Backtracking ScSPM Image Classifier for Weakly Supervised Top-down Saliency

Hisham Cholakkal Jubin Johnson Deepu Rajan
Multimedia Lab, School of Computer Science and Engineering
Nanyang Technological University Singapore
{hisham002, jubin001, asdrajan}@ntu.edu.sg

Abstract

Top-down saliency models produce a probability map that peaks at target locations specified by a task/goal such as object detection. They are usually trained in a supervised setting involving annotations of objects. We propose a weakly supervised top-down saliency framework using only binary labels that indicate the presence/absence of an object in an image. First, the probabilistic contribution of each image patch to the confidence of an ScSPM-based classifier produces a Reverse-ScSPM (R-ScSPM) saliency map. Neighborhood information is then incorporated through a contextual saliency map which is estimated using logistic regression learnt on patches having high R-ScSPM saliency. Both the saliency maps are combined to obtain the final saliency map. We evaluate the performance of the proposed weakly supervised top-down saliency and achieves comparable performance with fully supervised approaches. Experiments are carried out on 5 challenging datasets across 3 different applications.

1. Introduction

A saliency map can be thought of as a probability map in which the probabilities of pixels belonging to salient regions are mapped to intensities. It helps to reduce the search space for further processing, for example, in object segmentation. Bottom-up saliency aims to locate regions in an image that capture human fixations within first few milliseconds after the stimulus is presented [4, 14]. Here, feature contrast at a location plays the central role, with no regard to the notion of an object, although high-level concepts like face have been used in conjunction with visual cues like color and shape [23]. Lack of prior knowledge about the target in goal-oriented applications such as object detection and object class segmentation limits its utility. For example, the saliency maps produced by recent bottom-up approaches [10, 39] cannot discriminate between bus, person and bicycle in Fig. 1(b, c).

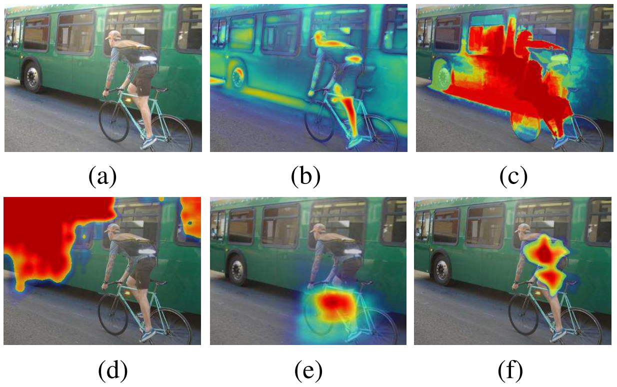


Figure 1. Comparison of our top-down saliency with bottom-up methods. (a) Input image, (b) saliency map of [10], (c) [39]; proposed top-down saliency maps for (d) bus (e) bicycle and (f) person categories.

On the contrary, top-down approaches utilize prior knowledge about the target for better estimation of saliency. The top-down saliency models produce a probability map that peaks at target/object locations [5, 36]. Our objective is to generate top-down saliency maps like those shown in Fig. 1(d, e, f). They were generated using the proposed method to identify probable image regions that belong to bus, bicycle and person separately.

Most methods for top-down saliency detection learn object classes in a fully supervised manner, where an exact object annotation is available. The learning component enables either discrimination between object classes or generation of saliency models of objects. Weakly supervised learning (WSL) alleviates the need for such user-intensive annotation by providing only class labels for an image during learning. The top-down saliency method using WSL [25] employs iterative refinement of object hypothesis on a training image. The proposed weakly supervised top-down saliency approach (Fig. 2 (e)) does not require these iterative steps, but produces a saliency map that is even comparable to fully supervised approaches as shown in Fig. 2(b, c, d).

In the proposed method, first, an ScSPM-based image classifier [37] is trained for an object category. On a val-

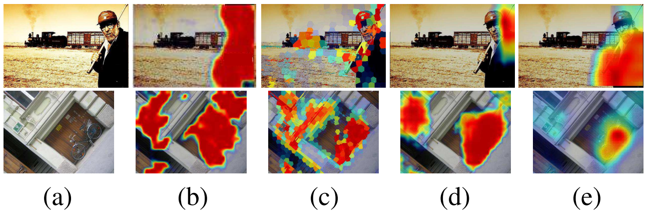


Figure 2. Our weakly supervised top-down saliency map in comparison with fully supervised methods. (a) Input images, person and bicycle saliency maps of (b) [36], (c) [20], (d) [5] and (e) proposed method are shown in row 1 and row 2 respectively.

idation/test image, the classifier gives a confidence score indicating the presence of the object. The probabilistic contribution of each patch in the image to this confidence score is analyzed to estimate its Reverse-ScSPM (R-ScSPM) saliency. The patches having high R-ScSPM saliency are generally from object regions, but they lack contextual information. For high-level understanding of the surrounding spatial region, contextual information of the patch is required. Hence, we incorporate a contextual saliency module that computes the probability of object presence in a patch using logistic regression trained on contextual max-pooled vectors [5]. The training of contextual saliency needs a set of positive patches from the object region and a set of random negative patches from images that do not contain the object. Since a patch-level annotation is not available, we use patches from the positive training images having high R-ScSPM saliency to train the contextual saliency. The contextual saliency inferred on a test image is combined with the R-ScSPM saliency to form the final saliency map. R-ScSPM saliency considers the spatial location of patch through backtracking max-pooled vector whereas contextual saliency considers its spatial neighborhood information, thereby complementing one another. We also propose a classifier confidence-based refinement to the saliency map. Besides illustrating the accuracy of saliency maps produced by the proposed method, we demonstrate its effectiveness in applications like weakly supervised object annotation and class segmentation.

