Pattern Recognition Letters
journal homepage: www.elsevier.com

Revisiting Unsupervised Object Localization: A Simple yet Efficient Method

H.L Praveen Raj[a,**], Shanmuganathan Raman[b]

[a]*Department of Computer Science and Engineering, National Institute of Technology Karnataka, Surathkal,India*
[b]*Electrical Engineering & Computer Science and Engineering, Indian Institute of Technology Gandhinagar, Palaj, Gandhinagar*

ABSTRACT

Object localization is an important problem in computer vision. Object localization in a 2D image is the task of estimating the precise bounding box around the object present in the image. In this paper, we propose an unsupervised approach towards localizing a single object in an image. Previously, many supervised and weakly supervised approaches were proposed, however, the major disadvantages of these techniques were, they required a huge amount of training data along with human annotations. In this work, we propose an unsupervised object localization algorithm which along with saliency information as well as spatial pyramid matching helps in efficient proposal selection from a huge set of object proposals. After this, a proposal grouping technique is proposed from which the final localization window is estimated. We tested our algorithm on two of the most used object recognition datasets namely, PASCAL VOC 2007 and PASCAL VOC 2012. The results show that our approach achieves an average CorLoc of 60.91% when evaluated on the PASCAL VOC 2007 dataset. The experiments show that our algorithm performs significantly better than other unsupervised approaches, and performs comparably to the state of the art weakly-supervised approach (which achieves an average CorLoc of 64.60%).

1. Introduction

Object localization is an important problem in computer vision. The task here is to estimate precise bounding box around the objects present in a 2D image. Because of the presence of intra-class variations, background clutter, occlusion, diversity in view points, object localization becomes a difficult and challenging task. Object localization is widely used in a variety of computer vision problems, such as, object detection, segmentation, separation of foreground from background, to name a few. This has been a prime area of research for a long time now. There are many approaches proposed previously, all of which can be classified into three major categories, namely (1) Fully-supervised, (2) Weakly-supervised and (3) Unsupervised approaches. Fully-supervised approaches [Brizard (2015); Villamizar et al. (2012)] involve training a prediction model using annotated images and other additional data. Even though these approaches give extremely high accuracy, they require strong supervision involving human annotations, and acquiring such quality annotated dataset is very difficult. Since these approaches require a lot of human resource, they are not suitable for resource deficient environments. The other major disadvantage is the inclusion of external error, due to the involvement of humans in the critical steps of the process (eg. training). In order to overcome the huge manual effort required to annotate each object in an image, weakly-supervised approaches [Cinbis et al. (2017); Shi et al. (2013); Teh et al. (2016)] use simpler, image level annotations which tell about the presence or absence of an object and about the class of an object present in the image. But these approaches too require a considerable amount of human effort.

In order to completely eliminate the involvement of humans in the process, the unsupervised approaches [Cinbis et al. (2017); Shi et al. (2013); Teh et al. (2016)] were proposed. The early unsupervised approaches were co-localization algorithms which tried to localize an object belonging to the same object class across multiple images [Grauman and Darrell (2006); Kim and Torralba (2009); Tang et al. (2014); Zitnick and Dollár (2014)] without any supervision. Since the co-localization algorithms try to localize object of same class across multi-

**Corresponding author:
e-mail: hlpr98@gmail.com (H.L Praveen Raj)

ple images, it imparts certain supervision to the process and can be solved using techniques like proposal matching [Cho et al. (2015)] and clustering [Tang et al. (2014)] across the images. Unlike these approaches, this paper presents an algorithm which localizes a single object instance in an image in a completely unsupervised manner.

To achieve unsupervised object localization, we start by extracting the object proposals. Object proposals are the candidate regions where an object might be present. We filter these proposals to obtain a handful of proposals having a higher probability of containing an object. The filtering of the proposals is achieved using the saliency map and saliency contrast. Finally, we group similar proposals together, find the similarity index (the measure of similarity among the elements of a group) for each group, find the best group and use its proposals to compute the final localization window. The main contributions of this paper can be summarized as follows,

- A completely unsupervised single object localization algorithm which completely avoids human involvement during the process.
- A method to score the object proposals and hence filter them, so that we are left with only those proposals that tightly surround the object.
- A proposal grouping method which would help in estimating the final localization window.

In Section 2, we have discussed previous approaches to object localization, which include weakly-supervised approaches, co-localization and co-segmentation approaches and the other fully-unsupervised approaches. Section 3 provides a complete and detailed description of the proposed approach. Section 4 talks about the experimentation process and evaluation criteria used for the evaluation of our approach. Here, we also provide a comparison between the performance of our approach on the PASCAL VOC 2007 dataset [Everingham et al. (a) (2007)] and the performance of the other state of the art approaches on the same dataset. Section 5 provides a conclusion to this manuscript.

