Category-Independent Object-level Saliency Detection

Yangqing Jia
UC Berkeley
jiayq@eecs.berkeley.edu

Mei Han
Google Research
meihan@google.com

Abstract

It is known that purely low-level saliency cues such as frequency does not lead to a good salient object detection result, requiring high-level knowledge to be adopted for successful discovery of task-independent salient objects. In this paper, we propose an efficient way to combine such high-level saliency priors and low-level appearance models. We obtain the high-level saliency prior with the objectness algorithm to find potential object candidates without the need of category information, and then enforce the consistency among the salient regions using a Gaussian MRF with the weights scaled by diverse density that emphasizes the influence of potential foreground pixels. Our model obtains saliency maps that assign high scores for the whole salient object, and achieves state-of-the-art performance on benchmark datasets covering various foreground statistics.

1. Introduction

Many computer vision applications may benefit from understanding where humans focus given a scene. Other than cognitively understanding the way human perceive images and scenes, finding salient regions and objects in the images helps various tasks such as speeding up object detection [27, 23] and content-aware image editing [4].

There is a line of saliency detection work centered around visual attention models [13, 15, 12] that focuses on finding locations of images that capture early-stage human fixations before more complex object recognition or scene parsing takes place. While this bears much importance in understanding human visual systems, we focus on the problem of finding salient *objects*, aiming to find consistent foreground objects, which is often of interest in many further applications such as object detection. Existing work have suggested that purely low-level information (such as the frequency domain image signature [12]) often does not produce object-level saliency maps, and high-level information such as common object detectors [15, 28] and global image statistics [6] aid the selection of the most salient regions in the images.

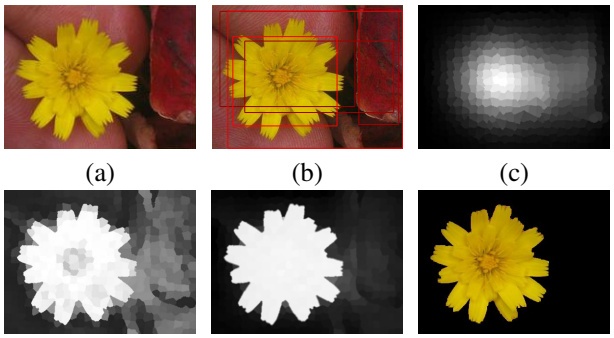


Figure 1. An illustration of our approach from images to the final saliency map: (a) Input Image (b) objectness detections, (c) saliency prior from objectness, (d) diverse density scores for pixels, (e) the final saliency map, and (f) the segmented object.

In this paper, we propose a novel approach that fuses top-down object level information and bottom-up pixel appearances to obtain a final saliency map that identifies the most interesting regions in the image. While early work such as [15] uses specific category-dependent object detectors, there has been several work focusing on learning category-independent object representations in the recent years [3, 7, 11], making it possible to detect objects without the need of knowing the object category present in the images. Specifically, we adopt the recent objectness framework [3] that finds potential object candidates in the image. Such objectness information is then passed onto the pixel level as a prior of the per-pixel saliency.

To fuse the high-level object information with the pixel appearances, we used a fully-connected Markov random field (MRF) that takes into consideration the overall agreement between salient regions over the whole image, with an explicit emphasis on nodes that are more likely to be foregrounds inspired by the work of graph-based multi instance learning [25]. Unlike classical frequency-based saliency maps that often focus on edges of the images and objects, we will show that our method returns a more object-centric saliency map that is consistent over the salient object, achieving state-of-the-art performance on the bench-

mark MSRA and Weizmann datasets.

We will start by reviewing the related work in the saliency detection field, and then formally describe our algorithm in Section 3 and 4, including the employment of the objectness cue in our model, and the Markov random field that fuses high-level object information and low-level appearance. The experiments on the MSRA and Weizmann datasets, as well as the analysis of performance, are presented in Section 5.

2. Related Work

Pre-attentive bottom up saliency algorithms have been extensively studied from biological and computational perspectives. These algorithms used low-level information including biologically inspired center-surrounding operators [13], gradient information [29], local contrast features [17, 20], frequency-domain information [12, 2], etc. We note that such information is not neglected in our framework, but is rather incorporated into the objectness detection component.

On combining high-level object knowledge and low-level appearances, Chang *et al*. [5] was the first to adopt a high-level object information as saliency prior. However, the prior is combined with pixel-wise scores from another low-level saliency model, which then creates an arbitrary bias towards the specific algorithms' behaviors, such as favoring high frequency areas, and may in some cases hurt the final performance. Other work, especially in segmentation [26, 16], adopt parameterized models such as Gaussian Mixture Model (GMM) to model the foreground and to cut out the foreground region with coarse supervised information. While such tasks (such as cosegmentation) explicitly need to identify the mixture components of the foreground, it may not be necessary in finding saliency regions, and the multiple parameters to be tuned in these models may hurt performance. We empirically tested a parameterized mixture model for foreground modeling, and found the MRF approach in our paper to better fit the saliency problem.

