Multi-criteria Energy Minimization with Boundedness, Edge-density and Rarity, for Object Saliency in Natural Images

Sudeshna Roy
Indian Institute Technology, Madras, India
sudeshna@cse.iitm.ac.in

Sukhendu Das
Indian Institute Technology, Madras, India
sdas@iitm.ac.in

ABSTRACT

Recent methods of bottom-up salient object detection have attempted to either: (i) obtain a probability map with a 'contrast rarity' based functional, formed using low level cues; or (ii) Minimize an objective function, to detect the object. Most of these methods fail for complex, natural scenes, such as the PASCAL-VOC challenge dataset which contains images with diverse appearances, illumination conditions, multiple distracting objects and varying scene environments. We thus formulate a novel multi-criteria objective function which captures many dependencies and the scene structure for correct spatial propagation of low-level priors to perform salient object segmentation, in such cases. Our proposed formulation is based on CRF modeling where the minimization is performed using graph cut and the optimal parameters of the objective function are learned using a max-margin framework from the training set, without the use of class labels. Hence the method proposed is unsupervised, and works efficiently when compared to the very recent state-of-the art methods of saliency map detection and object proposals. Results, compared using F-measure and intersection-over-union scores, show that the proposed method exhibits superior performance in case of the complex PASCAL-VOC 2012 object segmentation dataset as well as the traditional MSRA-B saliency dataset.

Keywords

Saliency, Object Segmentation, Objectness, CRF

1. INTRODUCTION

Human visual system has an amazing capability to localize objects even before recognizing them. This comes from the ability to select regions with important visual information during early vision. This ability of human visual system is known as *Visual Saliency*. Again, cognitive science literature describes that spatial groupings of a small set of simple primitives give the early description of an image[1]. Localization of multiple objects in an image happens as a part of early visual processing. This indicates that saliency can be substantially utilized for localizing objects, thus imitating the human visual system. *Salient object segmentation* can then be successfully used as a pre-processing step to accomplish low-level tasks, e.g., shape-based feature extraction, and high-level vision tasks such as, object recognition, scene understanding, object tracking etc, as it reduces search space and minimizes information overload.

Category dependent object localization algorithms work only for a predefined set of objects and is practically infeasible given the huge number of classes existing in reality. Studies show that human beings can localize objects even when the identification or recognition system is impaired [2]. There has been thorough research in class specific object detection and localization. Sliding window approaches [3, 4] try to find objects at different windows at different scale and orientation. Therefore, these methods incur huge computational cost. Again, state-of-the-art segmentation methods [5, 6, 7] are not suitable to extract object specific image regions. So it is important to devise a system which can localize objects without any prior knowledge about it. Later, more features can be learned from these extracted object regions and recognition algorithms can also benefit from this spatial filtering. In this paper, we propose a Salient Object Segmentation method so that the same visual processing hierarchy as in humans can be employed by computer vision techniques. Our goal is to localize objects independent of its category.

Class independent object segmentation has recently gained importance in the Computer Vision community [8, 9]. Methods in this class typically give a bag of binary maps or masks where each map gives a region in the image so that each object is represented by at least one of these maps. Both CPMC [8] and Object Proposal [9] methods start with many seeds to predict a bag of masks as object proposals. Then, they rank order these maps based on precision of representing an object. Since, these methods give as many as few hundred maps, they give a very good recall. However their precision is very low, as a lot of background regions are proposed as objects. Optimization is based on intersection-over-union criteria [9] to rank the maps, but the results show that the top-most map generally contains almost half of the image (refer last row in Fig. 1). [10, 11, 12, 13] also address the problem of detecting generic objects, but give bounding boxes rather than pixel level segmentation output. In this work we concentrate on pixel accurate segmentation maps. [12] emphasizes on recall and does not intend for a pixel-

ICVGIP '14, December 14-18, 2014, Bangalore, India

http://dx.doi.org/10.1145/2683483.2683538

accurate map as they aim for object recognition.

On the other hand, saliency detection in images has been an area of research interest for long time. Researchers have taken two different approaches- fixation prediction and salient object segmentation. In fixation prediction, saliency is depicted as eye gaze fixation points [14, 15]. Whereas, salient object detection or segmentation methods give a pixel accurate saliency map where a pixel value expresses the probability of that pixel being salient [16, 17, 18, 19, 20, 21]. Although fixation prediction methods establish the fundamental principals of saliency detection, they are less suitable compared to saliency maps, for the object segmentation purpose. Recent saliency detection methods show high performance in saliency datasets, but they fail to perform when tested in natural image datasets like PASCAL [22]. There are two reasons behind this. First, these methods use only low-level perceptual cues such as, center surround operations [14], local and global contrast [17, 18], uniqueness and color distribution [19, 21] and boundary prior [20, 21, 23]. Second, there is typically a huge dataset bias which ensures the presence of only a single object at the center of an image. Moreover, in saliency datasets the objects are in high contrast with respect to the background. Hence, this class of methods do not scale up for more natural images such as in PASCAL segmentation dataset [22].