2. Related Work

Kanan *et al.* [17] proposed a top-down saliency approach which uses object appearance cues along with location prior. It fails if the object appears at random locations. Closer to our framework, [36] proposed a fully supervised top-down saliency model that jointly learns a conditional random field (CRF) and dictionary using sparse codes of SIFT features as latent variables. The inability to discriminate similar objects (*e.g.* car and bike) and lack of contextual information causes a large number of false detections. [20] improves upon this by considering the first and second order statistics of color, edge orientation and pixel location within a superpixel, along with objectness [3] instead of SIFT features. Khan and Tappan [18] use label and location-dependent smoothness constraint in a sparse code formulation to improve the pixel-level accuracy of [36] at the expense of increased computational complexity. Zhu *et al.* [41] proposed a contextual pooling based approach where LLC [35] codes of SIFT features are max-pooled in a local neighborhood followed by log-linear model learning. By replacing LLC codes with locality-constrained contextual sparse coding (LCCSC), [5] improves [41] with a carefully chosen category-specific dictionary learned from the annotated object area. The proposed method uses a smaller dictionary which is not category-specific. Discriminative models [38, 32] often represent a few patches on the object as salient and not the entire object. Hence, such models end up with low recall rates as compared to generative models [20, 36]. The proposed framework addresses this using contextual saliency. In [32], the task of image classification is improved using discriminative spatial saliency to weight visual features. A fully supervised CNN-based approach [40] achieves high accuracy in category-independent saliency datasets such as PASCAL-S [22] by training on large datasets such as ImageNet [30], which is known to be computationally intense.

The use of weak supervision in top-down saliency has largely been left unexamined. DSD [12] uses a weakly supervised setting where bottom-up features are combined with discriminative features that maximize the mutual information to the category label. The lack of contextual feature information limits its performance in images containing background clutter. In [25], a joint framework using classifier and top-down saliency is used for object categorization by sampling representative windows containing the object. Their iterative strategy leads to inaccurate saliency estimation if the initialized windows do not contain the object.

Since we demonstrate the usefulness of our saliency maps for object annotation and class segmentation, we review closely related works in these areas. The class segmentation approach of [11] uses superpixels as the basic unit to build a classifier using histogram of local features within each superpixel. Aldavert *et al.* [2] computes multiclass object segmentation of an image through an integral linear classifier. A much larger codebook of 500,000 atoms is used compared to only 1536 atoms in our framework. Training of shape mask [24] requires images to be marked as *difficult* or *truncated* in addition to the object annotation. The proposed method produces better results by just using a binary label indicating the presence or absence of object of interest. Co-segmentation approaches [29, 13, 15, 16, 19] segment out the common objects among a set of input images. Semantic object selection [1] collects images with white background from the internet using tag-based image retrieval. A saliency model is proposed in [33] to address the object annotation task [26, 34].

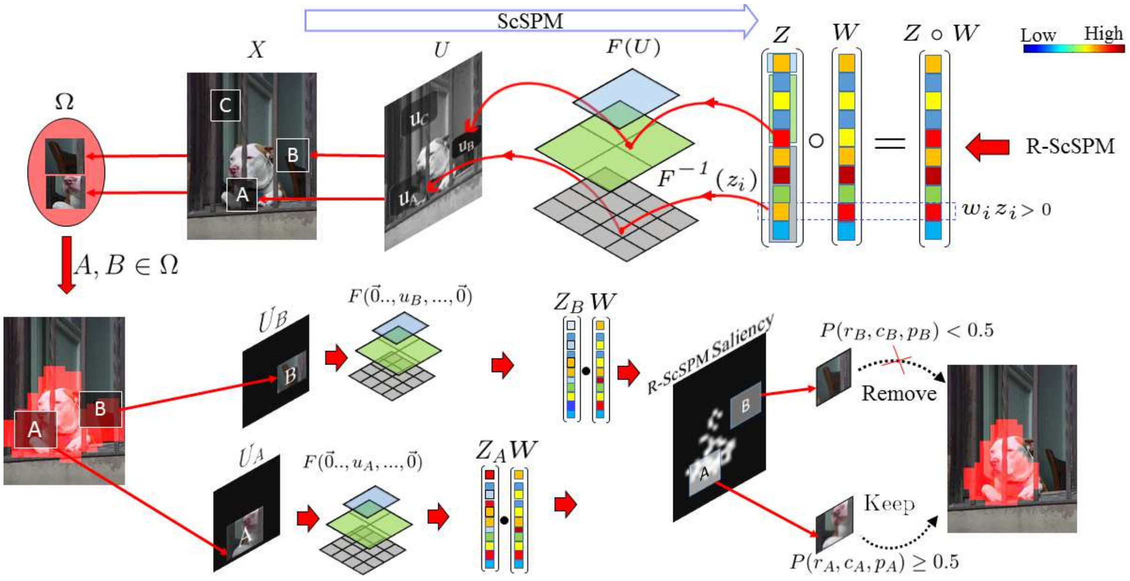


Figure 3. Illustration of our R-ScSPM saliency estimation and patch selection for *dog* category. Red arrows indicate the proposed R-ScSPM framework. The elements z_i of Z having $(w_i z_i > 0)$ are traced back to the image patches A, B and are added to Ω. The patch $C \notin \Omega$ as it does not contribute positively to classifier confidence. For a patch $A \in \Omega$, R-ScSPM saliency $P(r_A, c_A, p_A)$ is evaluated by setting the sparse codes of all patches except A to $\vec{0}$ forming U_A followed by a scalar product $(\cdot)$ with the classifier weight W. Similar procedure is followed for all patches in Ω. The patch A is selected as object patch since $P(r_A, c_A, p_A) \geq 0.5$.