To obtain a consistent salient object detection, an important structural choice is to use a fully-connected graph, rather than a locally connected graph as many previous approaches do [5], as locally connected graph may lead to overly smoothed saliency maps. We note that this is not the first time fully connected graphs have shown advantages. Previous work such as [18] have shown a significant performance gain over locally connected graphs. With the help of superpixels, inference takes only a short time for a reasonably sized image. Note that efficient inference with such graphs exists even for large-scale graphs.

Finally, several previous works have explored the choice of feature extraction from pixels or superpixels, such as pixel values and Gabor filter responses [28, 6]. There are also work on weighting or encoding for more discriminative features [15, 28, 20]. What we will show in the paper is that, despite its simplicity, a purely top-down prior and a fully-connected graph built on simple color features could achieve state-of-the-art performance, without the need of additional bottom-up saliency prior or additional handcrafted features.

3. Saliency Detection with Object-level Information

In this section, we formally describe the algorithm we proposed to perform saliency detection based on high-level object information.

3.1. Object Detection

Our method starts with finding an informative prior that captures the potential salient regions from images. While specific object detectors such as faces and vehicles have been adopted to help finding good prior knowledge of salient objects [15, 28], we focus on algorithms that are able to handle general object appearances without category-specific information. To this end, we adopted the objectness algorithm as proposed in [3] to find a set of object candidates in input images.

Specifically, the objectness algorithm finds a set of object candidates represented by bounding boxes, together with the confidence scores, for each input image. It adopts four different low-level cues to learn if a certain bounding box contains an object or not, which we give a brief explanation to the cues adopted in the method as follows for completeness.

1. *Multiscale Saliency (MS)*: this cue utilizes the spectral residual of the Fourier Transform on multiple scales to find regions with unique appearances within the image.
2. *Color Contrast*: this cue computes the dissimilarity of the color distribution of a candidate bounding box with that of its surrounding area. The idea is that an object would look sufficiently different from the background that surrounds it.
3. *Edge Density*: this cue computes the density of the edges (computed by Canny edge detection) near the borders of the candidate bounding box. The idea is that an object would have a dense edge distribution around its boundary.
4. *Superpixel Straddling*: this cue computes the agreement between the candidate bounding box and the super pixels obtained by [8]. Since pixels in the same superpixel often belong to the same semantic group (either the object or the background), for a good object candidate most superpixels should lie mostly either inside or outside the bounding box, and should not cross the boundary.

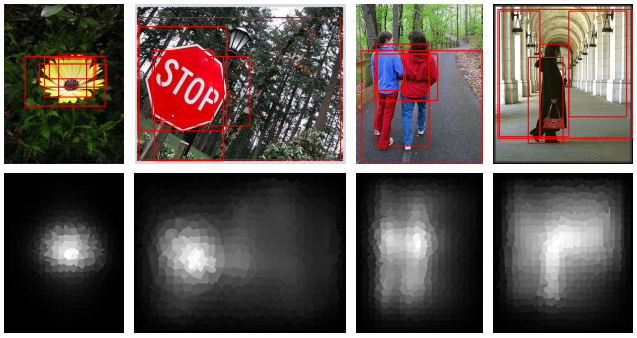


Figure 2. The top row shows sample images with the most confident 5 bounding boxes, with the bottom row the corresponding saliency map priors obtained from objectness.

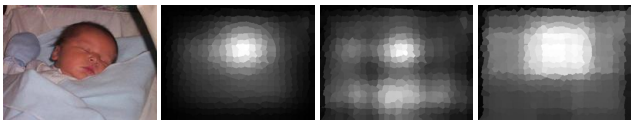


Image rss ave no smooth

Figure 3. The per-superpixel image saliency obtained directly from the objectness detection. Computing the rss of object scores allow us to favor areas with densely detected objects, and smoothing allows us to reduce boundary artifacts.

We note that unlike previous works such as [5], we keep the low-level, frequency based saliency component (MS) in the objectness pipeline. This allows the objectness method to more accurately identify possible objects in the image, and as we will discuss in the next section, low-level saliency may impose a negative impact on the final saliency measure when used alone.

We trained the objectness parameters on a randomly selected subset of ImageNet images that are separate from our testing data. For each input $\mathbf{I}$, we then performed objectness detection on each image with K top object candidates[1], denoted by $\{(B_1, b_1), (B_2, b_2), \cdots, (B_K, b_K)\}$ where B_k is the bounding box and b_k is the corresponding confidence score. Figure 2 shows exemplar results with the most confident 5 bounding boxes shown. In most cases, the algorithm is able to capture the correct location of the salient object, although in rare cases such as the last column example, where the large number of vertical lines in the background building causes the objectness to bias towards it.

