Looking Beyond the Image: Unsupervised Learning for Object Saliency and Detection

Parthipan Siva*, Chris Russell†, Tao Xiang*, Lourdes Agapito†
School of EECS, Queen Mary, University of London, UK
psiva7@gmail.com, [chrisr,txiang,lourdes]@eecs.qmul.ac.uk

Abstract

We propose a principled probabilistic formulation of object saliency as a sampling problem. This novel formulation allows us to learn, from a large corpus of unlabelled images, which patches of an image are of the greatest interest and most likely to correspond to an object. We then sample the object saliency map to propose object locations. We show that using only a single object location proposal per image, we are able to correctly select an object in over 42% *of the images in the* PASCAL VOC *2007 dataset, substantially outperforming existing approaches. Furthermore, we show that our object proposal can be used as a simple unsupervised approach to the weakly supervised annotation problem. Our simple unsupervised approach to annotating objects of interest in images achieves a higher annotation accuracy than most weakly supervised approaches.*

1. Introduction

With the prevalence of media sharing websites such as Flicker, researchers have easy access to terabytes of liberally licensed images. The primary bottleneck that prohibits the use of this data lies in the difficulty of annotating it. In this paper we show how such images can be automatically annotated. Our primary focus lies on two types of annotation: Given an image, *(i)* find a bounding box tightly containing one object of interest (this unsupervised annotation is comparable to the weakly supervised multi-instance learning [8, 28, 29] approaches), and *(ii)* produce a binary mask highlighting regions of interest. Unlike other annotations (e.g. find 1000 boxes covering every object in the image [3, 25]) annotation *(i)* can be easily validated with a simple "Yes/No" from a human, and can be directly used to learn an object detector (see section 5.3).

As the word "saliency" is widespread in the literature and used to refer to whatever a researcher currently considers interesting, it is important to distinguish between different usages of the word. We will use "human saliency" to refer to methods that predict where a human looks, or what they will label as interesting in an image, and "object saliency" to refer to methods that annotate the location of a predefined set of object types. As humans look at objects and find them interesting, there is substantial overlap between the two problems, and methods for one problem may be applied to the other.

Human saliency was first formulated as a predictor of human fixation in images [16]. Recent applications in computer vision have led to an increased interest in object saliency formulations [3, 6, 13, 15, 31] that propose salient bounding boxes in images as potential object locations. These boxes can be used to speed up object detection [3, 31] or weakly supervised object annotation for training a detector [8, 29].

Most existing approaches for object saliency can be characterised as extensions of expert-driven human saliency methods or supervised learning methods. Object saliency methods that build on expert-driven human saliency approaches [6, 13, 15] tend to use cognitive psychological knowledge of the human visual system and finds image patches on edges and junctions as salient using local contrast or global unique frequencies. Recently, object saliency approaches based on supervised learning have emerged [3, 20, 25]. In these approaches, data from manual annotation of images are used to mark patches of interest. These annotations can then be used to train a saliency model (based on global and local image features) to predict patches of interest in unseen images.

We propose an unsupervised approach to object saliency (fig.1) that does not rely on any information outside of a large corpus of unlabelled images. As it is not possible to predict what a person will find salient, without either asking or observing them, our research attempts to answer the related question *"What should a person be interested in?"* We show that an answer lies in the most surprising patches of an image, or those that have the least probability of being sampled from a corpus of similar images.

To understand the relationship between our approach and

*Author funded by EPSRC under the EP/G063974/1 grant
†Authors funded by the European Research Council under the ERC Starting Grant agreement 204871-HUMANIS

1063-6919/13 $26.00 © 2013 Crown Copyright
DOI 10.1109/CVPR.2013.416

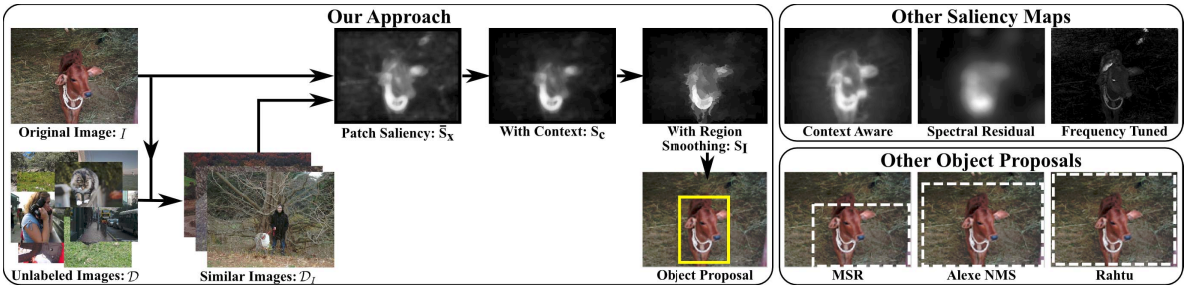


Figure 1. *Our approach to object saliency and object location proposal in comparison with some existing techniques. Other methods include: context aware [13], spectral residual [15], frequency tuned [1], MSR [12], Alexe NMS [3], and Rahtu [25].*

expert driven approaches, consider two broad scenarios:

Things and Stuff Addleson [2] observed that image patches can be loosely categorised as belonging to one of two types of objects, *Things* and *Stuff*. Things being individual objects such as a person or a car, and stuff being amorphous object-classes such as road or grass that can be recognised as reoccurring stochastic patterns. The majority of the natural world is "stuff", and as such, if we use descriptors of the world that are robust to the small amount of stochastic variation that "stuff" classes exhibit, we will invariably find "things" or foreground objects as being more salient. This tendency to find "things" as being salient is intensified by sampling from similar images. With the majority of an image being "stuff", images sharing the same dominant patterns typically contain similar "stuff" rather than similar "things".