Our method of Salient Object Segmentation uses saliency feature and objectness criteria as two important cues to generate a single salient object segmentation map. The aim of the map is to depict all the object regions with high probability values. Natural images exhibit spatial interactions, e.g., neighboring pixels are likely to belong to the same object. Graph-based methods can capture these dependencies and do good spatial propagation of saliency information. Hence, we employ a graph-based approach and model our method as a conditional random field (CRF) based optimization approach. We perform all our processing at the superpixel level. To determine the superpixels of an image, we use the SLIC algorithm [24] which preserves primitive information like color, object boundary and edges. Since, number of superpixels are significantly less than that of pixels in an image, this makes our method fast and suitable as a pre-processor. We take account of low-level perceptual cues using saliency prediction method. Again, the objectness factors are incorporated based on geometric constraint and distribution of edges in the image. *Objectness*, as first defined by Alexe et al. [10], are the features that predict the likelihood of a superpixel belonging to any object. These in combination with saliency give the likelihood of each superpixel belonging to a salient object and forms the unary potential for CRF in the proposed method. Since, many objects are roughly homogeneous in appearance [9], CRF smoothness constraint gives a benefit. In following two sections, we first describe the image cues, namely saliency and objectness, followed by, our graphical model formulation, inference and learning. We have tested our method on challenging PASCAL 2012 dataset [22] and it shows better performance than saliency as well as object proposal methods, in terms of both F-measure and intersection-over-union scores.

2. IMAGE CUES

The aim of our approach is to segment all salient objects

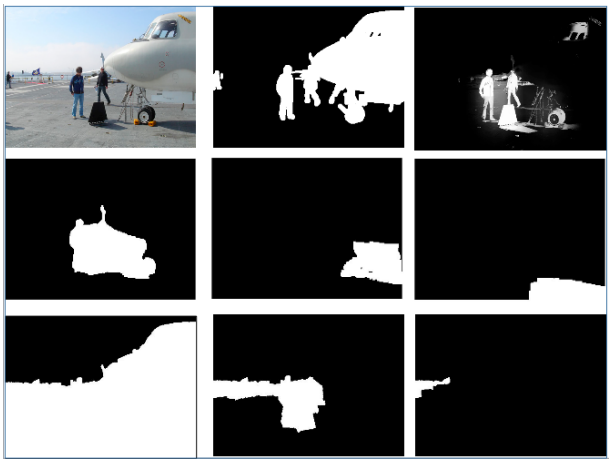

Figure 1: Examples of category independent models on a sample image from PASCAL VOC 2012 segmentation dataset [22]. From left to right, first row shows the image, its binarized ground truth and output of proposed method (refer Section 3). Second and third rows show top 3 ranked maps of CPMC [8] and Object Proposal [9] methods respectively.

in an image. Saliency methods alone produce a probabilistic saliency map. Therefore, to segment out the objects from an image, we use objectness criteria in conjunction with saliency. We characterize objectness by two constituent factors: geometric constraint and distribution of edges in the image. These two features are respectively modeled and termed by us as boundedness and edge-density. To find the image cues, an image is first segmented into a set of Nsuperpixels, $\{sp_i\}$, i = 1,...,N using the SLIC algorithm [24]. Then all the image cues are computed over superpixels as described in the following subsections.

2.1 Saliency as a Cue

Motivated by biological factors of human vision, as described in section 1, we employ saliency as a primary factor in our algorithm. Saliency detection methods mostly rely on low-level cues, such as, center-surround response [14], frequency domain features [26, 27, 28] or local and global contrast based information [17, 18, 19, 20, 21, 29]. All these methods try to find the rare or unique information in an image and represent that as salient. This kind of approach is known as feature rarity based approach. Authors in [21, 23] show that the feature rarity alone is not enough to describe the salient object. Therefore they introduce a background prior term. Background prior argues that most of the area in the boundary are occupied by non-salient regions and these regions are connected with each other. These are termed as boundary prior and connectedness prior and distance from boundary gives a measure of saliency. These methods in saliency literature have shown considerably improved performance in terms of localizing region belonging to objects.