3.2. Pixel-level Objectness Scores

As our goal is to obtain a saliency map for the whole image, we transfer the objectness scores from the bounding boxes to the pixel level. To this end, we propose to directly adopt a pooling approach similar to the ones used in image classification, by computing each pixel p's objectness score (denoted by s_p) as the square root of the summed squares (rss) of the scores from all the bounding boxes that covers it, weighted by a Gaussian function for smoothness:

$$s_p = \Big[\sum_{i=1}^{N} b_i^2 I(p \in B_i) \exp\{-\lambda d(p, B_i)\} \Big]^{1/2} \quad (1)$$

where b_i is the objectness score for the bounding box, $I(p \in B_i)$ is the 0-1 indicator function denoting whether p is inside the bounding box, and $d(p, B_i)$ is the normalized distance between the pixel p and the center of the bounding box B_i measured in a scaled coordinate space where the bounding box is normalized to length 1 on both axes. The exponent term provides a discounting factor so that pixels far from the bounding box center receives less contribution from the bounding box than pixels near the bounding center do.

To reduce the computation cost for subsequent steps we adopted the idea of superpixels and averaged the saliency values of pixels inside each superpixel. In our experiments, we adopted the Turbopixel algorithm [19] to produce superpixels that have similar sizes[2]. Further, since the scale of the objectness scores from Eqn. 1 may vary due to different objectness detections, we re-normalize the per-pixel scores based on each image so that the maximum score is 1 and the minimum is 0 over the whole image.

Figure 3 shows the resulting per-pixel objectness score with summed pooling and two baseline choices, average pooling (as often used in classification), and without smoothing. It could be observed that although the saliency map is still coarse, it provides a reasonable initialization for the final saliency map as it correctly identifies the salient object location. More importantly, such saliency prior is not biased towards specific low-level appearances such as high-frequency regions, which often misses the inside region of the salient objects.

4. Saliency Computation with Graph-based Foreground Agreement

The pixel level prior gives us a reasonably informative result on the salient regions of the images. However, due to the fact that objectness bounding boxes are often over-complete, the saliency map is often very coarse, and one would expect low-level appearance based information to be helpful in refining the saliency maps.

Thus, in our model we propose to extract features for individual pixels, and use a Markov random field to enforce

[1] We fixed K to be 1,000 in all our experiments to obtain a good estimation of the object locations in the image.

[2] While one may also want to use alternative algorithms such as the one in [8], we found the Turbopixel algorithm to empirically work much better in our algorithm, due to the fact that methods like [8] may produce many small superpixels (usually in highly textured areas). Such statistics will bias our further inference algorithm towards small highly textured areas, a negative effect for saliency detection.

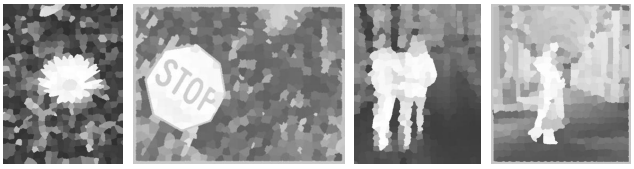


Figure 4. The superpixel diverse density values for the images shown in Figure 2.

agreement between salient regions in the image, based on the similarities between pixel level features. The idea is that if a pixel has a high salient prior, then pixels that appear similar in the image should also receive high salient scores even if it lies in a region with low contrast, thus ensuring a consistent saliency score assignment in the whole region of the salient object. As a simple example, consider a red object in a random background - while conventional frequency-based saliency detections will only be able to capture the edges of the object, enforcing consistency between salient pixels makes it possible to give high saliency values for the inside of the object as well.

4.1. MRF with Diverse Density

To construct the Markov random field, a classical approach is to consider each superpixel as a node and have the edges represent the spatial connection with them. While this enforces spatial consistency, it does not help in *appearance consistency*, which is of more interest in our work. Thus, we use a fully connected MRF where any two superpixels are connected, with the corresponding edge weight computed as

$$W_{ij} = \exp\left(\frac{-\|p_i, p_j\|_2^2}{2\sigma^2}\right) \tag{2}$$

between pixels i and j, where p_i and p_j are the features associated to the pixels. This leads to an $N \times N$ weight matrix $\mathbf{W}$ where N is the number of superpixels. Despite its simplicity, we found the raw color for the pixels to work best in our case against other choices such as Gabor outputs. The colors are represented in the LAB colorspace as distance in LAB matches human perception well.

A potential issue with the direct computation of the weight is that with images of small foreground regions, the large number of background superpixels will dominate the spectral characteristics of W, making it relatively hard to identify the foreground object. Intuitively, a foreground pixel should have more influence in propagating the saliency information than pixels in the background.

Since we do not have foreground and background labels in the first place, we could use the saliency prior as an approximation, and evaluate the "influence" of each pixel with a diverse density measure, which is inspired by the work of multi-instance learning [22, 25], to evaluate the agreement between potentially foreground pixels.