3. Proposed method

In this section, we first present the weakly supervised R-ScSPM framework (Fig. 3) to obtain R-ScSPM saliency. We then introduce contextual saliency (Sec. 3.2) that estimates object presence in a patch by considering its neighborhood information. Training of contextual saliency requires object patches that are selected using the R-ScSPM saliency map. Finally, during inference (Sec. 3.3), our framework combines both the saliency maps to generate the final saliency map.

3.1. R-ScSPM saliency

3.1.1 Our ScSPM implementation and notations

In our ScSPM-based classifier, dense-SIFT features are extracted from gray-scale image patches. K-means clustering of the SIFT features from training images are used to form a dictionary D of d elements (atoms). The SIFT features $X = [x_1, x_2 ... x_M]$ from M patches of an image are sparse coded using D to $U = [u_1, u_2 ... u_m, ... u_M]$. Here, u_m is a d-dimensional vector representing the sparse code of a feature x_m from the m^{th} image patch. The spatial distribution of the features in the image is encoded in the max-pooled image vector Z through a multi-scale max-pooling operation $F(u_1, u_2, ..., u_M)$ of the sparse codes on a 3-level spatial pyramid [21] as shown in Fig. 3. The i^{th} element z_i of Z is a max-pooled value derived using maximum operation on j^{th} elements of all patches in a spatial region $\mathcal{R}$ defined by i and $j = 1 + (i - 1) \ mod \ d$. It is represented as

$$z_i = max\{|u_{1j}|, |u_{2j}|, ...|u_{qj}|\}, \quad s.t. \quad 1, 2...q \in \mathcal{R}. \quad (1)$$

Let the label $Y_k \in \{1, -1\}$ indicate the presence or absence of an object O in the k^{th} image. If $Y_k = 1$ it is a positive image, else it is a negative image. Image-label pairs (Z_k, Y_k) of L training images are used to train a linear binary SVM classifier by minimizing following objective function [8]

$$\arg\min_{W} \|W\|^2 + \mathcal{C} \sum_{k=1}^{L} max(0, 1 - Y_k(W^\top Z_k + bias)), \quad (2)$$

where $W = [w_1, w_2 .... w_N]^\top$ and $bias$ are the SVM weight vector and bias respectively. W is learnt separately for each object category. N is the length of the max-pooled image vector Z_k and $\mathcal{C}$ is a constant. Given a validation/test image with max-pooled vector Z, the classifier score $W^\top Z + bias$ indicates the confidence of the presence of object O in it.

3.1.2 R-ScSPM saliency formulation

In an ScSPM image classifier, both the linear-SVM and multi-scale max-pooling operations can be traced back to the patch level. This enables us to analyze the contribution of each patch towards the final classifier score which is then utilized to generate the R-ScSPM saliency map for an object. Since a common dictionary D is learned for all objects by unsupervised clustering of random SIFT features from training images, the correspondence of a particular dictionary atom to an object or background is unknown. ScSPM stipulates that a patch is representative of its image if its sparse code makes the largest contribution (max-pooling) to a particular dictionary atom among other patches in the same spatial region $\mathcal{R}$. The representativeness, r_m, of a patch m for an image is indicated by the number of times the elements of that patch's sparse code made it to the max-pooled vector. Representative patches may either contribute positively or negatively to the classifier score with higher contribution indicating more relevance of the patch to an object O. The relevance of the patch to the object is denoted c_m.

It is possible that among the elements of the sparse code of a patch that contributes positively to the classifier confidence, there are other elements that may contribute negatively. For example, let $[u_{m1}, 0, ... \; u_{mj} \; ...0, u_{md}]^\top$ be the sparse code of a patch m with its j^{th} element u_{mj} being a local maximum in its spatial pyramid region. Although u_{mj} contributes positively to the classifier confidence $W^\top Z + bias$, the other non-zero elements u_{m1} or u_{md} may contribute negatively, indicating absence of the object in that patch. So, the relevance of a patch to the object requires its contribution to be computed in the absence of other patches; this relevance is denoted p_m. The probability of a patch m belonging to an object, which in turn indicates the saliency G of the object, depends on the three parameters–r_m, c_m and p_m as

$$G = P(r_m, c_m, p_m) = P(p_m|r_m, c_m)P(c_m|r_m)P(r_m). \tag{3}$$

The representative elements of the sparse code u_m is identified by

$$\Psi_m = \{i\delta(F^{-1}(z_i), u_{mj})\}, \quad \forall i \in \{1, 2, ..N\}, \tag{4}$$

where δ is the Kronecker delta function and F^{-1} is the inverse operation of spatial pyramid max-pooling and the location of z_i in Z identifies the region $\mathcal{R}$ in the spatial pyramid and its position j in the sparse code u_m. The probability of representativeness of the m^{th} patch to the image is then defined as

$$P(r_m) = \mathbf{card}(\Psi_m)/N, \tag{5}$$

where $\mathbf{card}(.)$ represents cardinality and N is the length of Z.

The classifier confidence is a score indicating the presence of the object in the image, which proportionally increases from a definite absence ($score \leq -1$) to definite presence ($score \geq 1$). Normalizing the confidence scores between 0 and 1 using parameters $\beta = 0.5$ and $b = \beta(bias + 1)$, we can represent it as the probability

$$\begin{aligned}
&P(Y = 1|F(u_1, u_2, ..., u_M)) \\
&= \beta W^\top F(u_1, u_2, ..., u_M) + b, \\
&= \beta W^\top Z + b = \beta \sum_{\forall i \in \{1,..N\}} w_i z_i + b, \\
&= \beta \sum_{\forall i \in \Psi_m} w_i z_i \;\; + \;\; \beta \sum_{\forall i \in \{1,..N\} \setminus \Psi_m} w_i z_i + b, \\
&= f(c_m|r_m) \;\; + \;\; \beta \sum_{\forall i \in \{1,..N\} \setminus \Psi_m} w_i z_i + b,
\end{aligned}$$

where $f(c_m|r_m)$ is the contribution of the patch m to the image classifier confidence.