2. Related Work

This section describes the previous works which address object localization. The approaches are listed in the decreasing order of supervision.

2.1. Weakly-Supervised approaches

Weakly supervised approaches can be divided broadly into, (1) Multiple instance learning (MIL) based, (2) CNN based and (3) Saliency detection based methods. A conventional way of solving this problem is using a MIL based technique. Here, an image is treated as a collection of object proposals. The idea is that, when an object is labeled positive, at least one of the object proposals must enclose the object [Cinbis et al. (2017); Shi and Ferrari (2016); Zitnick and Dollár (2014)]. However, the MIL based techniques lead to a non-convex optimization problem and tend to get stuck in local optima. CNN based techniques generally learn the object classifiers along with the localization features which help in the selection of the best object proposal [Cinbis et al. (2017); Russakovsky et al. (2015)]. [Cinbis et al. (2017)] use CNN to represent the features of object proposals, and hence use it as a knowledge transfer media for other tasks in the localization process. CNNs are also used for end-to-end learning, for example in [Bilen and Vedaldi (2016)]. Saliency detection automatically highlights the object in an image, and hence can be used for localization, as in [Lai and Gong (2016); Shimoda and Yanai (2016); Zhang et al. (2017)]. Weakly supervised approaches provide very good accuracy by utilizing very few resources.

2.2. Co-localization and co-segmentation

Co-segmentation is the problem, in which a common region among two images is segmented. This technique was first proposed by Rother et al. [Rother et al. (2006)], where they used Markov random fields and colour histograms to segment object common to two images. There has been a huge advancement of this technique since then and [Cho et al. (2010); Joulin et al. (2010); Rubinstein et al. (2013)] provide its application in a general case.

Co-localization is a technique similar to co-segmentation, but here the task is to localize a target object class common to both the images. Tang et al. [Tang et al. (2014)] proposed a discriminative clustering approach to localize a common object class in a set of noisy images. [Kim and Torralba (2009)] propose a technique, which analyses links to find the region of interest among the images. [Grauman and Darrell (2006)] proposed an approach that used partial correspondences and clustering of local features. Cho et al. [Cho et al. (2015)] proposed a bottom-up approach probabilistic approach with part based region matching, for multiple class object localization. Choe et al. [Choe et al. (2018)] uses GAN (generative adversarial network) for object co-localization. They use the fact that, GANs implicitly learn certain useful, unknown information about the distribution of data.

2.3. Unsupervised approaches

The fully-unsupervised instance of this problem is the most difficult version of it. It has been tackled by Vora and Raman [Vora and Raman (2018)] using iterative spectral clustering of the object proposals to naturally find the region of the object. [Vora and Raman (2018)] propose the use of iterative spectral clustering to narrow the set of all the object proposals into a set of proposals which tightly bound the object and later these proposals are grouped to obtain the final localization window.

3. Proposed Algorithm

In this section, we discuss the entire pipeline of our algorithm. The summary of the pipeline is shown in Figure 1. Each subsection here describes the different stages of the pipeline.

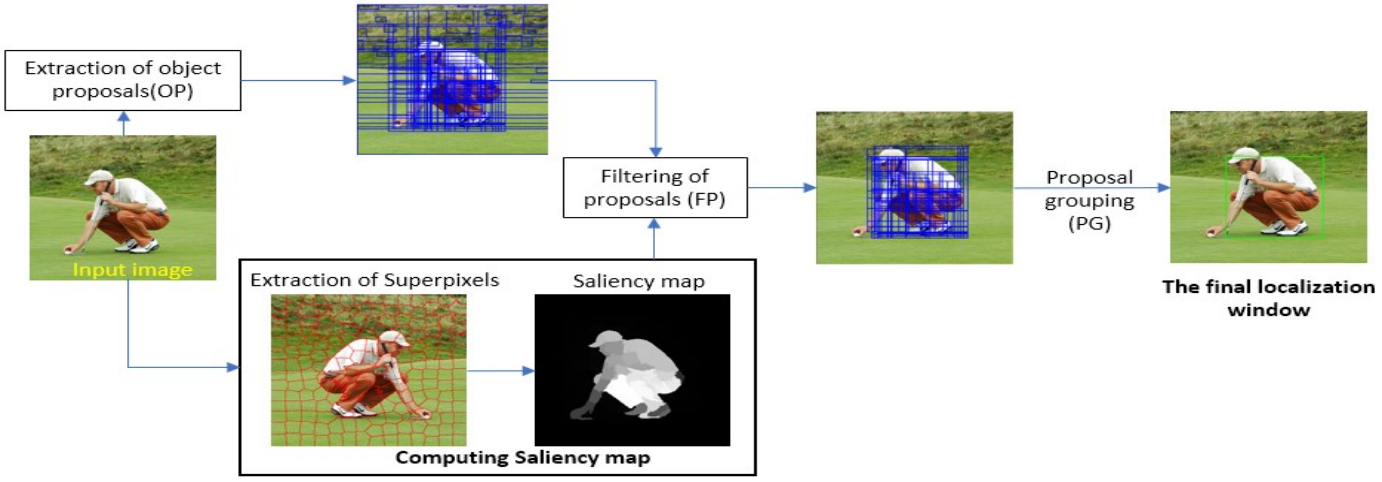


Fig. 1: An overview of the proposed algorithm.