Specifically, diverse density suppresses the weight values associated to pixels that are less likely to be foregrounds: given the objectness scores $\mathbf{s}$ for all the pixels, the diverse density of a pixel i is computed as

$$\mathrm{DD}_i = \sum_j \Big(W_{ij} s_j + (1 - W_{ij})(1 - s_j)\Big) \tag{3}$$

where W_{ij} is the weight between the pixels i and j computed as Eqn. 2. The diverse density models how near other salient regions are to it, and how far other non-salient regions are from it, with the saliency approximated by the prior information s_j. For normalization purposes, we then normalize all diverse density values by

$$\mathrm{DD}_i \leftarrow \left(\frac{\mathrm{DD}_i}{\max_j \mathrm{DD}_j}\right)^{\gamma} \tag{4}$$

where γ is a scaling factor that controls the peakness of the diverse density measure. In practice we found $\gamma = 4$ to work well under various image appearances. Figure 4 shows an example of the diverse density scores obtained. Then, we define the weight of the MRF, denoted by $\mathbf{G}$, to be the conventional weight $\mathbf{W}$ scaled by the DD value of each pixel. The weight between the pixels indexed i and j is computed as

$$G_{ij} = \frac{\mathrm{DD}_i + \mathrm{DD}_j}{2} W_{ij} \tag{5}$$

To introduce the saliency prior obtained from objectness, we then add two abstract nodes: one source node with saliency value 1 that connects to each pixel p with weight s_p, and one sink node with saliency value 0 that connects to each pixel weight value $1 - s_p$. Then, we solve for the improved saliency value for each pixel by viewing the graph as a Gaussian MRF, which leads to an efficient computation of the final saliency values $\hat{\mathbf{s}}$ as

$$\hat{\mathbf{s}} = \Big(\mathrm{diag}(\mathbf{G1}) - \mathbf{G}\Big)^{+} \begin{bmatrix}\mathbf{s} & 1 - \mathbf{s}\end{bmatrix} \begin{bmatrix}1 \\ 0\end{bmatrix} \tag{6}$$

where $\mathbf{1}$ is the all-one vector of length N. Examples of the final saliency map can be seen in Figure 7.

4.2. Analysis of Performance Contributions

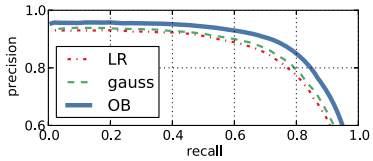
With the multiple stages of many current saliency detection algorithms, it would be interesting to observe how much each component contributes to the overall performance. Specifically, we are interested in finding how much the two key components, *i.e.* the objectness based prior and the diverse density based inference algorithms, contribute to the overall performance of the algorithm.

To analyze the effect of prior, we replace the objectness based prior with three baselines: a uniform prior over the

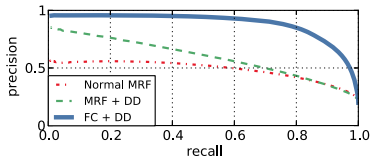
whole image, a Gaussian prior that favors the center of the image, and a more sophisticated prior from [28] that combines multiple cues, such as location, semantics and color to learn a final informed prior saliency. We then use them as the initialization of our graph, and perform GMRF inference to get the final saliency measure.

Figure 5(a) illustrates the precision-recall curves resulting from the priors. A prior that captures coarse locations of the foreground object does bear importance, as the uninformed priors do a very poor job of identifying the salient region. It is interesting that despite its simplicity, a Gaussian prior works as well as the prior from [28], partially as the latter also derives from combination of location and appearances. Both are still worse than the proposed objectness prior, which gives us a 4% average further precision increase, suggesting that the general objectness measure serves as a good heuristics in saliency detection.

Figure 5(b) shows different choices of the graph construction methods. We start from a normal MRF construction, where only spatially connected superpixels are connected in the graph. The diverse density (DD) term is then imposed on computing the edge weights of the graph. Both baselines are then compared against our method that uses both a diverse density term and a fully connected graph. The results show that diverse density provides with a significant precision gain in the low recall area, possibly resulting from preventing background superpixels to have a too strong influence on neighboring superpixels. Using a fully connected graph allows us to obtain a significant performance gain, suggesting that although spatial relationships are crucial in obtaining saliency priors (as is both the case in objectness detection and works like [28]), they should not play an important role in later stages of saliency computation, possibly due to the many irregular and even disjoint foreground objects (see *e.g.* the last 3 rows of Figure 7).



(a) Different Prior



(b) Different Graph

Figure 5. (a) PR curves under different choices of priors. The flat prior is not shown as it achieves only 0.2 average precision and is much below the shown priors (note the y axis is from 0.6 to 1). (b) PR curves under different choice of graph construction. FC means fully connected and DD means diverse density weighted.