Given that the patch is representative of the image, the probability of it belonging to the object is

$$P(c_m|r_m) = \begin{cases} f(c_m|r_m), & \text{if} \quad f(c_m|r_m) \geq 0, \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

Using the above probabilities, we select a set Ω of all patches that contribute positively to the classifier confidence as

$$\Omega = \{P(c_t|r_t)P(r_t) > 0\}, \quad \forall t = 1, 2, ..., \; M. \tag{7}$$

The net contribution of a patch $m \in \Omega$ in the absence of other patches is

$$P(p_m|r_m, c_m) = \beta W^\top F(\vec{0}.., u_m, ..., \vec{0}) + b, \tag{8}$$

where $F(\vec{0}.., u_m, ..., \vec{0})$ is the max-pooling operation performed by replacing the sparse codes of all other patches with a zero vector $\vec{0}$ of size d to form a max-pooled vector Z_m.

Implementation details. Fig. 3 illustrates three patches A, B and C on an image and their corresponding sparse codes. The classifier score $W^\top Z + bias$ indicates the confidence of the presence of object as mentioned earlier. Each element z_i of Z has a corresponding weight w_i. The elements from the Hadamard product $W \circ Z$ with $w_i z_i > 0$ mark the patches A and B that contribute positively to the classifier confidence through a $F^{-1}(.)$ operation, i.e the set Ω. The contribution of patch A in the absence of other patches is evaluated using max-pooling operations $F(\vec{0}.., u_A, .., \vec{0})$ on sparse code vectors U_A in which sparse codes of all other patches except u_A are replaced with $\vec{0}$ forming max-pooled vector Z_A. The R-ScSPM saliency of a patch m is given by

$$P(r_A, c_A, p_A) = \begin{cases} \beta W^\top F(\vec{0}.., u_A, .., \vec{0}) + b & \text{if } A \in \Omega, \\ 0 & \text{otherwise.} \end{cases} \tag{9}$$

3.2. Contextual saliency training

The purpose of contextual saliency is to include neighborhood information of a patch. Previous top-down saliency approaches [5, 41] use a fully supervised setting to select object patches to train their contextual saliency module. In our approach, we remove this requirement by using object patches that are extracted by R-ScSPM saliency. From positive training images, patches with R-ScSPM saliency $G > 0.5$ are selected as positive patches with label $l = +1$, while random patches are selected from negative images with patch label $l = -1$. The selected patches are indicative of belongingness to an object category. In fig 3, patch A having $P(r_A, c_A, p_A) \geq 0.5$ is selected for contextual model training while patch B having $P(r_B, c_B, p_B) < 0.5$ is removed.

For the selected patches, a 13×13 neighborhood of surrounding patches are divided into a 3×3 spatial grid followed by max pooling of sparse codes over each grid and concatenated to form a context max pooled vector ρ. A logistic regression model with weight v and bias b_v is learned using positive and negative patches from the training images to form the contextual saliency model [5]. Since the sparse codes for every patch is already computed for R-ScSPM, max pooling over the context of a patch followed by logistic regression learning is the only additional computation required for this contextual saliency model.

3.3. Saliency Inference

On a test image, the contextual saliency L is inferred using the logistic regression by

$$P(l = 1 \mid \rho, v) = \frac{1}{1 + exp(-(v^T \rho + b_v)}, \quad (10)$$

where $P(l = 1 \mid \rho, v)$ indicates the probability of presence of an object in a patch and ρ is the contextual max-pooled vector for a test patch. For each patch, the contextual and R-ScSPM saliency values are combined as $GL + 0.5(G + L)$ and normalized to values in $[0, 1]$ to form the saliency S. We choose this combination criteria instead of a product between the two, since the R-ScPSM saliency values are non-zero only for R-ScSPM selected patches.

Classifier-based refinement. The saliency map is refined using the same image classifier used for R-ScSPM saliency having SVM parameters $(W, bias)$. Given a test image, we compute its classifier confidence $W^\top Z + bias$. The test image could either contain a single class or multiple classes. For the former, as in the Graz-02 dataset, the classifier estimates the presence or absence of an object. However, for multiple classes, thresholding of classifier confidence determines the presence or absence of an object. During training, we compute the average classifier confidences $W^\top Z_j$ for all positive training images j in each run of K-fold learning (K=15). The mean of K such values is used as the final threshold th_O to indicate the presence of object O on a test image. To avoid situations where false negative values drastically reduce the threshold, we maintain the lowest possible confidence value as -0.5. If $W^\top Z < th_O$, it is less probable that the object O is present in that image, and therefore, there will be no salient object marked in the image. However, if $W^\top Z > th_O$, the saliency of the patch is retained as S. Pixel-level saliency maps are generated from patch-level saliency S using Gaussian-weighted interpolation as in [5].

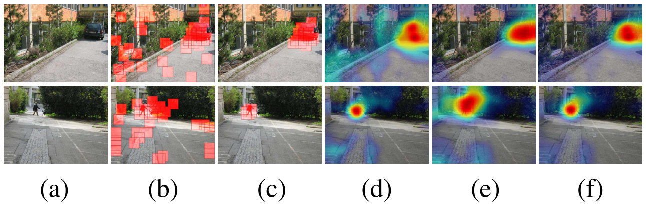


Figure 4. Illustration of individual stages of the proposed model. (a) Input image, (b) patches in Ω and (c) patches selected by thresholding (d) R-ScSPM saliency map. (e) contextual saliency map and (f) final saliency map.