3.1. Extraction of Object Proposals (OP)

Object proposals are the region boxes or the windows that have a higher probability of containing an object. Previously, people used to segment the images by partitioning all the pixels in the hope of localizing the object, but segmentation is a very difficult problem and has not been successfully solved, in completely automatic manner, yet. Hence, basing future steps of object localization on accurate segmentation is not a good strategy. Hence, the strategy shifted to generating not a single, fixed, perfect segmentation solution, but a number of probable segmentations (possibly 100s). Here we do not need to segment pixel by pixel, but just need to find the bounding box, where an object "might" exist. These proposed segments or bounding-boxes are called object proposals. Today object proposals are generally one of the initial steps in almost every approach for object localization and detection.

We obtain the object proposals using an off-the-shelf algorithm, which uses Randomized Prim's algorithm [Manen et al. (2013)]. This algorithm first segments the image into its superpixels and obtains a connectivity graph, G, from them. Then a set of groups, $S = \{g_1, g_2, ..., g_k\}$, of connected superpixels are obtained by applying Randomized Prim's algorithm on G. The bounding boxes of each group in S are then proposed as the object proposals, $O = \{P_1, P_2, P_3, ..., P_k\}$. Each proposal is simply a rectangular box.

3.2. Filtering of Proposals (FP)

After the obtaining the object proposals, we observe that the size of the set O is very large, generally $|O|$ is in the order of 1000s. And we also note that not every element in O is a viable bounding box solution since some are too small, some are too large, some which do not bound the object at all. Hence, after the extraction of the object proposals, $O = \{P_1, P_2, P_3, ..., P_k\}$, the next task is to score each proposal and then efficiently select a subset of O that have a high probability of containing an object.

We use the saliency measure of each pixel for scoring. The scoring process is as follows. First, we compute the saliency map of the image using the algorithm proposed by [Yang et al. (2013)]. This algorithm uses the superpixels for the generation of saliency map. We use the SLIC algorithm [Achanta et al. (2012)] for the generation of the superpixels since it is very fast. Thus we obtain a set of superpixels, $SP = \{sp_1, sp_2, ..., sp_N\}$. We restrict the maximum number of superpixels to be N = 200, as suggested in [Yang et al. (2013)]. Once the superpixels are obtained, the saliency map is generated according to the algorithm in [Yang et al. (2013)]. The saliency map contains saliency measure for each pixel, higher the saliency value higher is the probability that the pixel is part of an object. Here we calculate the saliency densities instead of just values in order to obtain the proposals as tight to the object as possible. Saliency density is helpful to choose the tighter proposals, in some cases. Let P_i and P_j be two object proposals such that $P_i \subset P_j$ and $Area(P_i) < Area(P_j)$ and $RS_i = RS_j$. Here it is evident that P_i is the tighter bounding box to the object and hence has a higher probability of the presence object as supposed to P_j.

We calculate the saliency measure of each object proposal region in O as:

$$RS_i = \frac{1}{Area(P_i)} \sum_{\forall p \in P_i} S_c(p) \ , \ \forall P_i \in O \ . \quad (1)$$

Here, $Area(P_i)$ is the area of the proposal P_i, $S_c(p)$ is the saliency of the pixel p that belongs to the proposal P_i. Area of the proposal P_i is the number of pixels enclosed by P_i.

We, then, calculate the saliency measure of the collection of all the superpixels adjacent to the respective proposal. Let us denote the set of all superpixels adjacent to the proposal P_i as $Adj(P_i)$. Hence, the saliency measure of $Adj(P_i)$ is:

$$AdjS_i = \frac{1}{Area(Adj(P_i))} \sum_{\forall p \in Adj(P_i)} S_c(p) \ , \ \forall P_i \in O \ . \quad (2)$$

Here, $Area(Adj(P_i))$ is the cumulative area of all the adjacent superpixels of the proposal P_i and $S_c(p_i)$ is the saliency of the pixel p that belongs to $Adj(P_i)$. Here, "area" implies the number of pixels a region encloses.