Edges and Junctions The two most important types of filters in expert-based filter banks are those that detect edges and junctions, as these filters are highly useful in selecting foreground objects. To understand why our saliency measure exhibits the same bias towards edges and junctions, consider an image composed of a single, approximately homogeneous object with a smooth boundary, and a homogeneous background. We allow the resolution of the image to vary as n^2 dots per square inch. The number of pixels lying in the interior of the object, or the background, will be $O(n^2)$, while the number of the pixels adjacent to an object edge will be $O(n)$ and the number of pixels adjacent to a junction (corresponding to an edge intersection) will be $O(1)$. Consequently, for most choices of n, it is much less likely that either an edge or a junction will be sampled from an image and thus our approach will consider them to be more salient.

The combination of these attributes leads to an object saliency map with highly desirable properties: Our saliency map exhibits a bias towards selecting junctions or the intersection of objects as salient regions, and a secondary bias towards the objects themselves because they occur infrequently in the set of similar images. However, it remains robust to the presence of junctions in reoccurring "stuff", such as brick work or tree branches, that frequently confuse filter-bank driven approaches.

Sampling Bounding Boxes It remains an open problem as to how bounding boxes (boxes that propose the location of objects) should be sampled from a per pixel object saliency map. Each sequentially selected box should tightly fit around one object, and never around an object that has been sampled before. However, the presence of an object in one box can cause neighbouring boxes to appear salient, leading to the selection of boxes which highly overlap each other and only partially overlap the actual object.

Suppression based sampling techniques, such as non-maximum suppression, are commonly used to avoid such oversampling. Under such formulations [3, 7] the selection of a box will act as a hard [3] or soft [7] constraint that blocks heavily overlapping boxes from being simultaneously selected. However, non-maximum suppression carries its share of disadvantages [12]. In particular, if a selected box narrowly misses an object it may block the future selection of a box that overlaps this object. To avoid these near misses, we propose a novel sampling method which encourages the selection of a box that "explains away" possible bounding boxes in the area blocked by non-maximum suppression.

Pipeline: Figure 1 illustrates our approach to object saliency. Our probabilistic patch based approach allows us to leverage the use of a corpus of unlabelled images. Furthermore, our object proposals, based on sampling our object saliency map, correctly locate objects in many images on the first proposal. This is an ideal behaviour for using our object proposal as an unsupervised approach to annotating objects of interest in weakly labelled data.

2. Prior work

Early works on human saliency were developed from biological models of the human visual system, and estimated fixation points where a human viewer would initially focus. These methods made use of the feature-integration theory of attention [30] to predict human fixation points in images and such are ill-suited for finding regions of interest. Our interest, motivated by applications in object detection, lies in object saliency approaches that can detect salient regions

as potential object locations.

Object saliency methods have made use of global frequency-based features [1, 15], which finds regions characterised by rare frequencies in the Fourier domain as salient. However, due to their global nature, they have difficulty in finding the full extent of objects [13]. More recently, global and local features have been combined to identify regions of interest in images [3, 6, 12, 13]. In [6, 13] local patches or segments are compared against all other patches or segments in the image, using colour distances. Saliency is then defined as the uniqueness of local patches or segments compared to the rest of the image. Unlike our approach, these methods are evaluated on simple datasets (e.g. MSRA [1]) with a single salient object per image, they do not provide means of proposing multiple object locations in an image, and they do not consider the use of other similar images.

Recently three approaches [3, 12, 25] have provided object location proposals on the challenging PASCAL VOC dataset [9]. [12] develops an unsupervised approach that integrates both saliency computation and object location proposal. Object locations are proposed as rectangular regions which contain pixels that can not be reconstructed using the pixels outside the region (based on colour). [3] starts by sampling rectangular regions based on the global frequency saliency map of [15] then adds additional cues such as colour contrast and super-pixel straddling. Parameters and weights for the different cues are learned on a fully annotated auxiliary dataset. Similar to [3], [25] also proposes a bounding box selection method based on supervised learning. We evaluate directly against these three methods on the PASCAL VOC dataset [9]. However, unlike these existing approaches to saliency, our method builds knowledge about the current image from similar unlabelled images. In particular we define a patch as salient if it is uncommon not only in the current image, but also in other similar images drawn from a large corpus of unlabelled images.