4. Experimental evaluation

We evaluate the performance of our weakly supervised top-down saliency model on 5 challenging datasets across three applications. We compare with other top-down saliency approaches using Graz-02 [27] and PASCAL VOC-07 [6] segmentation datasets. The top-down saliency map is applied to the tasks of class segmentation, object annotation and action-specific patch discovery on Object Discovery dataset [29], PASCAL VOC-07 detection dataset and PASCAL VOC-2010 action dataset [7] respectively. All these datasets are challenging, especially from a weakly supervised training perspective, due to heavy background clutter, occlusion and viewpoint variation.

We maintain the same parameters across the datasets. Following [36], SIFT features are extracted from 64×64 patches on a grayscale image with grid spacing of 16 pixels. The dictionary size for sparse coding is set to 1536 disregarding individual object categories whereas in [36, 20] separate dictionaries of size 512, corresponding to each object category are iteratively learned. The size of the context-pooled vector is $9 \times 1536 = 13824$.

4.1. Analysis of individual components

Fig. 4 shows a visual comparison of the effect of each stage in our proposed method. For a test image, the patches in Ω (refer Fig. 3) from the R-ScSPM pipeline are shown in Fig. 4(b) and the R-ScSPM saliency map is shown in Fig. 4(d). This saliency map is thresholded at 0.5 to obtain the most relevant patches weeding out the false detection in Ω as shown in Fig. 4(c). Fig. 4(f) shows the final saliency map formed by combining the contextual (Fig. 4(e)) and R-ScSPM saliency maps. At non-textured patches of the car (top), the lower R-ScSPM saliency is boosted by high con-

Table 1. Components analysis : Pixel-level precision rates at EER (%).

	Bike	Car	Person	Mean
Random trained contextual saliency	51.2	27.3	38.3	39
R-ScSPM trained contextual saliency	66.9	55.3	54.5	58.9
R-ScSPM saliency	61.6	46.6	54.8	54.3
Complete (R-ScSPM trained contextual saliency + R-ScSPM saliency)	67.5	56.48	57.56	60.52

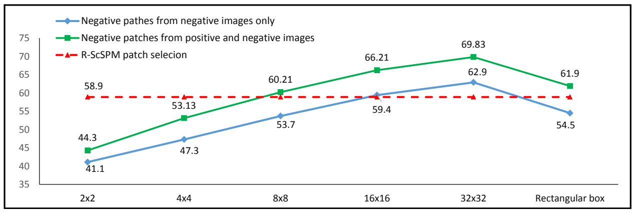


Figure 5. Effect of patch selection strategy for training. X-axis specifies the supervision settings, Y-axis denotes the mean of precision at EER (%) across 3 categories.

textual saliency in the final saliency map. The smearing of saliency in the contextual saliency map for small objects (bottom) is removed when combined with the R-ScSPM saliency map.

Table 1 analyzes the contribution of each component of the proposed saliency model on Graz-02 dataset (dataset details are in Sec. 4.3). The effectiveness of the proposed method in selecting positive patches is demonstrated by comparing its performance to that using random selection of patches from positive images in training contextual saliency. The mean precision rate at EER (%) of 39% is much lower than 60.52% obtained using the complete framework. This can be attributed to poor model learning in categories like car where the object size could be much smaller relative to the image, whereby random selections are more probable to pick out patches from the background. By training the contextual model with the R-ScSPM selected patches, the result improved to 58.9%, which shows that the R-ScSPM patch selection is effective in localizing object patches in an unannotated positive image. The contribution of R-ScSPM saliency is studied by removing the contextual saliency component from the framework. The results are poorer compared to 'R-ScSPM trained contextual saliency' due to lack of contextual information. Our complete framework gives 60.52% which shows that both the contextual and R-ScSPM saliency maps complement one another. R-ScSPM utilizes the patch location and contextual saliency utilizes its neighborhood information and hence they complement each other.

4.2. Comparison with various levels of supervision

Previous WSL localization and top-down saliency works [31, 25] select initial negative patches from either the boundaries or at random locations from the postive training images. They need to iteratively refine their model in order to remove potentially erroneous negative patches. Since the training of the proposed method is not iterative, we need to select negative patches only from negative images. We analyze the influence of negative patches extracted from positive images on the performance of contextual saliency using different supervision settings in Graz-02. Each positive training image is divided into regular sized grids varying from 2 × 2 to 32 × 32 and each grid is manually labeled to indicate if an object is present or not. We also consider the case of a rectangular bounding box around the object. The contextual saliency model is learned using the additional label information. We maintain the same number of positive and negative patches throughout the experiment. Each category's model is evaluated on its respective test images. Pixel-level precision rates at EER (%) is averaged over all categories and shown in Fig. 5. The model trained using negative patches from both positive and negative images (green) outperforms the result when only negative images are used (blue), with the performance increasing with increasing scale of supervision. It indicates that if an iterative learning is used in our method, the results can be improved considerably. The proposed weakly supervised method (red) matches the performance of the 16 × 16 supervised setting (a label for every 40 × 40 pixels) learned using negative patches from negative images despite having the label at the image level. We outperform the results of a labeled bounding box using the same learning settings.

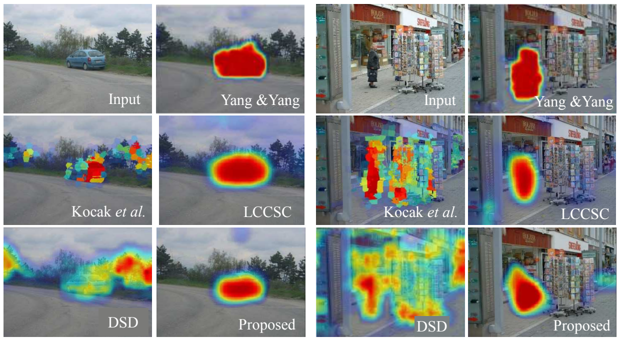


Figure 6. Comparison of the proposed weakly supervised method with other fully supervised (Yang&Yang [36] ,Kocak *et al.* [20], LCCSC [5]) and weakly supervised (DSD [12]) top-down saliency approaches on car and person images.