Finally, we calculate the score for each proposal by taking the difference between RS_i and $AdjS_i$ and also taking the $Area(P_i)$

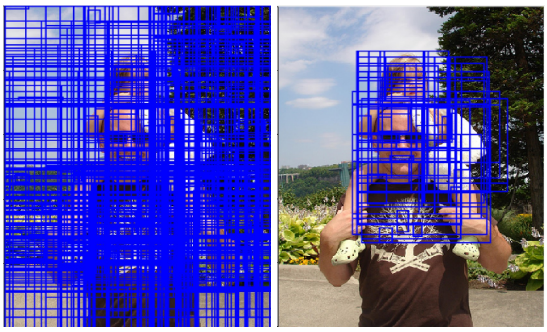
into consideration. We term this as saliency contrast (SC_i). The score for each proposal is given by:

$$SC_i = exp\left(\frac{Area(P_i)}{\sigma^2}\right)(RS_i - AdjS_i)\ ,\ \forall_{1\le i\le k} P_i \in O\ . \qquad (3)$$

Here, we note that SC_i is large when the object proposal has a higher saliency while its local context is not salient. This scoring system also takes into account the area of the object proposal to avoid the selection of proposals with a small area. We consider

$$\sigma = \frac{SDV(\{Area(P_i)\})}{10^{\tau}}$$

where, $\tau = maximum\ power\ of\ 10\ that\ divides\ \max(Area(P_i))$ and if W is a set of real numbers then, $SDV(W)$ provides standard deviation of the set W. After the scoring, the proposals with highest score are chosen to be the proposals with higher probability to contain the object and are considered in the following steps. Let these chosen object proposals be $C = \{op_1, op_2, op_3, ..., op_T\}$. Here, $T \le k$ and $C \subseteq O$, generally $|C| << |O|$. Another important thing to note here is that all the SC_i are 8-bit unsigned integers. We made this choice because the probability of choosing the right proposals increases when considered this way as supposed to when SC_i are floating point values. Further, integer values provide ease of use and higher speed of computation. Figure 2 describes the result of FP on an image from PASCAL VOC 2012.



(a) Object proposals extracted (b) Object proposals chosen after FP

Fig. 2: Result of Filtering of Proposals (FP) on an image for PASCAL VOC 2012 dataset.

3.3. Proposal Grouping (PG)

After filtering of proposals, we have C, a set of proposals which have the highest probability of containing an object. The next step involves grouping of one or more proposals in C to obtain a strict and precise bounding box around the object. To achieve this we use SIFT features and Spatial Pyramid Matching (SPM) algorithms. The idea behind this step is that both SIFT and SPM give a bag-of-orderless local feature descriptors which can further be used to perceive the similarity between two proposals which in turn would help in the grouping of proposals. This step is based on the SPM algorithm proposed by [Lazebnik et al. (2006)].

After filtering of proposals, for every proposal $op \in C$, we first extract the so called "weak features". These are the oriented edge points, i.e., the points whose gradient magnitude in a given direction exceeds a minimum threshold. Here we obtain the edge points at two scales and eight orientations, which amounts for a total of $M = 16$ channels. We designed these features to obtain a representation similar to a global SIFT descriptor [Lowe (1999); Lowe (2000)] of the image. For better discriminative power, we also utilize higher dimensional "strong features". These features are the SIFT descriptors of 16 x 16 pixels boxes, which are computed over a grid with 8 pixels' spacing. Intuitively, a dense image description is necessary to capture uniform regions such as sky or road surface or calm water (i.e. to deal with low-contrast regions). This step of extraction of features is based on [Lazebnik et al. (2006)]. Then, these feature vectors are used to generate a pyramid of $L = 2$ levels and $M = 200$ channels [Lazebnik et al. (2006)]. Later, we perform Spatial Pyramid Matching (SPM) proposed by [Lazebnik et al. (2006)] over the feature vectors obtained, which finally results in obtaining a 4200 dimensional histogram vector. Thus we obtain $SH = \{\vec{h_1}, \vec{h_2}, \vec{h_3}, ..., \vec{h_T}\}$, the set of all the Spatial pyramid histogram vectors and $\vec{h_i} \in \mathbb{R}^{4200}\ \forall \vec{h_i} \in SH$. We then construct a 4200 dimensional K-d tree, KD, using each element of SH as its node.

Next, we group the proposals in C into clusters of K similar proposals. Since the Spatial pyramid histogram vectors are a bag of orderless local features [Lazebnik et al. (2006)], it can effectively be used to measure the similarity between the given two proposals. Thus, we use each proposal in C as a seed and perform the operation, $g_i = \{\overline{op} \mid \overline{op} \in KNN(\vec{h_i})\}$. Here, i is the index of the seed proposal and $KNN(x)$ is the K-nearest neighbours of x. Here, we use $K = 10$ and it is found using KD for speeding up the searches. Thus we obtain the set of clusters $Gp = \{g_1, g_2, g_3, ..., g_T\}$. We, then, calculate the similarity index of each cluster in Gp, i.e the similarity among the elements of the cluster. The similarity is calculated as follows. For cluster $g_i \in Gp$,

$$SI_i = \sum_{j=1}^{K}\sum_{l=j+1}^{K}\left\|\left(\vec{h_{g_{i_j}}} - \vec{h_{g_{i_l}}}\right)\right\|_2\ . \qquad (4)$$

Here, g_{i_j} gives the index of the j^{th} proposal in the cluster g_i. Here, we note that the smaller the value of SI_i, the higher is the similarity between the constituent proposals.