Other methods have made use of multiple images for saliency. In [20] patches are classified as unique based on a support vector machine (SVM) learned from similar manually annotated images. In contrast, our method is unsupervised and does not need manually annotated images. In [32] the current image is registered to similar images and the difference between the registered image and the similar images are used as the saliency map. This requires the use of a very large auxiliary dataset which needs to contain similar images with the same background but without the salient object. Our patch based approach does not require near identical similar images.

Most object location proposal methods [3, 12, 25, 31] which report on the challenging PASCAL VOC dataset attempt to achieve a high recall rate given a large number of object location proposals. In this paper we are interested in the weakly supervised object annotation task [32], which requires high precision of a few object proposals.

In weakly supervised object annotation, an algorithm attempts to place a tight bounding box around objects of interest, after taking as input two sets of images: one of images not containing the objects, and the other set of images containing them. Most existing methods [8, 22, 28, 29] formulate this as a multiple instance learning problem. However, the simplest method to annotate the object of interest in an image is to assume that the most object like region in the image is the object of interest, i.e. to take the first location proposed as a potential object. This simple approach completely ignores the available weak labels (indicating which images contains the object of interest). Surprisingly, as Siva et al. [28][1] showed a relatively high accuracy for the weakly supervised annotation task can be achieved by this simple approach. In this paper we show that our saliency based object location proposal achieves higher weakly supervised annotation accuracy than other methods that propose object locations, or even those weakly supervised learning methods that make additional use of annotated weak labels.

Outside of saliency, object detection, and weakly supervised learning, there are several other related works. CMU has done exciting work on image in-painting [14] that motivated our decision to sample from related images and their more recent work [27] may provide a better method of finding related images; their work [17] finds related images and uses these images for object pop-up via background subtraction. Unlike our work, they used image warps to match patches taken from different views of the same scene. Also related, is the concept of abnormality detection in video and in images [4, 5, 33], as we consider abnormal datapoints with low *a priori* probability to be salient. We differ from [5, 33] in that we are interested in detecting abnormal *patches* rather than *scenes* and we make no use of video based cues, and from [4] in that we do not model the relationship between patches and we make use of a marginal density estimator rather than the MAP, giving us greater robustness, and allowing us to potentially detect salient regions using only a single image (see fig. 7).

3. Sampling-based Saliency

Given an image I and a large corpus of unlabelled images $\mathcal{D}$, we wish to find a saliency map S_I for image I. We define salient patches, as those belonging to image I, that have the least probability of being sampled from a set of images $\mathcal{D}_I$ similar to I. Here $\mathcal{D}_I$ includes the current image I and other images obtained from the corpus of unlabelled images $\mathcal{D}$ and patches are $n \times n$ regions around each image pixel. We must now compute p_x, a number proportional to the probability of sampling patch x from $\mathcal{D}_I$.

[1] The main method of [28] used the weak annotation but they show results for annotating object location using the most object like instance proposed by [3].

We make the assumption that the probability of sampling a patch x from an image $J \in \mathcal{D}_I$ can be formulated by uniformly selecting a patch y in J, and then perturbing it by some noise in an informative feature space. This gives us:

$$p_x \propto \Pr(X = x|\mathcal{D}_I) \tag{1}$$

$$= \int_{\mathcal{D}_I} \Pr(X = x|J)\,\mathrm{d}J \tag{2}$$

$$= \int_{\mathcal{D}_I}\int_{J} \Pr(X = x|y)\Pr(y|J)\,\mathrm{d}y\,\mathrm{d}J \tag{3}$$

$$\propto \int_{\mathcal{D}_I}\int_{J} \Pr(X = x|y)\,\mathrm{d}y\,\mathrm{d}J \tag{4}$$

Assuming the noise is uniform and Gaussian[2] over the space of image patches, we have:

$$p_x \propto \int_{\mathcal{D}_I}\int_{J} \exp\left(-\frac{d(x,y)^2}{\sigma^2}\right)\,\mathrm{d}y\,\mathrm{d}J \tag{5}$$

and we replace the proportionality sign in (5) with equality and take this as our definition of p_x. Here $d(x, y)$ is the Euclidean distance between the feature representations of patches x and y.

For efficient computation, it is important to note that the Gaussian distribution is *short-tailed*, and for our purposes p_x can be approximated as:

$$p_x \approx \sum_{y \in N_m(x, \mathcal{D}_I \setminus \{I\})} \exp\left(\frac{-d(x,y)^2}{\sigma^2}\right) + \sum_{y \in N_m(x,\{I\})} \exp\left(\frac{-d_I(x,y)}{\sigma^2}\right) \tag{6}$$

where $N_m(x, \mathcal{D}_I \setminus \{I\})$ are the m approximate nearest neighbours (ANNs) of patch x taken from all images using distance measure d in $\mathcal{D}_I$ except I and computed using Fast Library for Approximate Nearest Neighbours (FLANN) [21]. Note that that image set $\mathcal{D}_I$ includes the image I, and when selecting a patch from it some care must be taken not to sample from adjacent patches that always have similar appearance. Instead we want to find other patches y spatially far from the patch x as these matches will correspond to repeating patterns i.e. *stuff*. Following [13], a spatial distance bias is introduced to discourage matching spatially close patches in the same image, and we use

$$d_I(x,y) = \left(\frac{d(x,y)^2}{1 + c \cdot (l(x) - l(y))^2}\right) \tag{7}$$

where c is a constant, $l()$ is the location of patches in normalised image coordinates, and $c = 3$ per [13].