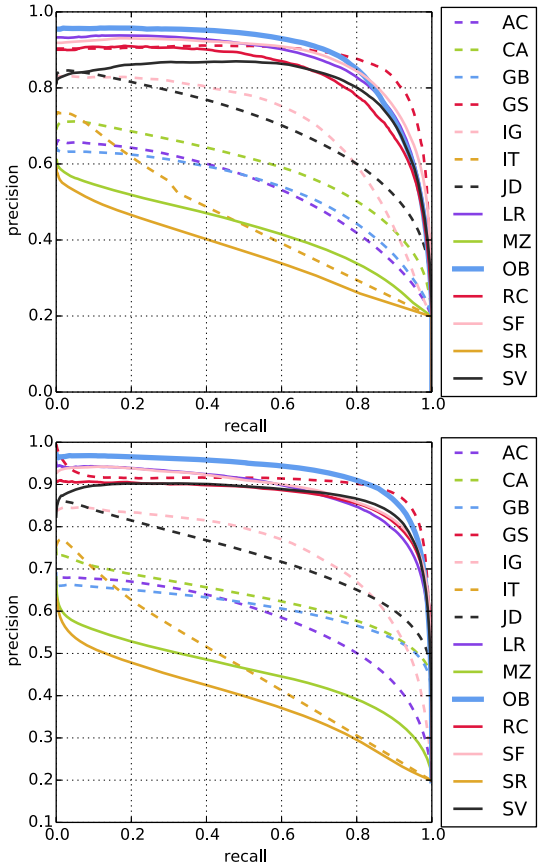


Figure 6. The precision-recall curves for our saliency detection algorithm and the baseline algorithms. The top figure shows `PR-overall` and the bottom figure shows `PR-individual`. The methods are sorted alphabetically.

5. Experiments

We evaluated our method on the MSRA saliency dataset containing 1000 images together with the salient object annotated by human participants as the ground-truth saliency map, and compared the performance against state-of-the-art algorithms. Our saliency maps on the MSRA dataset are publicly available at `http://www.eecs.berkeley.edu/~jiayq/` for benchmarking purpose.

5.1. Evaluation Criteria

We mainly adopted the criteria introduced in [2] to evaluate the performance of various saliency algorithms using precision recall (PR) curves. In addition, we used two different criteria to generate the PR curves. The first method, which we call `PR-overall`, follows the conventional criterion in the literature and uses a fixed threshold T to get precision and recall values over all images and then compute the average. Varying the T value between 0 and 255 gives us the average PR curve. The second method, which we call `PR-individual`, focuses on analyzing the