Now, we consider top $c = 5$ clusters, i.e., the first c elements when $SI = \{SI_i\}$ is sorted in ascending order. We, then, find the smallest rectangle that encloses every element in these c clusters. This is the final localization window, P_{final} of the object. This choice was purely experimental, because, we obtained higher accuracy with this assumption as supposed to when we considered the average of the corner coordinates. An intuitive explanation for this is that, since the filtering of proposals step would have already given fairly tightly bounding proposals, this step would select the proposals which are much tighter. Thus, taking the average of the corner coordinates would provide a bounding box which would bound a small portion in the median of the object and not the whole object, which is undesirable.

Algorithm 1: Pipeline of the proposed algorithm

Input: Image I
Result: Localization window, P_{final} of the object
Set maximum number of superpixels, N.
Set K in KNN search and c for number of clusters to select.
Set L, the number of levels and M, the number of channels in the spatial pyramid.
Extract Object Proposals, $O = \{P_1, P_2, P_3, ..., P_k\}$, from image I.
Compute Superpixels, $SP = \{sp_1, sp_2, ..., sp_N\}$, of I.
Obtain Saliency Map of I.
Set σ.
foreach $P \in O$ **do**
    Compute saliency measure of the region, RS.
    Compute saliency measure of $Adj(P)$, $AdjS$.
    Compute saliency contrast, SC.
end
Select highest valued proposals in $\{SC\}$ to obtain, $C = \{op_1, op_2, op_3, ..., op_T\}$.
foreach $op \in C$ **do**
    Compute the spatial pyramid histogram vector, $\vec{h}$.
end
Construct 4200 dimensional k-d tree, KD, using elements of $SH = \{\vec{h_1}, \vec{h_2}, \vec{h_3}, ..., \vec{h_T}\}$.
foreach $op \in C$ **do**
    Compute K-nearest neighbours clusters (g) using KD, $g = \{\overline{op} \mid \overline{op} \in KNN(\vec{h_{op}})\}$
end
Compute similarity index, SI, for each cluster in $Gp = \{g_1, g_2, g_3, ..., g_T\}$.
Select the Top-c clusters, and compute the smallest rectangle that encloses each proposal in these clusters, thus obtaining P_{final}.
Return the Object Localization window, P_{final}.

4. Evaluation

We evaluated our algorithm on two of the widely used object recognition datasets, namely, PASCAL VOC 2007 [Everingham et al. (a) (2007)] and PASCAL VOC 2012 [Everingham et al. (b) (2012)]. Both these datasets contain real-world images. We have compared our results with various other state-of-the-art unsupervised [Vora and Raman (2018)] and weakly-supervised[Siva and Xiang (2011); Shi et al. (2013); Wang et al. (2014); Teh et al. (2016); Cinbis et al. (2017)] approaches. During our experiments, the parameters were set as follows: (1) Maximum number of superpixels, $N = 200$, (2) $K = 10$ in KNN search, (3) Number of Top-clusters, $c = 5$, (4) Number of Spatial pyramid levels, $L = 2$ and (5) Number of Spatial pyramid channels, $M = 200$. During our experiments, we noted that taking higher values of K and c does not change the accuracy all that much, but does take longer processing time. Thus, we went with this choice of parameters.

4.1. Evaluation criteria and run time

Following the previous works of object localization, we used CorLoc (Correct Localization) as a metric for accuracy. CorLoc is defined as, the percentage of images correctly localized in the complete dataset. An image is said to be correctly localized if the intersection-over-union ratio of the predicted localization window and the ground truth is greater than 50%, i.e. $\frac{area(P_{predicted} \cap P_{gt})}{area(P_{predicted} \cup P_{gt})} > 0.5$. Here, $P_{predicted}$ is the final localization window and P_{gt} is the ground truth of the same object. CorLoc evaluation criterion was proposed by Everingham et al. in [Everingham et al. (2010)]. Since our algorithm is able to localize only one object, we face an issue when multiple objects are present in the image, which is fairly common in PASCAL VOC datasets. In order to measure CorLoc in such situations, we rule that, if any one of the objects in the image is localized correctly (i.e., satisfies the CorLoc condition), then consider that for evaluation.