We now have p_x the probability of a patch with the same feature response as x being sampled from $\mathcal{D}_I$. A high value of p_x indicates that the patch x is common in the image corpus, and the saliency of patch x is obtained as:

$$S_x = 1 - p_x \tag{8}$$

where p_x, over all patches x in the image I, was normalised to the range $[0, 1]$. To account for scale changes in salient objects, we compute saliency S_x (8) at four different image scales $[1, .8, .5, .3]$ and average the result over the four scales $\bar{S}_x$ as the patch saliency.

[2]This assumption is robust to choice of distribution. After normalisation of the distances, we get good performance using a standard deviation σ of 1; We tried also exponential and Cauchy distributions, empirically it made little difference.

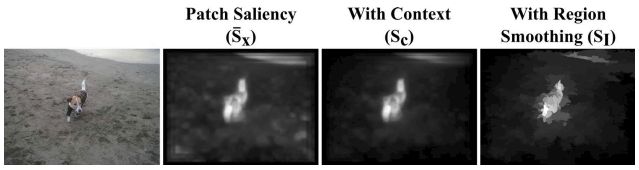


Figure 2. *Saliency ($\bar{S}_x$) is first computed based on the probability of sampling image patches from the current image or other similar images. Then each pixel is weighted by their distance to high salient pixels (S_c). Finally, the saliency map is smoothed based on image segmentation (S_I).*

Post-Processing Two post-processing steps are applied to $\bar{S}_x$. First, as in [13], immediate context information is included by weighting the saliency value of each pixel by their distance from the high salient pixel locations. Second, a segmentation based smoothing is applied to the saliency map to recover image boundary information.

To encode immediate context information, high salient pixel locations $\mathcal{F} = \bar{S}_x > T$ are found and the saliency value at all pixel location i is weighted by their distance to $\mathcal{F}$.

$$S_c(i) = \bar{S}_x(i)\left(\sum_{y \in N_{64}(i,\mathcal{F})} \exp\left(-\frac{(l(i) - l(y))^2}{\sigma_l}\right)\right) \tag{9}$$

where $N_{64}(i, \mathcal{F})$ are the 64 nearest neighbours of i in $\mathcal{F}$, $l()$ is the normalised image coordinate of pixels. As shown in fig. 2, the resulting saliency map S_c is blurred due to the use of overlapping patches and image boundaries (edges between objects and background) are not preserved. To recover some of the image boundary information, we segment image I using the segmentation technique of [11]. For each segment region, the average saliency from S_c is obtained and used as the final saliency value for that segment, producing our saliency map S_I.

Similar Images In (6), a set of similar images $\mathcal{D}_I$ to the current image I must be obtained from a corpus of unlabelled images $\mathcal{D}$. We follow the approach of [14] and select 20 similar images from $\mathcal{D}$, using Euclidean distance on GIST [23] descriptors and a 30×20 thumbnail image in Lab colour space.

Patch Features $d(x, y)$ is the Euclidean distance between the feature representation of patches x and y. We represent each $n \times n$ patch using the concatenation of two features. First, the $n \times n$ patch is represented as a vector of length $3n^2$

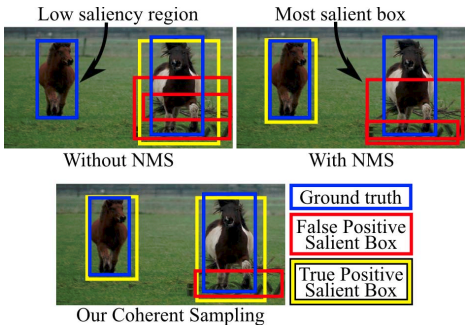


Figure 3. *An illustration of sampling. Sampling from the saliency map without non-maximum suppression (NMS) results in an over sampling of high saliency regions. While this allows exact alignment to the true object to be found in the first 3 salient boxes, objects in a lower salient region are missed. Sampling with NMS means that the lower saliency region will still be sampled from. However, the selection of a box that narrowly misses the object may cause the later rejection of the most salient box containing the object. Our coherent sampling recovers from many of these cases. Best viewed in colour.*

in Lab colour space. The Lab colour vector is concatenated to the 128 bin SIFT [18] descriptor of the $n \times n$ patch. The resulting vector of length $3n^2 + 128$ is used as the feature descriptor of the patch.

4. Bounding Box Sampling

Bounding boxes that should contain an object can be selected by sampling from a per-pixel saliency map. In the past several options have been explored [19], such as thresholding the saliency map followed by connected region detection [15] or selecting a bounding box containing 95% of the image saliency [19]. Such approaches typically assume one object per image and select a single salient region. For proposing multiple bounding boxes per image the saliency map may be randomly sampled from [3], or sampled from the highest score to the lowest score with non-maximum suppression (NMS) [12]. Random sampling based on saliency map density results in over-sampling regions of high saliency. This may be desirable if it is difficult to find the exact object location. However, in this case, low saliency regions containing objects will be missed. While non-maximum suppression ensures that even low salient regions are sampled from, it does not allow for the repeated sampling of high salient regions. This can cause true object locations to be narrowly missed even if the object has been successfully detected. A comparison between the two approaches can be seen in fig. 3.

