respectively.

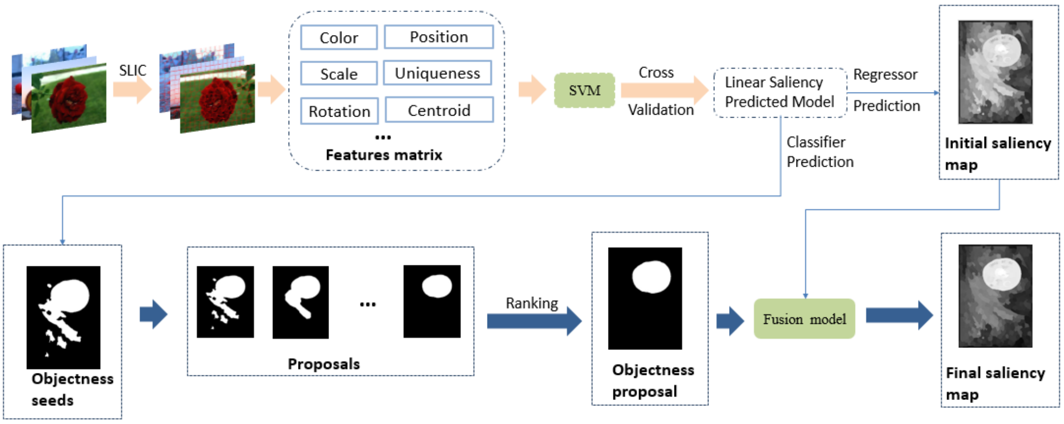


Fig. 1 The framework of our proposal algorithm

Linear saliency predicted model

Superpixel Features. We employ the simple linear iterative clustering (SLIC) superpixels [8] algorithm to segment images into regular regions and extract the regional features. To learn more precise models, it is necessary to present a set of features on the superpixel level. In this work, we compile a small set of features and these features are totally 51-dimensional, describing each superpixel. The details are shown in Table 1. The compiled feature vectors extracted high-level information and are used to train the predicted models which can find the statistical principles via these properties.

Table 1 Features in our model

Feature Types	Feature Names	Dim	Feature Types	Feature Names	Dim
Boundary	Boundary connectivity	1	**Color**	RGB histogram	27
	Length of boundary	1		Mean RGB	3
	Probability of Boundary	1		Mean Lab	3
Scale	Area Span	1	**Translation**	Position	2
	Minor Axis Length	1		Centroid	2
	Major Axis Length	1	**Appearance**	Uniqueness	1
	Bouding box	4	**Rotation**	Orientation	1
	Convex area	1	**Scale**	Eccentricity	1

Linear Support Vector Regression and Classification. To compute the saliency map reliably, we first calculate a coarse saliency map by SVM. We take [16] for SVM training and testing just for its high speed. The L2-regularized L2-loss support vector regression and classification are suitable to obtain ideal results. The learnt regression model is used to predict the probability that a superpixel being foreground or background, and an initial saliency prediction map can be generated according to the regression output value. We utilize the learnt classification model to feature the possible attributes for background or foreground, labelled 0 or 1, and take them as the seeds in the following sections. What's more, we put together different datasets that are employed, and take 5-folder cross validation to overcome overfitting.

Objectness based on min-cut

The initialization of foreground or background seeds received great interests in the field of recent vision literature especially in saliency detection. The simple selection of the background and foreground seeds restricts the effect of objectness proposals,.In this paper, we formulate a lucid model for seeds selection as a classification problem. Learning a classifier coarsely maps the foreground and background regions.

In CPMC [12], alternative sets of pixel-level seeds are hypothesized. Their foreground seeds are placed on a regular grid geometry, whereas background seeds are related to sets of pixels along with the image boundaries. Compared to CPMC, our selection for foreground and background seeds is based on the outputs of the SVM which are all on SLIC seperpixel level. Hence our algorithm has following advantages: firstly, SLIC superpixel level. Hence our algorithm has following advantages: firstly, our algorithm can have a better segmentation for images that salient objects are along with the image boundaries. In other words, ours can cut much more appropriately regardless of the location of the object regions. If the foreground seeds can be chosen within a target area in the image exactly, in most cases, we are able to get an ideal segmentation result by adjusting the amount of bias to a λ value. The worst case is that no foreground seed belongs to the target area. Thus, knowing the approximate location of the foreground region in an image can greatly improve the efficiency and accuracy of the algorithm. Secondly, SLIC superpixel can play a better performance in constraining the noise, while the noise on the pixel-level is inevitably. Thirdly, selection in superpixel-level has higher efficiency than pixel-level.