The algorithm was implemented in MATLAB-2017 and C++. We performed all our experiments in Windows 10 environment, on a computer with 7^{th}-Gen Intel Core-i7 processor and 940MX NVIDIA graphics card. The images considered were of resolution 500 x 375 and were from PASCAL VOC 2007 and PASCAL VOC 2012 datasets. The average time taken for object localization for an image was $\approx$ 22 sec. In order to reduce the run time, we have used certain faster data-structures like Hashmaps, K-d trees, and algorithms like Randomized Prim's and SLIC.

4.2. Experiments and results

The experiments were carried out using the images taken from two of the widely used object recognition datasets, namely PASCAL VOC 2007 [Everingham et al. (a) (2007)] and PASCAL VOC 2012 [Everingham et al. (b) (2012)]. Both these datasets contain challenging images taken in real-time scenarios with a considerable amount of occlusion, clutter and diverse view points. We chose to use these datasets in order to compare our results with other previous approaches since these datasets are the most popular ones in the community.

All the images in PASCAL VOC datasets have a resolution of 500 x 375. The images in these datasets are very diverse in terms of the classes of objects, view points, the location of the object, etc. PASCAL VOC datasets consist of 20 different classes. For our experiments, we considered the "train+val" images in PASCAL VOC datasets. The number of images in these image classes ranges from 96 (sheep) to 2008 (person) in the case of PASCAL VOC 2007 and from 303 (cow) to 4087 (person) in case of PASCAL VOC 2012. For an extensive benchmarking experiment, we considered all of these images during our experiments. The results of the experiments on PASCAL VOC 2007 dataset are presented in Table 1 and the results of experiments on PASCAL VOC 2012 are represented in Table 2. Sample results of localization on the PASCAL VOC 2007 and PASCAL VOC 2012 are given in Figure 3 and Figure 4, respectively.

Table 1: CorLoc(%) performance on PASCAL VOC 2007

Class →	Aero	Cycle	Bird	Boat	Bottle	Bus	Car	Cat	Chair	Cow	Table	Dog	Horse	Bike	Person	Plant	Sheep	Sofa	Train	TV	Avg
CorLoc(%)	56.11	34.0	50.0	61.67	21.30	61.48	64.69	62.93	33.71	64.4	52.22	65.55	70.96	56.82	69.97	32.22	69.91	53.33	65.29	70.24	**60.91**

Table 2: CorLoc(%) performance on PASCAL VOC 2012

Class →	Aero	Cycle	Bird	Boat	Bottle	Bus	Car	Cat	Chair	Cow	Table	Dog	Horse	Bike	Person	Plant	Sheep	Sofa	Train	TV	Avg
CorLoc(%)	57.6	46.0	51.47	62.61	21.05	67.80	61.4	65.79	35.56	60.47	53.33	69.57	71.67	61.54	70.2	34.44	66.67	55.06	69.94	70.12	**59.81**

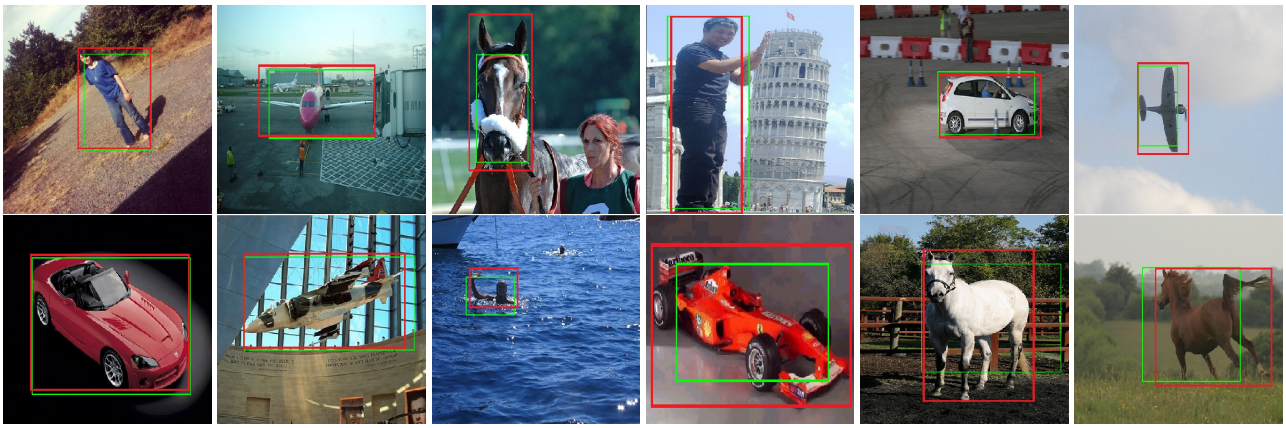


Fig. 3: Results of localization on PASCAL VOC 2007 dataset, here Green box = Predicted window, Red box = Ground truth.