We propose coherent sampling, as a variant of non-maximum suppression designed to avoid narrowly missing a detected object (see fig. 3). Consider an image in which we have already selected a set T of object locations, and we wish to add one more location to it. As with NMS, we select the box with the highest saliency score (b_0) that is not near the other T locations. Unlike standard NMS sampling we do not automatically add b_0 (the box with the highest saliency score) to the top T proposed boxes. Instead we consider $\mathcal{B}$, the set of all boxes that would be blocked by b_0, including itself, and seek $b_* \in \mathcal{B}$, the box that best explains the saliency of all bounding boxes in $\mathcal{B}$.

To find such a box, we describe the region from which the boxes in $\mathcal{B}$ are drawn using a saliency weighted average BoW SIFT histogram:

$$\mu^{\text{SIFT}} = \frac{1}{\sum_{i=0}^{N} d_i} \sum_{i=0}^{N} d_i f^{\text{SIFT}}(b_i) \tag{10}$$

where $f^{\text{SIFT}}(b_i)$ is the dense SIFT BoW histogram representation of b_i and d_i is the saliency score of box b_i. Then to maximise the overlap with the salient boxes in $\mathcal{B}$ that will be suppressed, b_* is chosen as the box with the closest histogram to μ^{SIFT}.

$$b_* = \arg\min_{b_i} \parallel f^{\text{SIFT}}(b_i) - \mu^{\text{SIFT}} \parallel_2 \tag{11}$$

The saliency score d_i for box b_i is defined as:

$$d_i = \frac{1}{|b_i|^r} \sum_{p \in b_i} S(p) - \frac{1}{|u_i|^r} \sum_{p \in u_i} S(p) \tag{12}$$

$|\cdot|$ refers to the size of the box in pixels, u_i is a buffer around the box b_i that ensures we select local maxima. It is chosen to be a maximum of 10 pixels wide, and r is a soft bias on the box size. When $r = 0$ the highest density box fills the image and if $r = 1$ the highest density box is typically only a single pixel wide. To sample boxes at different scales, instead of alternating between 4 explicit choices of scale [3], we alternate between sampling with a soft bias towards large scales with $r = 0.5$ and a bias towards smaller patches with $r = 0.75$.

5. Experiments

All results are reported on the PASCAL VOC 2007 [9] Train and Validation set, the standard dataset used for the weakly supervised annotation task [8, 24, 28, 29]. Our corpus of unlabelled images $\mathcal{D}$ consists of $98,000$ images obtained from LABELME [26], PASCAL VOC 2007, and 2012 [9] datasets.

For all experiments we fixed $\sigma = 1$ for (6), $\sigma_l = 0.2$ for (9), and at each pixel location a patch of size 7×7 pixels was used as the Lab colour representation and a 4×4 cell SIFT descriptor with each cell being 4 pixel was used.

5.1. Object Proposals

Performance Metric: The precision recall curve (PRC) is used to evaluate the performance of the object location proposals as it captures the behaviour of both precision and recall as the number of proposed boxes increases. Alternatively, the recall rate as a function of the number of object

location proposals is used by [3]. Note that recall rate vs object proposals is good for comparing the recall rate at high number of proposed locations but not for evaluating the precision when only one object location is proposed. For completeness, we report both PRC and the recall rate vs object proposals.

We report the precision and recall as function of the number of objects proposed per image following the PASCAL challenge [9]. This differs from [12], in that it treats multiple detections of the same object as false positives. Correct detection is also per PASCAL challenge [9] and is defined as the area of intersection of the two boxes divided by the area of union is greater than 0.5.

We are also interested in detecting only one object from each image because this is important for the weakly supervised annotation task (see section 5.2). As a result, we also report recall and precision based on detecting one object per image. Let $D_{i,j} \in \{0,1\}$ be a vector indicating if the j^{th} box proposed by the saliency algorithm correctly detects an object in the i^{th} image, then:

$$R^{\text{one}}(j) = \frac{\sum_{i=1}^{N} \max(D_{i,1}, \ldots, D_{i,j})}{N} \tag{13}$$

$$P^{\text{one}}(j) = \frac{\sum_{i=1}^{N} \max(D_{i,1}, \ldots, D_{i,j})}{jN} \tag{14}$$

where N is the number of images in the dataset, j is the number of boxes proposed per image, R^{one} and P^{one} are the recall and precision assuming one object per image.

We compare coherent sampling (**Our**) to:

Alexe NMS Objectness method of [3][3] using NMS sampling. This is a supervised approach that uses 50 manually annotated images.

Alexe MN The same supervised method of [3] using multinomial sampling.

MSR The unsupervised method of [12]. Boxes were obtained from the authors and has less than 100 boxes per image (hence the flat line for MSR in fig. 4(c)).