It is necessary to construct a graph *G* = (*V*, *E*) with N nodes for each image. We firstly use SLIC algorithm[8] to segment the image into superpixels $\{v_i\},(i=1,2,...N)$, the union of these superpixels is the nodes collection *V*, The set of all direct edges $\{\mathrm{e}_{ij}\},(i=1,2,...N)$, between adjacent superpixels is *E*. The feature value of the superpixels is represented by the average value. The seed regions serve as the starting points for object proposals. We can get a label $\{l_i\},(i=1,2,...N)$ for every image, where 1 and 0 represent the foreground and background seeds respectively. The label is the input of next process including min-cut and manifold ranking. The appearances and boundaries near the seeds are used to identify other superpixels that might belong to the same category, either the foreground or the background. By changing foreground bias, we can obtain a series of results using segmentation method based on graph-cut.

For given seed nodes v_b and v_f, the idea of objectness proposal is to minimize energy equation given below:

$$E(\lambda,X)=\sum_{i\in V}D(X_i)+\sum_{(i,j)\in\varepsilon}V_{ij}(x_i,x_j) \tag{1}$$

In this equation, $\lambda\in R$ and λ means the foreground bias. Unary term is also called data term. Superpixels value D_i represents the shortest geodesic distance between each superpixel and foreground or background seeds, defined as:

$$D(x_i)=\begin{cases}0, & x_i=1, i\notin v_b\\ \infty, & x_i=1, i\in v_b\\ \infty, & x_i=0, i\in v_f\\ f(x_i,\lambda), & x_i=0, i\notin v_f\end{cases} \tag{2}$$

In this equation, $f(x_i,\lambda)$ represents geodesic distance of superpixel x_i under different values of foreground bias: $f(x_i,\lambda)=d_{geo}(v_i,T)+\lambda$, where *T* is pre-selected image "reference ground", which means the background seeds regions in our experiment. Geodesic distance is the shortest weighted path from v_i to the node v_j of "reference ground". On the basis of the overall performance, geodesic distance is better than Euclidean distance in maintenance of the internal geometrical characteristics of the data. Using geodesic distance instead of Euclidean distance can illustrate global consistency among super-pixels. Geodesic distance is defined as: $d_{geo}(v_i,v_j)=\min_{C_{v_i,v_j}}\sum\left|w_{ij}\bullet C_{v_i,v_j}\right|$,*where* C_{v_i,v_j} *is* a path connecting the nodes v_i to v_j (equals to 0 or 1 for the node whether on the path or not).

In Eq.1, binary term, which is also known as smooth term, is represented by $\left(l_i, a_i, b_i\right)$ in the CIELab color space between two adjacent superpixels. The feature $f_i = \left[l_i, a_i, b_i, x_i, y_i\right]$ is constituted by combining with the superpixel spatial position $\left(x_i, y_i\right)$ with the binary term. Furthermore, it treats Euclidean distance between adjacent superpixels in five-dimensional vector space as undirected edge weight, namely:

$$V_{ij}\left(x_i, y_i\right) = \exp\left(-\sigma \bullet \left\| f_i - f_j \right\|\right) \quad (3)$$

In Eq. 3 described as above, σ is set as 10 in our experiment. Binary term well reflects the similarity between adjacent superpixels in appearance features. Smaller difference leads to the smaller binary term, which also means smaller cost of belonging to the same label.

Eq. 1 can be solved by the Gcmex solver in [2]. In the energy equation, the degree of segmentation among superpixels in SLIC algorithm affect the number of proposals directly. We choose 25 different $\lambda \in \left(-0.5, 0.8\right)$ values to obtain a pool of segments that may contain the proposal regions. The proposal generated using all these 25 different λ values must have the redundancy or do not have the salient object in it. We firstly reject very small segments by retaining the regions that are bigger than 30*30 pixels in our experiment. After the first rejection, the number of proposals decreases. We cast the ranking problem of these proposals as shown in [3].

After completing this step, we can get a pool of segmentation proposals for every image. The results of different λ values are illustrated as Fig. 2. We take the top 1 proposal for each image denoted that $\left\{\mathrm{Pro}_i, i = 1, \ldots.n\right\}$ (n as the number of images).

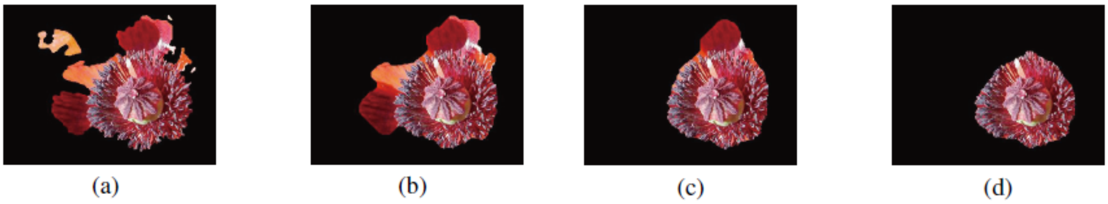


Fig. 2 Different cut results according to λ

Multi-scale initial saliency maps and fusion model

The accuracy of the saliency map is sensitive to the number of superpixels as salient objects are likely to appear at different scales. To deal with the scale problem, we generate four layers of superpixels with different granularities, where N = 200, 300, 400, 500 respectively. We represent the initial saliency map at each scale as $\left\{M_j, j = 1, \ldots, 4\right\}$ and we take the mean map of these four map as the multi-scale initial saliency map. As such, the proposed method is robust to scale variation.