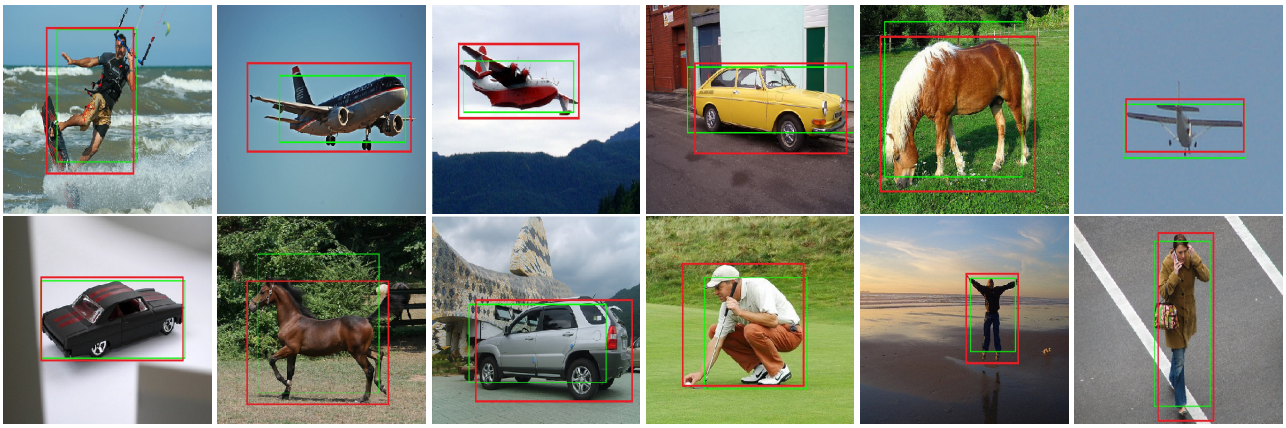


Fig. 4: Results of localization on PASCAL VOC 2012 dataset, here Green box = Predicted window, Red box = Ground truth.

4.3. Comparative analysis

We compared our approach with other state-of-the art weakly-supervised [Siva and Xiang (2011); Shi et al. (2013); Wang et al. (2014); Teh et al. (2016); Cinbis et al. (2017)] and unsupervised [Vora and Raman (2018)] approaches. The comparison was based on the CorLoc scores achieved on PASCAL VOC 2007. The statistics for the comparison were obtained from the quoted results in the respective papers. Table 3 contains a comparison between our approach and other state-of-the-art unsupervised and weakly-supervised approaches. In Table 3, the third column represents the type of approach, namely weakly-supervised (WS) or unsupervised (US), and the fourth column represents the types of data utilized during training. Here "P" represents positive images, "N" represents negative images and "A" represents certain addition data inputs. From the comparative analysis, it can be inferred that our approach performs better than [Siva and Xiang (2011); Shi et al. (2013); Wang et al. (2014); Cinbis et al. (2017); Vora and Raman (2018)] and performs comparably to the state-of-the-art weakly-supervised approach [Teh et al. (2016)] (CorLoc score = 64.60%). The better performance of our algorithm is mainly due to the fact that our algorithm

Table 3: Comparison of performance on PASCAL VOC 2007 dataset

Method	Type	Data	Avg CorLoc(%)
Siva and Xiang (2011)	WS	P+N	30.4
Shi et al. (2013)	WS	P+N	36.3
Cinbis et al. (2017)	WS	P	47.3
Wang et al. (2014)	WS	P+N+A	48.5
Teh et al. (2016)	WS	P	64.60
Vora and Raman (2018)	US	-	35.08
Ours	US	-	**60.91**

concentrates on capturing and utilizing the properties (such as gradients, saliency, and similarity) of the pixels of the object and the pixels adjacent to the object, unlike the other methods which concentrate on capturing the features of the object and their effects on the pixels. Our algorithm works in a bottom-up fashion, where it ratifies the presence of the object in a region if the pixels in that region follow certain properties (as explained in Section 2). While the other approaches function in a top-down fashion searching for regions containing the features of an object (which they would have already learned through training). The performance enhancement can also be attributed to the better extraction of object proposals and saliency maps.

5. Conclusion

We have presented an efficient yet simple method for unsupervised single object localization. Our approach uses object proposals, saliency map and spatial pyramid matching to obtain the final localization window. The experimental results show that our approach achieves an average CorLoc of 60.91% when evaluated on the PASCAL VOC 2007 dataset. The experiments also show that our algorithm performs significantly better than other unsupervised approaches, and performs comparably to the state of the art weakly-supervised approach. The performance of the algorithm is improvable using better saliency map and object proposal algorithms. This algorithm can be directly used in other computer vision pipelines, such as the pipeline for object detection. We plan to extend this work to unsupervised object localization in multiple object instance cases.

References

Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., Süsstrunk, S., et al., 2012. Slic superpixels compared to state-of-the-art superpixel methods. IEEE transactions on pattern analysis and machine intelligence 34, 2274–2282.