All the methods discussed above are bottom-up or stimulus driven methods and do not exploit any prior information about any specific object. There are on the other hand, top-down saliency methods [30, 31] which learn the features of the objects it has to find and given that object class as an

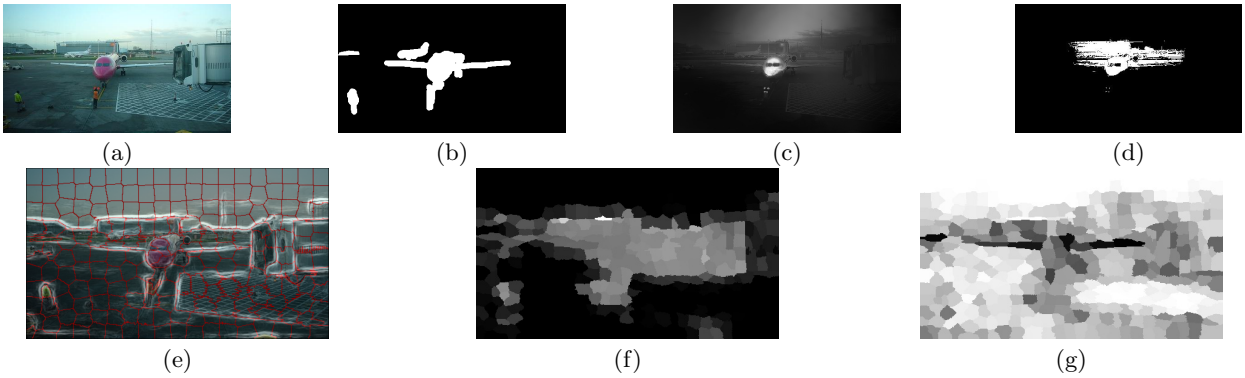

Figure 2: (a) Image; (b) it's binary ground truth and (c) saliency map obtained using the method proposed in [21]; (d) extracted airplanes by the proposed Salient Object Segmentation method. The bottom row illustrates the objectness factors: (e) the edge map [25] overlaid on superpixel level image and the two cues: (f) boundedness and (g) edge-density.

input, the object becomes salient for it [31]. However, similar to category dependent object segmentation models these methods cannot localize an object before recognizing it. So, they do not conform with our objective.

For the purpose of our salient object segmentation task we require a bottom-up saliency method which better predicts object regions employing both the feature rarity and boundary prior cues. Authors in [21] use these two factors effectively and in a time efficient way. Hence, we use the map produced by [21] as our saliency cue. The saliency detection method [21] produces a probabilistic saliency map of superpixels which is then upsampled to pixels. The superpixel saliency map in [21] is used as the saliency probability value of ith superpixel and denoted as s_i.

Figure 2(c) shows the saliency map of [21]. Clearly, the saliency map visually delineates that the rare features are depicted as salient. It uses compactness in color space and distinctness from boundary as the prominent cues for saliency. Hence, only the purple head of the airplane is filtered as salient. Wings on the other hand are completely ignored due to color similarity with sky on the top boundary. Also, white color may not be detected as a compact color in the image. Due to this the small white airplane in the background is also not identified. Moreover, partly the sky is predicted as salient which is not correct. Actually the work in [21] uses color information in great detail but shape and edge information are not exploited. As, human eye is most sensitive to color and brightness, the saliency method proposed in [21] performs good on saliency datasets (e.g., MSRA-B [26]), but fails when tested on natural image datasets such as PASCAL VOC dataset [22].

2.2 Objectness Features

Objects typically have well defined boundaries [10] and many objects are mostly homogeneous in appearance [9]. Superpixels preserve object boundaries as superpixel algorithms (e.g., [24]) group the pixels with homogeneous color and texture as a single superpixel. So, there should be no superpixel straddling by edges in an image [10]. Recently, Dollar and Zitnick have described an efficient edge detection algorithm in [25]. Their method gives a high-quality edge probability map using structured learning [32] prediction on random forest. Since, they do a direct inference, the method is computationally efficient than other edge detection techniques. We use this algorithm to generate an edge map (Fig. 2(e)). Next we compute the boundedness and edge-density factors.

2.2.1 *Boundedness*

Since there should be no superpixel straddling, we assume that the strong edges, the pixels with high edge probability value ($> T = 0.8$) in an edge map, mostly correspond to object boundaries. We define boundedness b_i of a superpixel sp_i based on the extent to which it is bounded by strong edges in all four directions. Boundedness is calculated for each pixel first and then averaged to superpixels. Edge contours on an edge map may be discontinuous and exist with small gaps. Due to this, some bounded pixels belonging to an object, may score low on boundedness. But most of the other pixels within the particular superpixel would have high boundedness score, if the superpixel belongs to an object. Hence, averaging over all the pixels makes it insensitive to noise in the edge contour. Moreover, as it can handle the discontinuities in the object boundary in an edge map, a computationally expensive high quality edge map [6] is not required.
Boundedness of superpixel sp_i is formulated as:

$$b_i = \frac{1}{|sp_i|} \sum_{p \in sp_i} (l_{p(x,y)} + t_{p(x,y)} + r_{p(x,y)} + d_{p(x,y)}) \cdot \mathcal{I}(l_{p(x,y)}, t_{p(x,y)}, r_{p(x,y)}, d_{p(x,y)}) \quad (1)$$

where, $l_{p(x,y)}, t_{p(x,y)}, r_{p(x,y)}, d_{p(x,y)} \in [0, 1]$ denote the strength of the left, top, right and bottom boundaries respectively, obtained from the edge probability map, of the particular pixel p with spatial location (x, y). $|sp_i|$ denotes the number of pixels in that particular superpixel. Also,

$$\mathcal{I}(l_{p(x,y)}, t_{p(x,y)}, r_{p(x,y)}, d_{p(x,y)}) = \begin{cases} 1, & if\ l_{p(x,y)}, t_{p(x,y)}, r_{p(x,y)}, d_{p(x,y)} > 0 \\ 0, & otherwise \end{cases} \quad (2)$$

is an indicator function and represents whether the pixel is close-bounded. The bounded pixels then gets the boundedness value dictated by the edge strength of it's boundaries.

The edge map gives a probability map where each pixel value denotes the strength of an edge passing through that particular pixel. Let the edge map value for a pixel at spatial location (x, y) be $\mathcal{P}e(x, y)$. We use a dynamic programming approach and compute the boundedness in order of number of pixels and recursively define the strength of the left boundary as:

$$l_{p(x,y)} = \begin{cases} l_{p(x-1,y)} & if\ \mathcal{P}e(x,y) < T \\ 0 & if\ x = 0 \\ \mathcal{P}e(x-1,y) & otherwise \end{cases} \quad (3)$$

Similarly, boundary strengths, $t_{p(x,y)}, r_{p(x,y)}, d_{p(x,y)}$ can be computed. For the whole image all the values of $l_{p(x,y)}$, $t_{p(x,y)}$, $r_{p(x,y)}$, $d_{p(x,y)}$ is computed only once in O(number of pixels) time. While calculating b_i, thus these values are accessed in O(1). High boundedness value implies that most of the pixels in the superpixel are bounded by strong edges, thus it is likely to belong to an object.

2.2.2 Edge Density

The distribution of edges in an image is captured as a cue using our edge-density term. Since, objects are mostly homogeneous in appearance [9], there should be fewer edges inside the superpixels belonging to an object. High density of edges in a region generally implies a cluttered background, e.g., rippling river, grass or forest. Again, very low density or overly smooth regions which are also not bounded should be part of background, e.g., clear sky. As no strong edge crosses over a superpixel, that is image boundaries are respected by superpixels, there should be only weak edges inside a superpixel. So, we compute the density of the edges within superpixel sp_i as:

$$density_i = \frac{1}{|sp_i|} \sum_{p(x,y) \in sp_i} \mathcal{P}e(x, y) \quad (4)$$

Now, we compute the mean μ_d and standard deviation σ_d of the set of densities, $\{density\}_i$, i = 1,...,N. Then for ith superpixel, the edge-density ed_i is calculated as:

$$ed_i = \begin{cases} 1 - density_i & if\ |\mu_d - density_i| < \sigma_d \\ 0 & otherwise \end{cases} \quad (5)$$

We have noticed, superpixels with high edge-density value are again have less probability for belonging to an object. Thus, we use it as a negative prior in our energy formulation as mentioned in the next section.

Bottom row of Figure 2 shows a superpixel-level image with the edge map superposed on it, along with boundedness cue and edge-density results on an example from PASCAL Segmentation dataset [22]. In Figure 2(e), red boxes show the superpixels and white lines portray the edge map. It can be seen that the closed strong edges mostly depict the object boundaries and these are mostly respected by superpixels as well. Exploiting the edge map we generate the boundedness map and edge-density map as illustrated in Figure 2(f) and 2(g) respectively.

3. SALIENT OBJECT SEGMENTATION

Conditional Random Fields (CRF) [33] has the ability to concisely represent dependencies among multiple random variables. Thus it can capture the structure of the problem efficiently. Hence, we formulate a CRF over superpixels to estimate the MAP (Maximum a Posteriori) value of each of them belonging to an object. Now each superpixel has the three features, as computed in the previous section, along with color. In the following subsections, we discuss the CRF formulation and the label prediction task.