The proposed initial saliency maps and the proposals have complicated properties. The initial saliency maps are from a global perspective that detailed are ignored. In contrast, the proposals lay more emphasize on to the local information. In this case, we integrate this two results by a weighted combination:

$$S_i = \alpha * Sal_i + \left(1 - \alpha\right) * \mathrm{Pro}_i \quad (4)$$

Where i is the index of images and α is a balance parameter in the combination. In our experiment, α=0.65 can get better maps.

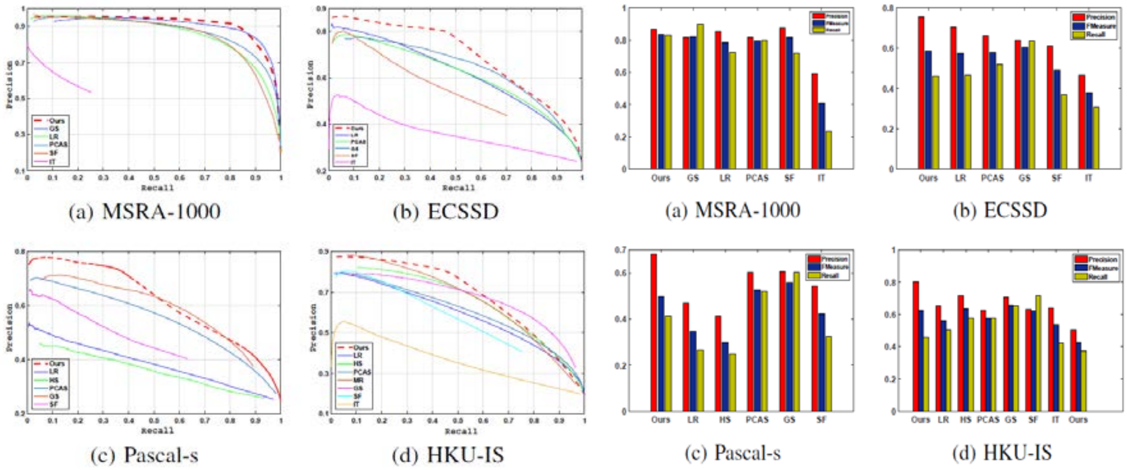
Experiment

We implement our proposed algorithm on four benchmark datasets including MSRA-1000, ECSSD, HKU-IS and Pascal-s. The MSRA-1000 is the most common dataset which is widely used in the field of image processing. The ECSSD dataset contains structurally complex images. The Pascal-S dataset contains 850 images which are also labeled with pixel-wise ground-truth. The

HKU-IS is recently released by Hong Kong University contained 4447 images. Additionally, we compare our algorithm with six representative saliency approach: IT[1],SF [14], HS [18] , LR [11] , GS [15] and PCAS [17] . The intuitional result comparison is shown in Fig. 3. We employ the F-measure and the precision and recall curve (PR curves) to evaluate the performance of our method. The F measureis computed as follows:

$$F_{\beta}=\frac{\left(1+\beta^{2}\right)\bullet precision\bullet recall}{\beta^{2}precision+recall} \tag{5}$$

The parameter β is set to 0.3 according to the literature. To compute the precision and recall, an adaptive threshold is calculated as twice of the mean saliency values of the saliency map, and then the saliency map is binarized to foreground and back-ground by this threshold. Our method achieves the better F-measure on four datasets.



(a) MSRA-1000 (b) ECSSD (a) MSRA-1000 (b) ECSSD

(c) Pascal-s (d) HKU-IS (c) Pascal-s (d) HKU-IS

Fig. 3 PR curve(left) and F-number(right)

Conclusion

As is shown in Fig.4, our saliency detection approach combined with the machine learning and objectness proposals performs well on the superpixel level. Even though only machine learning module can not work well in detecting salient regions, we can also use it in selecting seeds. In the framework of objectness proposal, we select seeds more effectively than the classcical gragh cut algorithm, in that better result as is shown below.