Rahtu The supervised approach of [25][4] in which a structured support vector machine (SVM) is used to generate a ranked list of rectangular regions.

Comparison with Competitors: The PRC curves for the first 1000 proposed boxes are shown in fig. 4 and a visualisation of the proposed bounding boxes is provided in fig. 5. Based on the average precision our proposed object locations substantially out-perform our competitors. Particularly, our first object proposal per image correctly locates an object in 42% of the images nearly 10% higher than our closest competitor (see table 1). However, as seen from the recall vs number of proposed windows, table 2, while our method has a higher recall than [3] at the first box, the recall at 1000 boxes is lower than that of [3]. The choice of [3] vs our proposed method or a hybrid approach would depend on the application, and whether high recall or precision is more important. We show in section 5.2 that our object proposals are particularly suitable for the task of annotating of weakly labelled data which requires maximal precision at the first proposed object location.

[3] http://www.vision.ee.ethz.ch/~calvin/objectness/

[4] http://www.cse.oulu.fi/CMV/Downloads/ObjectDetection

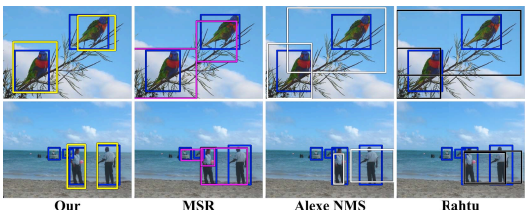


Figure 5. *Best bounding boxes taken from the top 10 proposed object locations by our coherent sampling method (Our), MSR [12], Alexe et al. NMS [3], and Rahtu et al. [25]. Blue is ground truth.*

	Our	Alexe MN [3]	Alexe NMS [3]	MSR [12]	Rahtu [25]
VOC07	**42.3**	20.4	30.8	32.6	32.5
VOC07-6x2	**42.8**	19.6	27.6	27.7	29.6

Table 1. *Percent of images in which an object is correctly located by the first object proposal.*

# Prop	Our	Alexe MN [3]	Alexe NMS [3]	MSR [12]	Rahtu [25]
1	**0.17**	0.08	0.12	0.13	0.13
2	**0.21**	0.14	0.19	0.20	0.16
10	0.34	0.32	**0.39**	0.35	0.26
100	0.57	0.50	**0.66**	0.42	0.51
1000	0.79	0.64	**0.86**	0.42	0.75

Table 2. *Recall vs # of object location proposed on the* PASCAL *2007 TrainVal dataset (excludes objects annotated as difficult).*

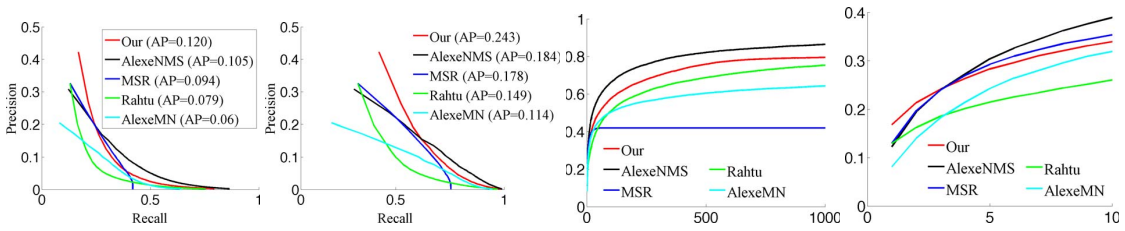
	Our	Alexe MN [3]	Alexe NMS [3]	MSR [12]	Rahtu [25]
VOC07	**31.1**	25.8	23.4	24.0	23.6
VOC07-6x2	**42.4**	33.8	28.8	29.6	29.0

Table 3. *Comparison of different object proposal based on the annotation of weakly labelled data.*

NMS Sampling vs Coherent Sampling: NMS sampling obtains an average precision of 0.117 vs the coherent sampling of 0.120. Overall the contribution of coherent sampling is small compared to NMS sampling. However, for the initial object proposal there is a 3% boost in precision which is beneficial for the weakly supervised annotation task (see section 5.2).

5.2. Weakly Supervised Object Annotation

In weakly supervised object annotation, a set of images with the object of interest and a set of images without the object of interest is given and the goal is to locate the object of interest in all images that contain it. As discussed in section 2, we select the first object location proposal in each image as the annotation of the object of interest. We test the weakly supervised annotation accuracy on the 20 classes of PASCAL VOC 2007 (VOC07) as defined in [29] and 6 classes (aeroplane, bicycle, boat, bus, horse, and motorbike) with Left and Right pose separately (VOC07-6x2) as defined in [8].