4.3. Graz-02 dataset

Graz dataset contains 3 object categories and a background category with 300 images per category. We split the images into training and testing sets following [36]. We report our results on 3 test set configurations. First, pixel-level results of the proposed saliency model and recent top-down saliency models [36, 20, 5] are evaluated on all 600 test images of the dataset. Second, for comparison with related approaches [24, 11], each object category is evaluated on test images from its respective category and the pixel-level results are reported. Finally, to compare with [12, 17], the patch-level results on 300 test images [36] is evaluated,

Table 2. Pixel-level precision rates at EER (%) on Graz-02.

Method	SV	**Test set**	**Bike**	**Car**	**Person**	**Mean**
1 - Zhang *et al.* [39]	US	All test images	31.77	18.66	30.71	27.1
2 - Yang and Yang [36]	FS		59.4	47.4	49.8	52.2
3 - Kocak *et al.* [20]	FS		59.92	45.18	51.52	52.21
4 - LCCSC [5]	FS		69.07	58.39	58.22	61.89
5 - Proposed WS	WS		63.96	45.11	55.21	54.76
6 - FS version	FS		71.5	56.6	62.3	63.51
7 - Zhang *et al.* [39]	US	Test images from respective category	54.67	39.03	52.04	48.58
8 - Aldavert *et al.* [2]	FS		71.9	64.9	58.6	65.13
9 - Fulkerson *et al.* [11]	FS		72.2	72.2	66.1	70.16
10 - Shape mask [24]	FS		61.8	53.8	44.1	53.23
11 - Yang and Yang [36]	FS		62.4	60	62	61.33
12 - Khan and Tappen [18]	FS		72.1	-	-	-
13 - Kocak *et al.* [20]	FS		73.9	68.4	68.2	70.16
14 - LCCSC [5]	FS		76.19	71.2	64.13	70.49
15 - Proposed WS	WS		67.5	56.48	57.56	60.52
16 - FS version	FS		77.61	71.91	66.95	72.16

where 150 test images are from a single category and the remaining 150 are from the background class.

Table 2 compares our pixel-level results with recent top-down saliency approaches [5, 36, 20] and related object segmentation and localization approaches [2, 24]. SV indicates supervision level with US, WS, FS referring to unsupervised, weakly supervised and fully supervised training respectively. [36] and [20] are fully supervised (FS), needing 20 iterations of CRF learning with sparse codes relearned at each iteration. Separate dictionaries are used for each object category. On the contrary, our weakly supervised method does not require any iterative learning and sparse codes are computed just once on a single dictionary. [5] uses a larger dictionary of 2048 atoms for Graz-02 as compared to 1536 atoms in our approach. When their model is evaluated on the entire 600 test images, the mean precision at EER is 52.2% and 52.21% respectively. The discriminative capability of [20] does not improve by incorporating objectness [3] and superpixel features to [36]. The proposed method achieves 54.76% with better discrimination against objects of other categories in a weakly supervised setting. [24, 2] reports results in which each model is tested on images from its own category. Pixel-level results of our proposed model is evaluated using same setting. It is seen that [20] is better (row 13) than proposed weakly supervised approach (row 15), however is inferior to our model (row 5) in removing false positives (row 3). [2] uses 500,000 dictionary atoms in their fully supervised framework to produce 65.13% accuracy as compared to 60.52% in our weakly supervised approach that uses only 1536 atoms. Our results are far superior compared to the fully supervised shape mask [24]. As expected, recent bottom-up saliency model [39] produces poor performance when compared to our result.

For fair comparison with fully supervised approaches, we report the results of our model in a fully supervised setting as well (FS version), i.e. the contextual saliency model is trained on object patches from training images using patch-level object annotations as in [36], instead of R-ScSPM. With this supervised setting, our model achieves state-of-the art results in top-down saliency.

Table 3. Patch-level precision rates at EER (%) on 300 test images.

	Bike	Car	Person	Mean
DSD [12]	62.5	37.6	48.2	49.4
SUN [17]	61.9	45.7	52.2	53.27
Proposed WS	76.0	53.7	66.7	65.43

Table 4. Patch-level precision rates at EER(%) on PASCAL VOC-07.

Method	Yang and Yang [36] (FS)	LCCSC [5] (FS)	Proposed (WS)
Mean of 20 classes	16.7	23.4	18.6

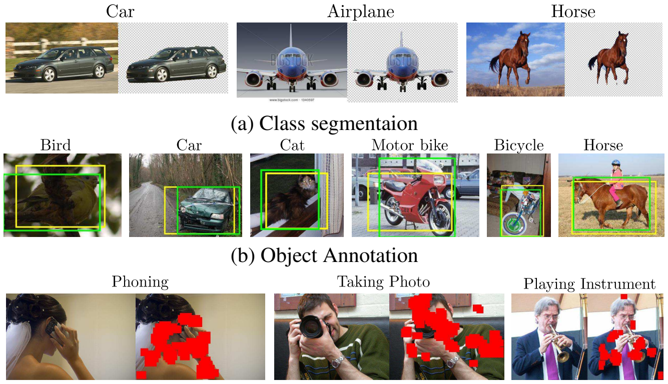
Table 3 compares the patch-level precision at EER of the proposed saliency model on 300 test images with other representative patch-level methods. As evident from Fig. 6, DSD [12] has limited capability to remove background clutter, resulting in poor performance of their model. Feature learning using independent component analysis helped SUN [17] to perform better than DSD, but substantially poorer than the proposed method.

4.4. PASCAL VOC-07 segmentation dataset

Following [5, 20, 36], training uses object detection images and testing is performed on 210 segmentation test images. Also, to reduce the computational complexity of sparse coding, a common dictionary of 1536 atoms is used for all object classes, which is much smaller than (20×512) atoms of [5] . For each object category, separate sparse codes are computed in [36, 20].

