Pattern Recognition Letters 106 (2018) 27–32

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Iterative spectral clustering for unsupervised object localization

Aditya Vora*, Shanmuganathan Raman

Indian Institute of Technology Gandhinagar, Gandhinagar, Gujarat, India

ARTICLE INFO

Article history:
Received 24 June 2017
Available online 21 February 2018

MSC:
41A05
41A10
65D05
65D17

Keywords:
Object localization
Spectral clustering
Unsupervised localization

ABSTRACT

This paper addresses the problem of unsupervised object localization in an image. Unlike previous supervised and weakly supervised algorithms that require bounding box or image level annotations for training classifiers, we propose a simple yet effective technique for localization using iterative spectral clustering. This iterative spectral clustering approach along with appropriate cluster selection strategy in each iteration naturally helps in searching of object region in the image. In order to estimate the final localization window, we group the proposals obtained from the iterative spectral clustering step based on the perceptual similarity, and average the coordinates of the proposals from the top scoring groups. We benchmark our algorithm on challenging datasets like Object Discovery and PASCAL VOC 2007, achieving an average CorLoc percentage of 51% and 35% respectively which is comparable to various other weakly supervised algorithms despite being completely unsupervised.

1. Introduction

Object localization is an important computer vision problem where the task is to estimate precise bounding boxes around all categories of the objects present in the given image. Due to the intra-class variations, occlusion, and background clutter present in real-world images, this becomes a challenging problem to solve. Object localization is useful in several image understanding tasks like separating foreground from background, object recognition, and segmentation. Previous fully-supervised approaches relied on sliding window search in order to search for an object in an image. Because of their inefficiency in terms of speed, several efficient sub-window search algorithms were proposed which work quite well in localizing the object in an image [21]. However, these techniques require strong supervision in the form of manually-annotated bounding boxes on locations of all the object categories in an image. Acquiring such human annotations for training accurate classifiers is a cumbersome task and is prone to human errors. As a result, supervised techniques for object localization do not prove to be useful in resource restricted settings. In order to overcome the huge manual efforts required in annotating objects in the image in supervised learning algorithms, several weakly supervised approaches were proposed. Rather than bounding box annotations of target instances, weakly supervised learning focuses on image level labeling which is based on the presence/absence of target object instances in an image [9,15,18,24,36]. Though these techniques work well in terms of localization accuracy, they still require human annotation efforts especially when the training data is large.

In an effort to make the task of object localization completely unsupervised, various object co-localization algorithms were proposed which try to localize an object across multiple images [6,16,19,29,35] without any supervision. As co-localization algorithms assume that each image has the same target object instance that needs to be localized [16,19], it imports some sort of supervision to the entire localization process thus making the entire task easier to solve using techniques like proposal matching [6] and clustering [37] across images. In contrast to these works, the work presented in this paper focuses on localizing a single object instance in an image in a completely unsupervised fashion. To the best of our knowledge, there is no previous work that tries to solve this problem in an unsupervised way. Unlike previous colocalization algorithms, in this work we do not make any assumptions like that in co-localization algorithms, thus making the entire problem more practical as well as challenging. Further, it is an important problem to be addressed because of the following reasons: (1) The proposed work is an unsupervised approach for object localization. As mentioned previously, all the manual labour required in annotating data with accurate bounding box regions around the target object instances will not be required, which saves the resources as well as training time. (2) Apart from being fully automatic and unsupervised, our technique is easy to fit in the current state-of-the-art object recognition pipelines like RCNN [14]. Thus unlike the current system, we do not need to classify each of the thousands of object proposals individually. Instead, we can localize the object

* Corresponding author.
E-mail address: aditya.vora@iitgn.ac.in (A. Vora).

https://doi.org/10.1016/j.patrec.2018.02.012
0167-8655/

directly in the input test image and then provide this localized object to the CNN pipeline that will classify the object appropriately. Such a type of functionality is not available with co-localization techniques. This restricts their applicability in real-world scenarios.