Bilen, H., Vedaldi, A., 2016. Weakly supervised deep detection networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2846–2854.

Brizard, A.J., 2015. Notes on the weierstrass elliptic function. arXiv preprint arXiv:1510.07818 .

Cho, M., Kwak, S., Schmid, C., Ponce, J., 2015. Unsupervised object discovery and localization in the wild: Part-based matching with bottom-up region proposals, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1201–1210.

Cho, M., Shin, Y.M., Lee, K.M., 2010. Unsupervised detection and segmentation of identical objects, in: Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, IEEE. pp. 1617–1624.

Choe, J., Park, J.H., Shim, H., 2018. Unsupervised object localization using generative adversarial networks. arXiv preprint arXiv:1806.00236 .

Cinbis, R.G., Verbeek, J., Schmid, C., 2017. Weakly supervised object localization with multi-fold multiple instance learning. IEEE transactions on pattern analysis and machine intelligence 39, 189–203.

Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A., 2010. The pascal visual object classes (voc) challenge. International journal of computer vision 88, 303–338.

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

Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A., b. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html.

Grauman, K., Darrell, T., 2006. Unsupervised learning of categories from sets of partially matching image features, in: Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on, IEEE. pp. 19–25.

Joulin, A., Bach, F., Ponce, J., 2010. Discriminative clustering for image co-segmentation, in: Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, IEEE. pp. 1943–1950.

Kim, G., Torralba, A., 2009. Unsupervised detection of regions of interest using iterative link analysis, in: Advances in neural information processing systems, pp. 961–969.

Lai, B., Gong, X., 2016. Saliency guided dictionary learning for weakly-supervised image parsing, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3630–3639.

Lazebnik, S., Schmid, C., Ponce, J., 2006. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories, in: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), IEEE. pp. 2169–2178.

Lowe, D.G., 1999. Object recognition from local scale-invariant features, in: Computer vision, 1999. The proceedings of the seventh IEEE international conference on, IEEE. pp. 1150–1157.

Lowe, D.G., 2000. Towards a computational model for object recognition in it cortex, in: International Workshop on Biologically Motivated Computer Vision, Springer. pp. 20–31.

Manen, S., Guillaumin, M., Van Gool, L., 2013. Prime object proposals with randomized prim's algorithm, in: Proceedings of the IEEE international conference on computer vision, pp. 2536–2543.

Rother, C., Minka, T., Blake, A., Kolmogorov, V., 2006. Cosegmentation of image pairs by histogram matching-incorporating a global constraint into mrfs, in: Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on, IEEE. pp. 993–1000.

Rubinstein, M., Joulin, A., Kopf, J., Liu, C., 2013. Unsupervised joint object discovery and segmentation in internet images, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1939–1946.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al., 2015. Imagenet large scale visual recognition challenge. International Journal of Computer Vision 115, 211–252.

Shi, M., Ferrari, V., 2016. Weakly supervised object localization using size estimates, in: European Conference on Computer Vision, Springer. pp. 105–121.

Shi, Z., Hospedales, T.M., Xiang, T., 2013. Bayesian joint topic modelling for weakly supervised object localisation, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 2984–2991.

Shimoda, W., Yanai, K., 2016. Distinct class-specific saliency maps for weakly supervised semantic segmentation, in: European Conference on Computer Vision, Springer. pp. 218–234.

Siva, P., Xiang, T., 2011. Weakly supervised object detector learning with model drift detection, in: Computer Vision (ICCV), 2011 IEEE International Conference on, IEEE. pp. 343–350.

Tang, K., Joulin, A., Li, L.J., Fei-Fei, L., 2014. Co-localization in real-world images, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1464–1471.

Teh, E.W., Rochan, M., Wang, Y., 2016. Attention networks for weakly supervised object localization., in: BMVC.

Villamizar, M., Andrade-Cetto, J., Sanfeliu, A., Moreno-Noguer, F., 2012. Bootstrapping boosted random ferns for discriminative and efficient object classification. Pattern Recognition 45, 3141–3153.

Vora, A., Raman, S., 2018. Iterative spectral clustering for unsupervised object localization. Pattern Recognition Letters 106, 27–32.

Wang, C., Ren, W., Huang, K., Tan, T., 2014. Weakly supervised object localization with latent category learning, in: European Conference on Computer Vision, Springer. pp. 431–445.

Yang, C., Zhang, L., Lu, H., Ruan, X., Yang, M.H., 2013. Saliency detection via graph-based manifold ranking, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3166–3173.

Zhang, D., Meng, D., Zhao, L., Han, J., 2017. Bridging saliency detection to weakly supervised object detection based on self-paced curriculum learning. arXiv preprint arXiv:1703.01290 .

Zitnick, C.L., Dollár, P., 2014. Edge boxes: Locating object proposals from edges, in: European conference on computer vision, Springer. pp. 391–405.