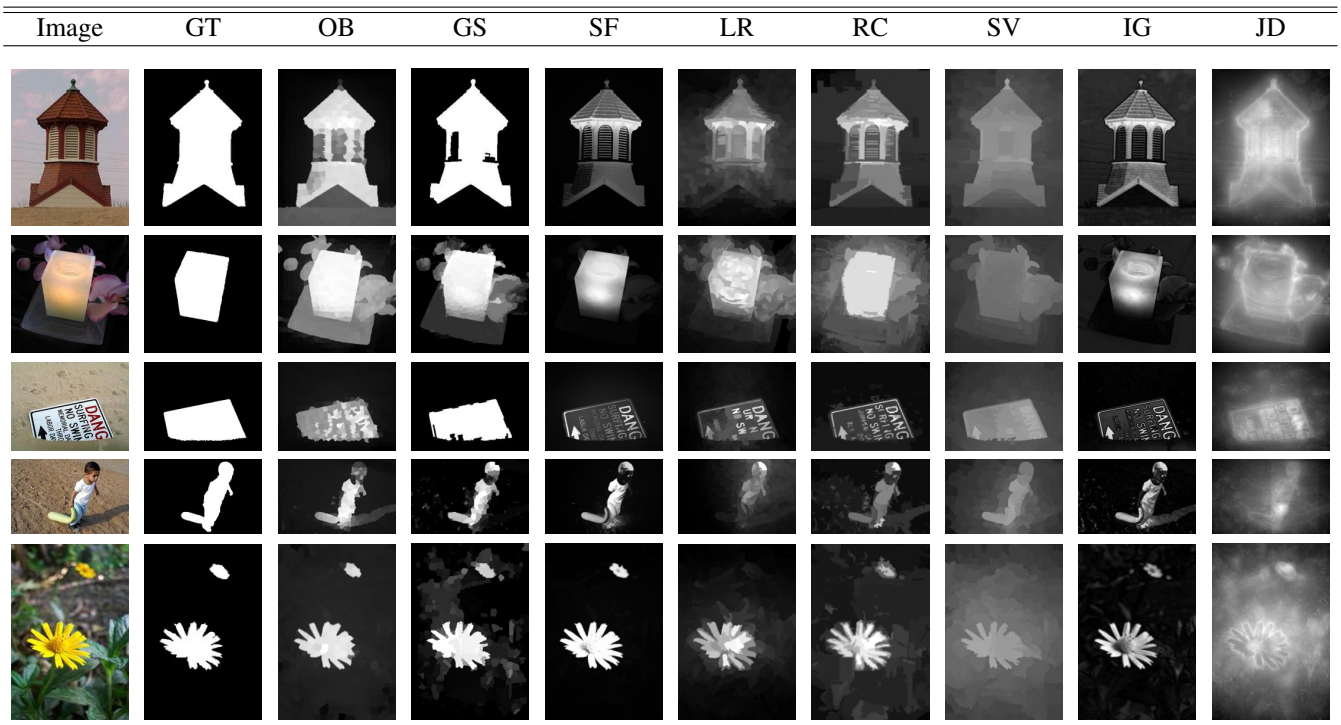


Figure 7. Results on the MSRA dataset with the saliency maps of the best 8 methods, ordered from left to right, where GT is the ground truth and OB is our approach. See the text for references to individual baselines.

saliency orders within each image: it uses the threshold to generate per-image PR curves, and then computes the average precision values at fixed recall (between 0 and 1 with a stepsize of 0.01) values to get the final PR curve. Intuitively, the first method represents the robustness of the algorithm in a cross-image fashion when an uninformative threshold is used, and the second focuses on checking the correct saliency order for pixels in a single image.

We also report the performance when we binarize the saliency map with an adaptive thresholding method. For binarization, we computed the mean m and the standard deviation σ of the saliency map, and then set all pixels whose saliency value is larger than $m + \sigma$ to be foreground and the rest to be background. We then followed the benchmark introduced in [2] and report the average precision and recall values over the images, as well as the F-measure computed as

$$F_\beta = \frac{(1 + \beta^2) \times \text{precision} \times \text{recall}}{\beta^2 \text{precision} + \text{recall}}$$

with $\beta^2 = 0.3$ as used in the literature [2].

5.2. Summary of Performance

We summarize in Figure 6 the performance of several baseline algorithms, which could be briefly separated to two categories. The first category of baselines uses low-level signals only, including Achanta *et al.* (AC)[1], context-aware saliency (CA)[9], graph-based visual saliency (GB)[10], frequency-tuned saliency (IG)[2], Itti *et al.* (IT)[13], contrast based attention model (MZ)[21], spectral residual approach (SR)[12], and saliency filters (SF)[24]. The second category is generally proposed more recently, including Judd *et al.* (JD) [15], global contrast based saliency (RC)[6], saliency by low rank recovery (LR)[28], Chang *et al.* (SV)[5], geodesic saliency (GS)[30] and also our method. Such methods utilize high-level object or global image information to create an informative prior for the saliency map. For all baseline methods, we used either the published implementations with their recommended parameters or the author-provided saliency maps.

In general, methods that utilize high-level information to obtain more informative saliency priors perform better than purely low-level approaches, and our method achieves the highest average precision on both PR curves over all baselines. When measured using the size of the area under the PR curve (aka. average precision, AP), our method achieves 92.70% with `PR-individual` and 87.63% with `PR-overall`, while the second best approach (GS) achieves 90.21% and 87.82% on the two criteria respectively, achieving a higher recall while our method has a higher precision. Figure 7 shows exemplar images and their corresponding saliency maps from various algorithms. Full results on the dataset can be found in the supplemen-

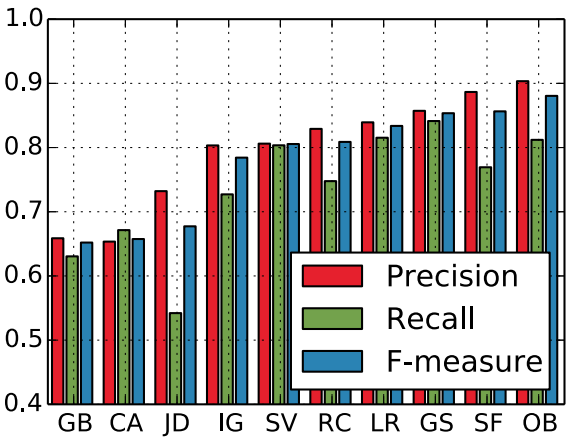


Figure 8. Precision, recall and F-measures of the adaptive thresholding method, sorted by the F-measure. Our method (the rightmost one) achieves the best overall performance (LR and GS obtains higher recalls despite lower F-measure). See supplementary material for numbers.

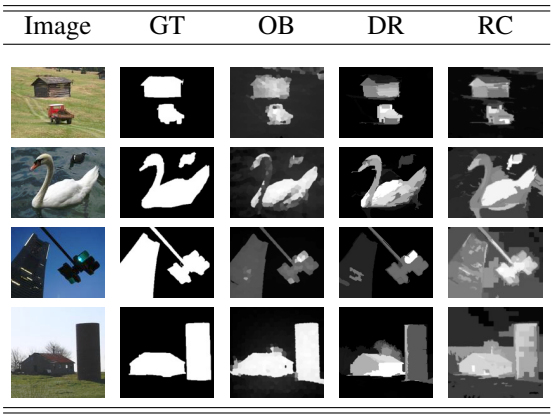


Figure 9. Results on the Weizmann Dataset with 2 foreground objects, showing the 3 best methods (OB, DR and RC). GT = ground truth.

tary material.