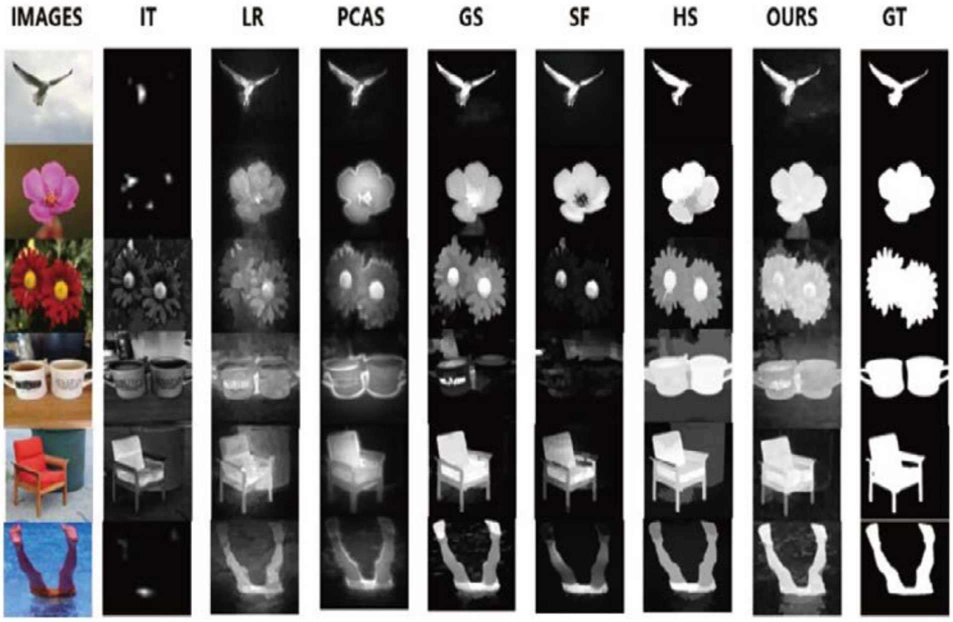


Fig. 4 The result comparison with some existed algorithms

References

[1] Itti L, Koch C, Niebur E. IEEE Transactions on Pattern Analysis & Machine Intelligence, 1998, 20(11):1254-1259.

[2] Boykov Y, Kolmogorov V. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2001, 11(26):1124-37.

[3] Endres I, Hoiem D. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2014, 36(2):222-34.

[4] Marichal X, Villegas P. Objective evaluation of segmentation masks in video sequences[C]// Signal Processing Conference, European. IEEE, 2015.

[5] Tao D, Cheng J, Song M, et al. IEEE Transactions on Neural Networks & Learning Systems, 2016, 27(6).

[6] Kapoor A, Biswas K K, Hanmandlu M. Visual Computer, 2016:1-21.

[7] Hemami, S., F. Estrada, and S. Susstrunk. "*Frequency-tuned salient region detection*." IEEE International Conference on Computer Vision and Pattern Recognition (CVPR 2009) 2009:1597-1604.

[8] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Susstrunk, EPFL Technical Report no. 149300, June 2010.

[9] Cheng M M, Mitra N J, Huang X, et al. The Visual Computer, 2014, 30(4):443-453.

[10] Wang, Wenguan, J. Shen, and F. Porikli. "Saliency-aware geodesic video object segmentation." Computer Vision and Pattern Recognition IEEE, 2015.

[11] Shen, Xiaohui. "A unified approach to salient object detection via low rank matrix recovery." IEEE Conference on Computer Vision & Pattern Recognition 2012:853-860.

[12] Carreira, J., and C. Sminchisescu. Pattern Analysis & Machine Intelligence IEEE Transactions on 34.7(2012):1312-1328.

[13] F. Shafieyan, N. Karimi, B. Mirmahboub, et al. Signal Processing: Image Communication, 2016.

[14] Perazzi F, Krahenbuhl P, Pritch Y, et al. 2012, 157(10):733-740.

[15] Wei, Yichen, et al. "Geodesic Saliency Using Background Priors." European Conference on Computer Vision 2012:29-42.

[16] Information on http://www.csie.ntu.edu.tw/~cjlin/liblinear/

[17] Margolin, R., Tal, A., and Zelnik-Manor, L. "What Makes a Patch Distinct." IEEE Conference on Computer Vision and Pattern Recognition IEEE Computer Society, 2014:1139-1146.

[18] Shi J, Yan Q, Li X, et al. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2015, 38(4):1.

[19] Yang, Chuan, et al. "Saliency Detection via Graph-Based Manifold Ranking." IEEE Conference on Computer Vision and Pattern Recognition IEEE Computer Society, 2013:3166-3173.

[20] Li, Yin, et al. "The Secrets of Salient Object Segmentation." Eprint Arxiv (2014):280 - 287.

[21] Li G, Yu Y. Visual Saliency Based on Multiscale Deep Features. 2015:5455-5463.