3.1 Random Field Model

Let $\mathbf{x} = \{\mathbf{x}_i\}_{i=1}^N$ be the feature vector set and $\mathbf{y}$ be the segmentation labels of all the superpixels, where N is the total number of superpixels. Conditional random field (CRF) model takes the form:

$$P(\mathbf{y}|\mathbf{x}, \mathbf{w}) = \frac{1}{Z} e^{-E(\mathbf{y},\mathbf{x};\mathbf{w})} \quad (6)$$

where, $\mathbf{w}$ is a parameter vector and Z is the partition function. The energy term E generally decomposes over nodes $\mathcal{V}$ (set of superpixels) and edges $\mathcal{E}$ (8-neighborhood of each of the superpixel). We consider the energy E with node and edge features as $\phi^{(1)}$ and $\phi^{(2)}$ respectively, as:

$$\begin{aligned} E(\mathbf{y}, \mathbf{x}; \mathbf{w}) = & \sum_{i \in \mathcal{V}} \phi_i^{(1)}(y_i, \mathbf{x_i^{(1)}}; \mathbf{w_1}) \\ & + \lambda \sum_{(i,j) \in \mathcal{E}} \phi_{i,j}^{(2)}(y_i, y_j, \mathbf{x_i^{(2)}}, \mathbf{x_j^{(2)}}; \mathbf{w_2}) \end{aligned} \quad (7)$$

where, $\mathbf{w_1}$ and $\mathbf{w_2}$ are the parameters in node and edge potential and $\mathbf{x^{(1)}}$, $\mathbf{x^{(2)}}$ are the features contributing to unary and pairwise terms respectively. Node potentials are considered as negative log-likelihoods. So, first we compute different features, such as, saliency, boundedness, edge-density for the node potential or the unary term. Then we define the edge cost or the pair-wise smoothness term to fully specify the CRF.

3.2 Salient Object Likelihood

The node potentials are obtained by combining the image cues that are defined in Section 2. The image features for node potential of ith node (superpixel) is denoted by $\mathbf{x_i^{(1)}}$ and the parameters by $\mathbf{w_1} = [w_s\ w_b\ w_e]^T$. s_i, b_i and ed_i are the image features defined in section 2, used to form the $\mathbf{x_i^{(1)}}$. These features correspond to the likelihood of a superpixel being part of an object. So, it penalizes when a superpixel with large likelihood is assigned a background label or a superpixel with low likelihood is assigned a foreground label. Here a particular label $y_i \in \{0, 1\}$ is assigned for foreground and background pixels. Hence, the node potential is written as:

$$\begin{aligned} \phi_i^{(1)}(y_i, \mathbf{x_i^{(1)}}; \mathbf{w_1}) = & \underbrace{w_s((1-y_i)(1-s_i) + y_i s_i)}_{saliency} \\ & + \underbrace{w_b((1-y_i)(1-b_i) + y_i b_i)}_{boundedness} \\ & + \underbrace{w_e((1-y_i)ed_i + y_i(1-ed_i))}_{edge\ density} \end{aligned} \quad (8)$$

3.3 Edge Cost

Edge cost enforces agreement between adjacent superpixels in an image. If two adjacent superpixels are similar in appearance, they should have same label, otherwise the objective function is penalized. $\mathbf{x}_i^{(2)}$ is the feature that accounts for the pairwise term of ith superpixel. Here it is

color in Lab space, denoted by c_i. We express the pairwise or smoothness term as:

$$\phi^{(2)}_{i,j}(y_i, y_j, \mathbf{x_i^{(2)}}, \mathbf{x_j^{(2)}}) = |y_i - y_j| e^{-(-k_c||c_i - c_j||^2)} \quad (9)$$

Hence, similarity in Lab color space specifies the edge cost. Here, k_c dictates the sensitivity of color similarity and is given a constant value in all our experiments.

3.4 Superpixel Label Prediction: Inference Problem

Now that the random field is fully specified, we have two tasks, inference and parameter learning.

3.4.1 Inference

The edge-cost defined by our model leads to a sub-modular CRF and hence, we perform an exact inference using graph cut. This makes our method time efficient. With more complex edge-cost and approximate inference technique, the label prediction task becomes computationally inefficient. To improve the results, we perform the graph cut iteratively by exponentiating and normalizing the unary prior refined by the predicted labels. That is, we take a feedback from the first graph cut output and combine that with the unary term and again perform graph cut to generate a better segmentation. Experimentally, maximum of 8 iterations are performed for all the images.

3.4.2 Parameter Learning

As per our formulation, the parameter vector $\mathbf{w}$ can be considered as $\mathbf{w} = [\mathbf{w_1}\ \mathbf{w_2}]^T = [w_s\ w_b\ w_e\ \lambda]^T$. This gives an energy function E which is linear in $\mathbf{w}$ and can be written as $\mathbf{w}^T\phi$, where $\phi = [\phi^{(1)}\phi^{(2)}]$. This is important because linearity in $\mathbf{w}$ ensures a convex learning problem and thus can be solved efficiently. We take a simple max-margin approach [34] for parameter learning. The benefit of max-margin method is that, it takes into account how far a predicted label is from its ground truth, and the margin is adapted based on how much competing labelings differ from ground truth. The formulation is given as follows:

$$\begin{aligned} &\min_{\mathbf{w}} \frac{1}{2}||\mathbf{w}||^2 + \frac{1}{M}\sum_{n=1}^{M}\xi_n \\ &subject\ to,\ \ \mathbf{w}^T\mathbf{v} = 1 \\ &\qquad \mathbf{w}^T\phi_n - \mathbf{w}^T\hat{\phi}_n \le \mathcal{L}(\mathbf{y}_n, \hat{\mathbf{y}}_n) + \xi_n \\ &\qquad \xi_n \ge 0 \\ &\qquad w_2 > 0 \end{aligned} \quad (10)$$

where,
$\mathcal{L}(\mathbf{y}_n, \hat{\mathbf{y}}_n) = \frac{False\ Positive + False\ Negative}{True\ Positive + False\ Positive + False\ Negative}$

is the loss function. n ranges over the M training instances, ξ is the slack variable, $\hat{\mathbf{y}}$ is the ground truth labeling and $\mathbf{w}^T\hat{\phi}$ represents the ground truth energy. $\mathbf{v}$ is considered as $[1\ 1\ 1\ 0]^T$. The formulation is a structured learning approach. It is similar to the structured SVM approach [35] and follows the margin re-scaled algorithm as given by [34]. The first constraint normalizes the unary term. The second one gives a penalty to the objective function if the calculated parameters do not lead to an energy close to ground truth which ideally should have minimum energy. $w_2 > 0$ ensures the submodularity. Thus, minimizing this objective function leads to estimating the optimal parameters for which the energy for predicted labels is close to ground truth energy. To solve the optimization problem we start with an initial guess of w and minimize for $\mathbf{y}$, so that we can compute $\mathcal{L}$ using the grountruth labels $\hat{\mathbf{y}}$. Now given the loss function, the objective function as well as all the constraints are convex. We have used *cvx* toolbox [36] to solve the optimization problem.

After we perform inference, we obtain a binary map at superpixel level. This superpixel map can be thought as a low resolution image and we upsample it to full resolution pixel accurate map [37, 19]. Now each pixel has a label $\in [0, 1]$. Pixel label values depict the probability value of that pixel belonging to a salient object. The top right image in Figure 1 shows an example of our proposed upsampled map. The upsampling algorithm is also a fast implementation [19] and performs in linear time. We threshold this upsampled salient object probability map to generate a salient object segmentation mask. The threshold is taken as the median of maximum and minimum probability values. This method of producing the masks are used for all the experiments in Section 4 and presented in Figure 3.

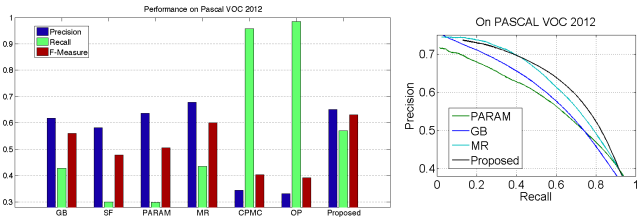

Figure 4: Comparative study of the performance of our proposed method using Precision-Recall and F-measure, on PASCAL VOC 2012 segmentation dataset [22]. The proposed method gives high precision even at high recall and outperforms all other methods in terms of f-measure.

4. EXPERIMENTS AND RESULTS

We measure the performance of our approach along with recent saliency detection methods as well as object proposal methods on both object segmentation dataset and saliency dataset. We evaluate using both F-measure and intersection-over-union score. All the results presented are generated using our optimal parameters, $\mathbf{w}$ learned from eqn. (10).

4.1 PASCAL Segmentation Dataset

We use the segmentation part of the PASCAL VOC 2012 dataset [22]. It has a segmentation part which has 2,913 images with object specific segmentation ground truth. To extract all the objects in an image we generate a binarized ground truth map as the second image in the top row of Figure 1 shows. We perform training on 1,464 images in the training set and testing is done on 1,449 images from the validation set.

Qualitative result on images from this dataset is presented in Figure 3. VOC 2012 has images from 20 different classes, Images from 12 different classes are shown in Figure 3 to illustrate the performance. The figure shows that our method performs better than the saliency methods, namely, SR [19], MR [20], PARAM [21]. Examples in first five rows clearly

Figure 3: Visual results of our Salient object segmentation method and three different saliency methods, on PASCAL VOC 2012 segmentation dataset [22], demonstrate the superiority of the proposed method. GT denotes the binarized ground truth mask. Last three rows depict the failure cases of all the methods.

depict the superiority of the salient object segmentation by the proposed method. All the methods fail to perform well on the samples in the last three rows. In case of the sample in the last row, for example, the object is not salient based on features used by saliency methods. In addition, as the edge map only captures the thick rod of the bi-cycle as strong edges, the boundedness and the edge-density cues also fail.