Figure 8 summarizes the performance of adaptive segmentation. As our binarization routine is different than the one used in some baseline algorithms, we re-ran both binarization on all baselines and reported the better result for a fair comparison. It can be observed that our method obtains the best result on the overall F-measure with a significant precision rate increase. This partially results from the fact that it correctly identifies foreground regions that have a consistent appearance, without being biased towards *e.g.* high-frequency low-level saliency predictions.

Finally, we note that the best results on the MSRA dataset by the time we write the paper is achieved by Jiang *et al.* with discriminative regional features (DR)[14], in which a pixel-wise saliency prediction model is trained on ground-truth saliency maps[3]. It is interesting to note that a major contribution is also due to the introduction of object-level information, further justifying the use of such approaches in saliency detection.

5.3. Performance on the Weizmann Dataset

The MSRA saliency dataset mainly contains single salient objects of medium sizes per image, which is the assumption made by several saliency detection algorithms, especially those with a high-level object appearance model. To evaluate the performance of our approach under more varied conditions such as multiple foreground objects, we used the Weizmann Dataset, which contains two subsets of images with single foreground object and two foreground objects respectively.

Figure 10 summarizes the precision-recall curves for our method and DR, LR, RC, SV, IG and CA whose results are published or implementations are available. It could be observed that, when there is only 1 object present (Figure 10(a)), the performance of the algorithms align with those on the MSRA dataset very well, and methods with high-level model (LR, SV) perform better than pure low-level models (RC, IG, CA). However, multiple foreground objects being present hurts some high-level model baselines (Figure 10(b)), leading to an even slightly worse performance than good low-level models (RC). This is possibly due to the fact that these models explicitly models one single foreground (LC) or favors a connected foreground (SV). Our method is free from such assumptions,

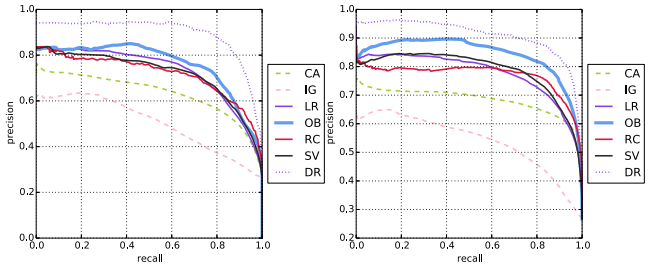
On contrary, our model does not make such assumptions, and is able to naturally cope with multiple foreground blobs, as we model the appearance of the foreground with a graph, implicitly allowing mixtures of foreground appearances[4]. As a result it performs on par with the supervised approach (DR) on the 2 object dataset. We visualize representative results from the Weizmann 2-object dataset in Figure 10.

6. Conclusion

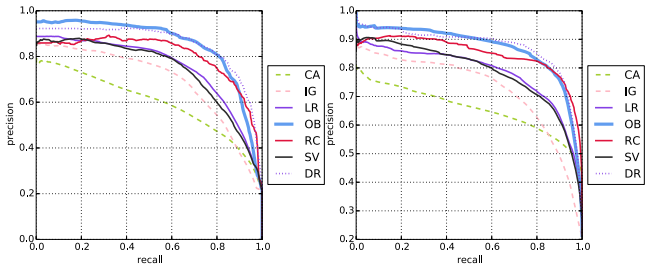
In this paper we proposed a novel image saliency algorithm that utilizes the object-level information to obtain better discovery of salient objects. In the model, we obtain the high-level saliency prior with the objectness algorithm to find potential object candidates without the need of category information, and then enforce the consistency among the salient regions using a Gaussian MRF with the weights scaled by diverse density, which emphasizes the influence of potential foreground pixels. Our model obtains saliency maps that assign high scores for the whole salient

[3]We report the performance on the MSRA dataset in the supplementary material, as their result is on a test subset and is not directly comparable.

[4]We note that similar benefits have also been observed in the multi-instance classification field [25].



(a) Weizmann 1 Object Dataset



(b) Weizmann 2 Object Dataset

Figure 10. The precision-recall curves for our saliency detection algorithm and the baseline algorithms on the Weizmann dataset.

object, and achieves state-of-the-art performance on benchmark datasets covering various foreground statistics.