(a) All Objects (b) One Object per Image (c) Hit Rate (d) Hit Rate

Figure 4. *Precision recall curve and recall vs number of object location proposal on the* PASCAL *2007 TrainVal dataset. (d) is a zoomed in view of (c). Best viewed in colour.*

	VOC07		VOC07-6x2	
	Init	*Final*	*Init*	*Final*
Nguyen [22]*	–	22.4	–	25
Siva and Xiang [29]	28.9	30.4	39.6	49
Siva et al. [28]	29.0	–	37.1	47
PandeyNoCrop [24]	–	–	36.7	59†
PandeyCrop [24]	–	–	**43.7**	**61**†
Deselaers1Feat [8]	–	–	35.0	40
Deselaers4Feat [8]	–	–	39.0	50
Our	**31.1**	**32.0**	42.4	55

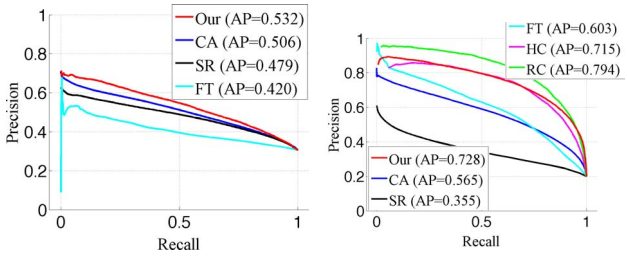
*As reported in [29]. † Requires aspect ratio to be set for initialisation.

Table 4. *Average annotation results for* PASCAL *datasets using different weakly supervised learning methods.*

Comparison with Other Saliency Methods: We compare our object proposal based annotation (*Our*) results against other object proposal methods in table 3. For all methods, except *Alexe MN*, we select the first object location proposal per image as the object of interest (annotation for weakly labelled data). For *Alexe MN*, we select the top 100 proposed bounding boxes and from these we select the bounding box with the highest objectness score, with objectness scores taken from [3]. Our object proposal has a relative improvement from all other object proposal methods of at least 21% on PASCAL07 and 25% on the PASCAL07-*6x2* datasets.

Comparison with Weakly Supervised Methods: As seen in [24, 28, 29], the initial annotation of the object of interest can be iteratively refined by training a deformable part-based model (DPM) detector [10] and applying the trained detector to the weakly annotated images known to contain the object of interest. We iteratively train the DPM using our object location proposals as the initial annotation, following [29]. Note for the iterative refinement we make use of weak annotation, while the initialisation is unsupervised, the final iterative annotation result is weakly supervised. A numeric evaluation of all methods can be found in table 4.

Overall our object proposal based annotation obtains high initial annotation accuracy and high iteratively refined annotation accuracy; outperforming almost all existing approaches using a much simpler approach. However, on the more restrictive single pose subset (*VOC07-6x2*) our annotation accuracy is lower than that of Pandey and Lazebnik [24] who make use of prior knowledge regarding the aspect ratio of bounding boxes.



(a) VOC07 (b) MSRA

Figure 6. *Per-pixel accuracy vs CA [13], SR [15], FT [1], HC [6], RC [6].*

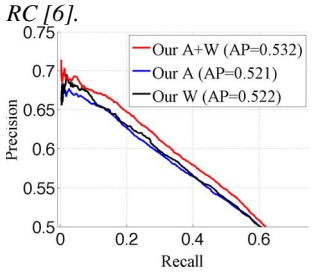


Figure 7. *Variations of our method: Our A - using similar images without current image $\mathcal{D}_I \setminus \{I\}$, Our W - using only the current image I, and Our A+W - using similar images with current image $\mathcal{D}_I$.*

5.3. Evaluation of Saliency Maps

As in [13] we evaluate the saliency map's ability to predict foreground pixels by reporting the precision recall curve (PRC) and average precision (AP) as a function of the saliency map threshold. We use the PASCAL 2007 segmentation data (422 images in the train and validation set), where all object segments are used as foreground pixels. We evaluate on the PASCAL dataset as it is a more challenging dataset and the common dataset used for the task of annotating weakly labelled object data. For completeness we also report PRCs for the MSRA saliency dataset [1].

Figure 6 shows the PRCs of our method and some other existing saliency approaches; some examples can be seen in fig. 8. Our approach perform better than many existing methods; particularly note the better performance over the spectral residual (SR) method [15], used in the object proposals of [3], and the context aware (CA) method [13], which is closest to our formulation.

On the more restrictive MSRA dataset (fig. 6b), texture based models are unneeded and better results can be achived using coarsely quantized color models such as histogram contrast (HC) or region contrast (RC) [6]. Of the two methods, HC has similar performance to our approach while RC

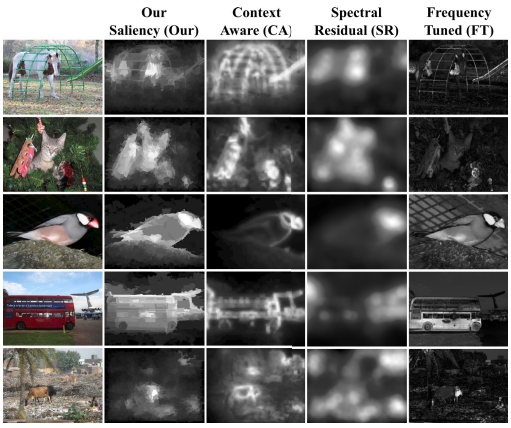


Figure 8. *Our image saliency in comparison to CA [13], SR [15], and FT [1] methods.*

has better performance. RC [6] explicitly targets segmentation regions of unique colour. In MSRA, unique colors often indicate salient objects but on VOC, unique color is often indicative of a small patch of sky (see fig. 8 bott. left).