Table. 4 shows the patch-level performance comparison between the proposed WS method and FS top-down saliency approaches [36, 5]. Category-level results and comparisons are available in the supplementary material. Knowledge about object presence inferred from the classifier refinement helped the saliency map to outperform [36, 5] in classes like *aeroplane* and *train*. However, the use of a fixed context neighborhood size and lack of object annotation limits the performance in smaller objects like *bottle* and *bird*. Our method outperforms the fully supervised approach [36] and achieves a higher mean precision rates at EER (%) computed over all 20 classes.

Khan and Tappen [18] report their pixel-level precision rates at EER only for cow category (8.5%) which is lower than the proposed weakly supervised approach (9.7%). We did not compare with [20] since they manually assign an all zero map if the object of interest is absent [5]. By simple thresholding at EER, the proposed saliency maps outperform segmentation approaches [11, 2]. The results are available in the supplementary material. The presence of multiple, visually similar object classes in a single image is challenging for a weakly supervised approach, yet we achieve patch-level precision rate at EER comparable to that of state-of-the art fully supervised approach [5].



(a) Class segmentaion

(b) Object Annotation

(c) Action specific patch selection

Figure 7. Applications of the proposed saliency maps for (a) class segmentation, (b) object annotation and (c) action-specific patch selection.

4.5. Computation time

Training of the proposed framework is significantly faster compared to [36, 20], since we do not use iterative dictionary learning. MATLAB implementations of all approaches were evaluated on a PC running on Intel Xeon 2.4GHz processor. Our unoptimized implementation needs only 3.5 seconds for inference on a test image, which is faster when compared to 5.5 seconds for [36] and 28 seconds for [20]. In our framework, all the saliency models share a common sparse code and contextual max-pooled vector. Inferring another model on same image needs an additional 1 second only. However, [36, 20] needs to calculate sparse codes for each model separately. [5] uses a larger dictionary of different sizes in both datasets. It took 3.85 seconds for inference in Graz-02 dataset and 17 seconds in PASCAL VOC-07.

4.6. Applications

4.6.1 Object class segmentation

The saliency maps obtained for a particular class are thresholded as in [13] followed by Grab-Cut [28]. Co-segmentation aims to segment the common object from a given set of images, which is similar to the image-level label provided in our weakly supervised training which enables a fair comparison with our approach. We train airplane, car and horse models using 130 images per category from PASCAL VOC 2010 detection dataset and evaluated on object discovery dataset [29]. Qualitative results[1] are shown in Fig. 7 and quantitative comparisons with co-segmentation approaches are shown in table 5. The jaccard similiarity, i.e, intersection over union (IOU) with the ground-truth is evaluated as in [29]. Although [29] performs better than our method on the horse class, we achieve better precision (84.09% vs 82.81%) which indicates that the proposed method can remove false detections on negative images.

Table 5. Comparison with segmentation approaches on Object Discovery dataset.

Method	Airplane	Car	Horse
Joulin *et al.* [15]	15.36	37.15	30.16
Joulin *et al.* [16]	11.72	35.15	29.53
Kim *et al.* [19]	7.9	0.04	6.43
Object Discovey [29]	55.81	64.42	51.65
Proposed	57.27	67.42	50.51

Table 6. Comparison with weakly supervised object annotation approaches on PASCAL -07 detection dataset

Method	Nguyen *et al.* [26]	Siva and Xing [34]	Siva *et al.* [33]	Proposed
Annotation accuracy (Avg. of 20 Classes)	22.4	30.4	32.0	36.22

4.6.2 Object annotation

We generated rectangular boxes from our saliency maps using coherent sampling [33] to annotate PASCAL VOC-07 detection images. As in [33] we select the first object location proposal in each image as the annotation of the object of interest. If $IOU > 0.5$ it is labeled as a correct annotation. Table 6 shows average annotation accuracy across 20 object categories. It illustrates that proposed approach is better than a deformable part-based model [9] trained on saliency maps of [33], which in turn uses several iterations of weakly supervised training to produce the reported result. It can be observed[1] from Fig. 7 that inspite of low contrast with the background, the proposed method can successfully annotate (yellow boxes) bird and cat images. Green colored rectangular box indicates the ground truth.

4.6.3 Action-specific patch discovery

We aim to automatically identify patches that help to describe the action. Qualitative evaluation of our R-ScSPM patch selection strategy on PASCAL VOC-2010 action dataset indicates that it is effective in identifying the most representative patches of different action categories as shown[1] in Fig. 7. The representative patches of an action category include class-specific objects as well as the action-specific orientation of human body parts. The patches corresponding to the body part performing the action, namely the hand, and the objects with which the hand interacts, namely the phone, camera and instrument have been extracted correctly.

5. Conclusion

In this paper, a weakly supervised top-down saliency approach is presented that requires just a binary label indicating the presence/absence of the object in an image for training. A novel R-ScSPM framework produces a saliency map that enables selection of representative patches for contextual saliency which is shown to improve the final saliency map. Extensive experimental evaluations show that the proposed method performs comparably with that of fully-supervised top-down saliency approaches.

Acknowledgment: The authors thank K.R Jerripothula for sharing the code of [13].

[1] More results are available at https://goo.gl/8bvHpZ.