Figure 1 shows an example with category independent object proposal methods. The figure shows 3 top ranked masks of CPMC [8] and Object Proposal (OP) [9] along with our segmentation output. It is qualitatively visible that their precision is very low even for top ranked masks and same has been found in quantitative analysis in section 4.2.

4.2 F-measure and Intersection-over-Union Score

F-measure is defined as [19, 26],

$$Fmeasure = \frac{(1+\beta).Precision.Recall}{\beta.Precision + Recall}$$

This score is widely used in all saliency detection methods and β is taken as 0.3 [19, 26]. Figure 4 shows the precision-recall-fmeasure values on VOC 2012 segmentation dataset [22], for different competing saliency methods. We also compare with the category independent object proposal methods, viz, CPMC [8] and OP [9] (top 10 masks are taken). GB [16], SF [19], MR [20] and PARAM [21] are among the competing saliency methods we compare with. Clearly, in terms of F-measure our proposed method outperforms all the other methods. In terms of precision and recall also our method is comparable to very recent saliency methods [20, 21]. CPMC and OP give high recall values as they generate a number of maps, but are very low on precision. This implies that they propose a lot of non-object parts of an image as object regions, even when top 10 maps are considered (see Figures 1 and 4).

Intersection-over-Union (IoU) score is computed as,

$$IoU = \frac{|Predicted\ Map \cap Ground\ Truth|}{|Predicted\ Map \cup Ground\ Truth|}$$

We compute the IoU score of CPMC, OP and our method against the binarized ground truth on PASCAL segmentation dataset. Results are presented in Table 1 . We take the top ranked 10 maps of the method CPMC and OP to compute the IoU score considering all the objects. Table 1 shows that the single map of our method produces 21% better object segmentation maps in terms of IoU score, compared to the object proposal methods. Visual results in Figure 1 also illustrate the same.

4.3 Performance on Saliency Dataset

In order to show the efficiency of our proposed method, we compare the performance with recent state-of-the-art saliency techniques on a saliency dataset. For this purpose we use MSRA-B dataset, a part of which was first used by [26]. It has 5000 images with publicly available pixel accurate binary ground truth masks. Figure 5 shows the precision-recall-fmeasure plot on MSRA-B dataset for our method along with recent saliency methods, viz, RC [18], GB [16], HFT [28], SF [19], MR [20] and PARAM [21]. It illustrates our result on MSRA-B with the parameters learned

Method Name	IoU Score
CPMC	0.3319
OP (Object Proposal)	0.3266
Proposed	**0.4097**

Table 1: Intersection-over-Union score of top 10 object maps of category independent generic object segmentation methods, viz., CPMC [8], OP [9] and our proposed method of Salient object segmentation, on the PASCAL 2012 segmentation dataset.Our proposed method produces much better segmentation results.

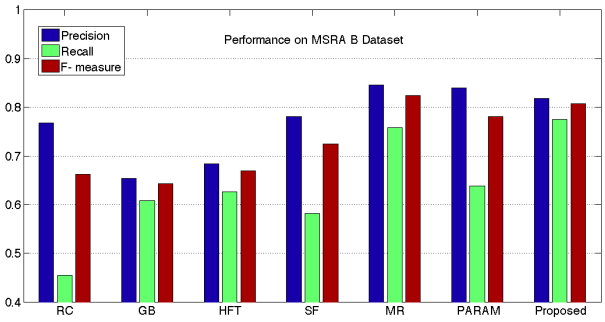

Figure 5: Precision, Recall, F-measure for proposed method and six other saliency methods on MSRA-B saliency dataset [26].

using PASCAL segmentation dataset. The results establish the dataset bias in saliency datasets, as explained in section 1. Recent saliency method MR [20] has shown to perform always with little higher precision. But it has very low recall when tested on natural image dataset like PASCAL segmentation dataset, i.e., it fails to extract many object regions and thus not suitable for current task.

4.4 Computational Efficiency

Our proposed method of salient object segmentation is deemed to work as a pixel-level pre-processing step for different computer vision tasks and must be suitable for a live system. Hence, computational efficiency is of real importance. However, since related category independent object proposal methods uses much complex procedure of generating a bag of outputs from different seeds and ranking them, they are less time efficient. We mainly have three components to be computed for the prediction task, viz., saliency, edge map and objectness cues. All of these happens in either order of pixel or even less (order of superpixels). The saliency [21] and edge map [25] extraction methods have described time efficiency in their paper. [25] is completely suitable for a live system. Also, [21] uses fast computations proposed by [19] and does all the operations at superpixel level and thus, is computationally efficient. Our computation of boundedness and edge-density is in order of number of pixels, compared to the objectness cues of [10] which are costly. Hence, our method is time efficient ($O(|superpixels|^3)$ which is even less than $|pixels|^2$) and is suitable as a precomputing technique.