References

[1] Radhakrishna Achanta, Francisco Estrada, Patricia Wils, and Sabine Süsstrunk. Salient region detection and segmentation. In *ICVS*, 2008. 6

[2] Radhakrishna Achanta, Sheila Hemami, Francisco Estrada, and Sabine Susstrunk. Frequency-tuned salient region detection. In *CVPR*, 2009. 2, 5, 6

[3] B. Alexe, T. Deselaers, and V. Ferrari. What is an object? In *CVPR*, 2010. 1, 2

[4] Shai Avidan and Ariel Shamir. Seam carving for content-aware image resizing. *ACM Trans on Graphics*, 26(3):10, 2007. 1

[5] K.Y. Chang, T.L. Liu, H.T. Chen, and S.H. Lai. Fusing generic objectness and visual saliency for salient object detection. In *CVPR*. IEEE, 2011. 2, 3, 6

[6] Ming-Ming Cheng, Guo-Xin Zhang, Niloy J Mitra, Xiaolei Huang, and Shi-Min Hu. Global contrast based salient region detection. In *CVPR*, 2011. 1, 2, 6

[7] Ian Endres and Derek Hoiem. Category independent object proposals. In *ECCV*, 2010. 1

[8] P.F. Felzenszwalb and D.P. Huttenlocher. Efficient graph-based image segmentation. *IJCV*, 59(2):167–181, 2004. 2, 3

[9] Stas Goferman, Lihi Zelnik-Manor, and Ayellet Tal. Context-aware saliency detection. *TPAMI*, 34(10):1915–1926, 2012. 6

[10] Jonathan Harel, Christof Koch, and Pietro Perona. Graph-based visual saliency. In *NIPS*, 2007. 6

[11] Geremy Heitz and Daphne Koller. Learning spatial context: Using stuff to find things. In *ECCV*, 2008. 1

[12] X. Hou, J. Harel, and C. Koch. Image signature: Highlighting sparse salient regions. *TPAMI*, 34(1):194–201, 2012. 1, 2, 6

[13] Laurent Itti, Christof Koch, and Ernst Niebur. A model of saliency-based visual attention for rapid scene analysis. *TPAMI*, 20(11):1254–1259, 1998. 1, 2, 6

[14] Huaizu Jiang, Jingdong Wang, Zejian Yuan, Yang Wu, Nanning Zheng, and Shipeng Li. Salient object detection: A discriminative regional feature integration approach. In *CVPR*, 2013. 7

[15] T. Judd, K. Ehinger, F. Durand, and A. Torralba. Learning to predict where humans look. In *ICCV*, 2009. 1, 2, 6

[16] Gunhee Kim, Eric P Xing, Li Fei-Fei, and Takeo Kanade. Distributed cosegmentation via submodular optimization on anisotropic diffusion. In *ICCV*, 2011. 2

[17] Dominik A Klein and Simone Frintrop. Center-surround divergence of feature statistics for salient object detection. In *ICCV*, 2011. 2

[18] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *NIPS*, pages 109–117, 2011. 2

[19] Alex Levinshtein, Adrian Stere, Kiriakos N Kutulakos, David J Fleet, Sven J Dickinson, and Kaleem Siddiqi. Turbopixels: Fast superpixels using geometric flows. *TPAMI*, 31(12):2290–2297, 2009. 3

[20] Tie Liu, Jian Sun, Nan-Ning Zheng, Xiaoou Tang, and Heung-Yeung Shum. Learning to detect a salient object. In *CVPR*, 2007. 2

[21] Yu-Fei Ma and Hong-Jiang Zhang. Contrast-based image attention analysis by using fuzzy growing. In *ACM MM*, 2003. 6

[22] Oded Maron and Tomás Lozano-Pérez. A framework for multiple-instance learning. In *NIPS*, 1998. 4

[23] Vidhya Navalpakkam and Laurent Itti. An integrated model of top-down and bottom-up attention for optimizing detection speed. In *CVPR*, 2006. 1

[24] Federico Perazzi, Philipp Krahenbuhl, Yael Pritch, and Alexander Hornung. Saliency filters: Contrast based filtering for salient region detection. In *CVPR*, 2012. 6

[25] R. Rahmani and S.A. Goldman. Missl: Multiple-instance semi-supervised learning. In *ICML*, 2006. 1, 4, 7

[26] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. Grabcut: Interactive foreground extraction using iterated graph cuts. *ACM Trans. Graphics*, 23(3):309–314, 2004. 2

[27] Ueli Rutishauser, Dirk Walther, Christof Koch, and Pietro Perona. Is bottom-up attention useful for object recognition? In *CVPR*, 2004. 1

[28] Xiaohui Shen and Ying Wu. A unified approach to salient object detection via low rank matrix recovery. In *CVPR*, 2012. 1, 2, 5, 6

[29] Roberto Valenti, Nicu Sebe, and Theo Gevers. Image saliency by isocentric curvedness and color. In *ICCV*, 2009. 2

[30] Yichen Wei, Fang Wen, Wangjiang Zhu, and Jian Sun. Geodesic saliency using background priors. In *ECCV*, 2012. 6