References

[1] E. Ahmed, S. Cohen, and B. Price. Semantic object selection. In *CVPR*, 2014. 2

[2] D. Aldavert, A. Ramisa, R. L. de Mantaras, and R. Toledo. Fast and robust object segmentation with the integral linear classifier. In *CVPR*, 2010. 2, 7

[3] B. Alexe, T. Deselaers, and V. Ferrari. What is an object? In *CVPR*, 2010. 2, 7

[4] A. Borji and H. R. Tavakoli. Analysis of scores, datasets, and models in visual saliency prediction. In *ICCV*, 2013. 1

[5] H. Cholakkal, D. Rajan, and J. Johnson. Top-down saliency with locality-constrained contextual sparse coding. In *BMVC*, 2015. 1, 2, 5, 6, 7, 8

[6] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html. 5

[7] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2010 (VOC2010) Results. http://www.pascal-network.org/challenges/VOC/voc2010/workshop/index.html. 5

[8] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. Liblinear: A library for large linear classification. *The Journal of Machine Learning Research*, 9:1871–1874, 2008. 3

[9] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *PAMI*, 2010. 8

[10] S. Frintrop, T. Werner, and G. M. García. Traditional saliency reloaded: A good old model in new shape. In *CVPR*, 2015. 1

[11] B. Fulkerson, A. Vedaldi, and S. Soatto. Class segmentation and object localization with superpixel neighborhoods. In *ICCV*, 2009. 2, 6, 7

[12] D. Gao, S. Han, and N. Vasconcelos. Discriminant saliency, the detection of suspicious coincidences, and applications to visual recognition. *PAMI*, 2009. 2, 6, 7

[13] K. R. Jerripothula, J. Cai, F. Meng, and J. Yuan. Automatic image co-segmentation using geometric mean saliency. In *ICIP*, 2014. 2, 8

[14] Y. Jia and M. Han. Category-independent object-level saliency detection. In *ICCV*, 2013. 1

[15] A. Joulin, F. Bach, and J. Ponce. Discriminative clustering for image co-segmentation. In *CVPR*, 2010. 2, 8

[16] A. Joulin, F. Bach, and J. Ponce. Multi-class cosegmentation. In *CVPR*, 2012. 2, 8

[17] C. Kanan, M. H. Tong, L. Zhang, and G. W. Cottrell. Sun: Top-down saliency using natural statistics. *Visual Cognition*, 17(6-7):979–1003, 2009. 2, 6, 7

[18] N. Khan and M. F. Tappen. Discriminative dictionary learning with spatial priors. In *ICIP*, 2013. 2, 7

[19] G. Kim, E. P. Xing, L. Fei-Fei, and T. Kanade. Distributed cosegmentation via submodular optimization on anisotropic diffusion. In *ICCV*, 2011. 2, 8

[20] A. Kocak, K. Cizmeciler, A. Erdem, and E. E. Top down saliency estimation via superpixel-based discriminative dictionaries. In *BMVC*, 2014. 2, 5, 6, 7, 8

[21] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006. 3

[22] Y. Li, X. Hou, C. Koch, J. M. Rehg, and A. L. Yuille. The secrets of salient object segmentation. In *CVPR*, 2014. 2

[23] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, and H.-Y. Shum. Learning to detect a salient object. *PAMI*, 2011. 1

[24] M. Marszałek and C. Schmid. Accurate object recognition with shape masks. *International journal of computer vision*, 97(2):191–209, 2012. 2, 6, 7

[25] F. Moosmann, D. Larlus, and F. Jurie. Learning saliency maps for object categorization. In *ECCV workshop*, 2006. 1, 2, 6

[26] M. H. Nguyen, L. Torresani, F. de la Torre, and C. Rother. Weakly supervised discriminative localization and classification: a joint learning process. In *ICCV*, 2009. 2, 8

[27] A. Opelt, A. Pinz, M. Fussenegger, and P. Auer. Generic object recognition with boosting. *PAMI*, 28(3):416–431, 2006. 5

[28] C. Rother, V. Kolmogorov, and A. Blake. Grabcut: Interactive foreground extraction using iterated graph cuts. In *ACM Transactions on Graphics (TOG)*, volume 23, pages 309–314. ACM, 2004. 8

[29] M. Rubinstein, A. Joulin, J. Kopf, and C. Liu. Unsupervised joint object discovery and segmentation in internet images. In *CVPR*, 2013. 2, 5, 8

[30] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *IJCV*, 2015. 2

[31] O. Russakovsky, Y. Lin, K. Yu, and L. Fei-Fei. Object-centric spatial pooling for image classification. In *ECCV*. 2012. 6

[32] G. Sharma, F. Jurie, and C. Schmid. Discriminative spatial saliency for image classification. In *CVPR*, 2012. 2

[33] P. Siva, C. Russell, T. Xiang, and L. Agapito. Looking beyond the image: Unsupervised learning for object saliency and detection. In *CVPR*, 2013. 2, 8

[34] P. Siva and T. Xiang. Weakly supervised object detector learning with model drift detection. In *ICCV*, 2011. 2, 8

[35] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong. Locality-constrained linear coding for image classification. In *CVPR*, 2010. 2

[36] J. Yang and M.-H. Yang. Top-down visual saliency via joint crf and dictionary learning. In *CVPR*, 2012. 1, 2, 5, 6, 7, 8

[37] J. Yang, K. Yu, Y. Gong, and T. Huang. Linear spatial pyramid matching using sparse coding for image classification. In *CVPR*, 2009. 1

[38] B. Yao, A. Khosla, and L. Fei-Fei. Combining randomization and discrimination for fine-grained image categorization. In *CVPR*, 2011. 2

[39] J. Zhang, S. Sclaroff, Z. Lin, X. Shen, B. Price, and R. Měch. Minimum barrier salient object detection at 80 fps. In *ICCV*, 2015. 1, 7

[40] R. Zhao, W. Ouyang, H. Li, and X. Wang. Saliency detection by multi-context deep learning. In *CVPR*, 2015. 2

[41] J. Zhu, Y. Qiu, R. Zhang, J. Huang, and W. Zhang. Top-down saliency detection via contextual pooling. *Journal of Signal Processing Systems*, 74(1):33–46, 2014. 2, 5