In section 2, we defined as salient patches with a low probability being sampled from a set of similar images $\mathcal{D}_I$. In fig. 7, we analyse the contribution of using the current image I in addition to other similar images when computing the saliency map. We plot the precision-recall curve of the saliency map computed using similar images without current image $\mathcal{D}_I \setminus \{I\}$ (across image saliency A), using just the current image I (within image saliency W), and both combined $\mathcal{D}_I$ (our combined A+W). Note that although across-image saliency and within-image saliency have similar performance, combining them provides a boost in performance, particularly in the region of high-precision we are most concerned with.

6. Conclusion

We have presented a novel unsupervised approach to the problems of saliency and bounding box annotation[5], and shown how it substantially outperforms all other saliency based approaches to bounding box annotation on real world data. In comparison to existing approaches tailored for the problem of detection from weak annotation, we outperform all existing methods on the full VOC dataset. The power and conceptual simplicity of our approach makes it an attractive candidate to be combined with supervised approaches and to be applied to a wide variety of problems.

References

[1] R. Achanta, S. Hemami, F. Estrada, and S. Süsstrunk. Frequency-tuned salient region detection. In *CVPR*, 2009.

[2] E. H. Adelson. On seeing stuff: the perception of materials by humans and machines. In *SPIE*, 2001.

[3] B. Alexe, T. Deselaers, and V. Ferrari. Measuring the objectness of image windows. *PAMI*, 2012.

[4] O. Boiman and M. Irani. Detecting irregularities in images and in video. *IJCV*, 74(1), 2007.

[5]Code and bounding boxes available at `http://www.psiva.ca/Papers/CVPR2013/CVPR2013.html`

[5] M. Breitenstein, H. Grabner, and L. Van Gool. Hunting nessie – real-time abnormality detection from webcams. In *IEEE Int. Workshop on Visual Surveillance*, 2009.

[6] M. Cheng, G. Zhang, N. Mitra, X. Huang, and S. Hu. Global contrast based salient region detection. In *CVPR*, 2011.

[7] C. Desai, D. Ramanan, and C. C. Fowlkes. Discriminative models for multi-class object layout. *IJCV*, 95(1), 2011.

[8] T. Deselaers, B. Alexe, and V. Ferrari. Localizing objects while learning their appearance. In *ECCV*, 2010.

[9] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 88(2), 2010.

[10] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *TPAMI*, 32(9), 2010.

[11] P. F. Felzenszwalb and D. P. Huttenlocher. Efficient graph-based image segmentation. *IJCV*, 59(2), 2004.

[12] J. Feng, Y. Wei, L. Tao, C. Zhang, and J. Sun. Salient object detection by composition. In *CVPR*, 2011.

[13] S. Goferman, L. Zelnik-Manor, and A. Tal. Context-aware saliency detection. *PAMI*, 2011.

[14] J. Hays and A. A. Efros. Scene completion using millions of photographs. In *SIGGRAPH*, 2007.

[15] X. Hou and L. Zhang. Saliency detection: A spectral residual approach. In *CVPR*, 2007.

[16] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *PAMI*, 20(11), 1998.

[17] H. Kang, A. A. Efros, M. Hebert, and T. Kanade. Image composition for object pop-out. In *ICCV Workshops*, 2009.

[18] D. G. Lowe. Object recognition from local scale invariant features. In *ICCV*, 1999.

[19] Y. Luo, J. Yuan, P. Xue, and Q. Tian. Saliency density maximization for efficient visual objects discovery. *IEEE Trans. Circuits Syst. Video Techn.*, 21(12), 2011.

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

[21] M. Muja. Flann, fast library for approximate nearest neighbors, 2011. `http://mloss.org/software/view/143/`.

[22] 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.

[23] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *IJCV*, 42(3), 2001.

[24] M. Pandey and S. Lazebnik. Scene recognition and weakly supervised object localization with deformable part-based model. In *ICCV*, 2011.

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

[26] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman. Labelme: A database and web-based tool for image annotation. *IJCV*, 77(1-3), 2008.

[27] A. Shrivastava, T. Malisiewicz, A. Gupta, and A. A. Efros. Data-driven visual similarity for cross-domain image matching. *ACM Trans. Graph.*, 30(6), 2011.

[28] P. Siva, C. Russell, and T. Xiang. In defence of negative mining for annotating weakly labelled data. In *ECCV*, 2012.

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

[30] A. M. Treisman and G. Gelade. A feature-integration theory of attention. *Cognitive psychology*, 12(1), 1980.

[31] K. van de Sande, J. Uijlings, T. Gevers, and A. Smeulders. Segmentation as selective search for object recognition. In *ICCV*, 2011.

[32] M. Wang, J. Konrad, P. Ishwar, K. Jing, and H. Rowley. Image saliency: From intrinsic to extrinsic context. In *CVPR*, 2011.

[33] H. Zhong, J. Shi, and M. Visontai. Detecting unusual activity in video. In *CVPR*, volume 2, 2004.