5. CONCLUSIONS

We attempt to solve the problem of category independent salient object segmentation using a multi-criteria objective function. We propose a time efficient approach which performs better than recent state-of-the-art methods. Motivated by saliency and category independent object segmentation methods, we propose to predict a segmentation which captures all the salient objects in an image. We devise two objectness factors which are computed in linear time and used with saliency as the priors. We demonstrate that graph-based methods can be used efficiently both in terms of inference and learning parameters. Proposed method can be easily utilized as a pre-processing step for many high-level computer vision tasks.

6. REFERENCES

[1] A. Treisman and S. Gormican, "Feature analysis in early vision: evidence from search asymmetries." *Psychological review*, vol. 95, no. 1, p. 15, 1988.

[2] M. A. Goodale, A. D. Milner, L. Jakobson, and D. Carey, "A neurological dissociation between perceiving objects and grasping them," *Nature*, vol. 349, 1991.

[3] P. Viola and M. J. Jones, "Robust real-time face detection," *IJCV*, vol. 57, 2004.

[4] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *TPAMI*, vol. 32, 2010.

[5] J. Shi and J. Malik, "Normalized cuts and image segmentation," *TPAMI*, vol. 22, 2000.

[6] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik, "Contour detection and hierarchical image segmentation," *TPAMI*, vol. 33, 2011.

[7] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient graph-based image segmentation," *IJCV*, vol. 59, 2004.

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

[9] I. Endres and D. Hoiem, "Category independent object proposals," in *ECCV*, 2010.

[10] B. Alexe, T. Deselaers, and V. Ferrari, "Measuring the objectness of image windows," *TPAMI*, 2012.

[11] E. Rahtu, J. Kannala, and M. Blaschko, "Learning a category independent object detection cascade," in *ICCV*, 2011.

[12] K. E. A. van de Sande, J. R. R. Uijlings, T. Gevers, and A. W. M. Smeulders, "Segmentation as selective search for object recognition," in *ICCV*, 2011, pp. 1879–1886.

[13] M.-M. Cheng, Z. Zhang, W.-Y. Lin, and P. Torr, "Bing: Binarized normed gradients for objectness estimation at 300fps," in *CVPR*, 2014.

[14] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *TPAMI*, vol. 20, 1998.

[15] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *CVPR*, 2007.

[16] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *NIPS*, 2006, pp. 545–552.

[17] S. Goferman, L. Zelnik-manor, and A.Tal, "Context-aware saliency detection," in *CVPR*, 2010.

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

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

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

[21] S. Roy and S. Das, "Saliency detection in images using graph-based rarity, spatial compactness and background prior," in *VISAPP*, 2014.

[22] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL Visual Object Classes Challenge 2012 (VOC 2012) Results," http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html.

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

[24] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk, "SLIC Superpixels," EPFL, Tech. Rep., June 2010.

[25] P. Dollár and C. L. Zitnick, "Structured forests for fast edge detection," in *ICCV*, 2013.

[26] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *CVPR*, 2009.

[27] X. Hou, J. Harel, and C. Koch, "Image signature: Highlighting sparse salient regions," *TPAMI*, vol. 34, no. 1, pp. 194–201, 2012.

[28] J. Li, M. D. Levine, X. An, X. Xu, and H. He, "Visual saliency based on scale-space analysis in the frequency domain," *TPAMI*, vol. 35, 2013.

[29] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, and H. Shum, "Learning to detect a salient object," *TPAMI*, vol. 33, 2011.

[30] L. Zhang, M. H. Tong, and e. a. Marks, T. K., "Sun: A bayesian framework for saliency using natural statistics," *Journal of Vision*, vol. 8, 2008.

[31] J. Yang and M.-H. Yang, "Top-down visual saliency via joint crf and dictionary learning," in *CVPR*, 2012.

[32] S. Nowozin and C. H. Lampert, "Structured learning and prediction in computer vision," *Foundations and Trends in Computer Graphics and Vision*, 2011.

[33] J. D. Lafferty, A. McCallum, and F. C. N. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *ICML*, 2001.

[34] M. Szummer, P. Kohli, and D. Hoiem, "Learning crfs using graph cuts," in *ECCV*, 2008.

[35] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun, "Large margin methods for structured and interdependent output variables," in *JMLR*, 2005.

[36] M. Grant and S. Boyd, "Cvx: Matlab software for disciplined convex programming, version 2.0," http://cvxr.com/cvx, September 2013.

[37] J. Dolson, B. Jongmin, C. Plagemann, and S. Thrun, "Upsampling range data in dynamic environments," in *CVPR*, 2010.