To solve this problem in an unsupervised manner, we start with extracting thousands of object proposals from the input image using an off-the-shelf object proposal algorithm [38,44]. We then try to filter out a number of object proposals effectively in such a way that after the entire proposal filtering process, a good set of object proposals that contain the object are retained. In order to achieve this, we formulate the problem as an undirected graph problem and perform spectral clustering on the constructed graph. This will split the set of proposals that can be discriminated based on the selected feature space. However, one iteration of spectral clustering would not be enough to filter the proposals by a significant amount. As a result, we repeat the process for a number of iterations after selecting an appropriate cluster for subsequent partitioning. We compute a cluster score after each iteration and select the cluster that has a higher score for further partitioning in the next iteration and discard all the proposals in the cluster which has a lower score. After this filtering step, we then estimate the final localized window by grouping the proposals based on the perceptual similarity among the proposals. We then pick top scoring groups and take the mean of coordinates of proposals present in that groups in order to get the final localized window.

The main contributions of this paper are summarized as follows: (1) A completely unsupervised object localization algorithm for an image containing a single object is presented and benchmarked on challenging datasets like Object Discovery [27] and PASCAL VOC 2007 [12]. (2) An iterative spectral clustering approach along with an appropriate cluster selection strategy is proposed which naturally helps to search for an object region in an image. This entire process takes place in a completely unsupervised fashion. (3) Proposal grouping technique is proposed which helps in estimating the final localized window in the image.

2. Related work

Previous works in object localization are described below based on the decreasing order of supervision.

Supervised approaches for object localization involves *sliding window* approaches that apply a classifier subsequently to subimages, thus obtaining a classification map. Indicator of the object region is obtained from the classification map as the region with the maximum score. As an average size image will have a lot of pixels, scanning all of them and deriving the classification map is a computationally expensive task. Lampert et al. [21] and Villamizar et al. [40] tried to come up with a more efficient solution by proposing an *efficient subwindow search* for object localization which does not suffer from the above mentioned drawbacks. This scheme helps to optimize the quality function over all the possible subregions of the image with fewer classification evaluations and thus making the algorithm run in linear time or faster. Blaschko and Lampert [3] introduced the concept of *global and local context kernels* that tries to combine different context models into a single discriminative classifier. Sermanet et al. [30] proposed an integrated framework for image classification, localization and detection. They efficiently implemented a multiscale and sliding window approach within a convolutional neural network (CNN). They treat localization as a regression problem where the final layer is involved in predicting the coordinates of the bounding box. This entire system is trained end-to-end with bounding box annotations from the ImageNet dataset [20].

Weakly-supervised approaches for localization can be divided into 4 categories: (i) Exhaustive search technique [18,24,25,43], (ii) multiple instance learning [7,13,26,33,39,41], (iii) inter-intra-class modelling [9,10,28,34,42], and (iv) topic model [32,35]. Exhaustive search techniques try to learn discriminative sub-window classifiers from the weakly labelled data and then based on the scores of the most discriminative local regions of the image, they try to estimate the final localization window. Multiple-instance learning approaches try to learn various object categories from the bag of positive and negative labelled images. Different multiple-instance learning algorithms try to exploit various aspects associated with the image. For example, Wang et al. [41] tries to model the latent categories of the image like sky and grass in order to improve the overall localization accuracy. Vijayanarasimhan and Grauman [39] proposed an alternative learning approach where they trained robust category models from images returned by keyword-based search engines. In order to improve the quality of the object regions, along with the inter-class models, researchers also model intra-class relations to improve the similarity of the regions within the same object class.

Co-localization approaches: Tang et al. [37] proposed an image-box formulation for solving object co-localization problem, where they simultaneously localize object of the same class across a set of images. Cho et al. [6] generalized the task of object localization by relaxing the condition that each image should contain the object from the same category. Kim and Torralba [19] proposed an iterative link analysis technique in order to estimate the ROI in the image. Grauman and Darrell [16] proposed a colocalization approach based on the partial correspondences and the clustering of local features.

Iterative Spectral Clutering: There are few works in the literature about recursive spectral bipartitioning [1,2,17] which have tried to come up with many variants of spectral partitioning approach. But most of the works target VLSI CAD applications [1] because of which it is difficult to model these algorithms for computer vision problems. Moreover, conventional spectral partitioning approaches has proved to give excellent results for many computer vision problems [31] because of which we build our iterative spectral partitioning approach over simple spectral clustering.

3. Proposed approach

In this section, we describe the entire pipeline of our approach. The summary of the pipeline is shown in Fig. 1.

3.1. Object proposals extraction and scoring

We generate object proposals from the input image using an off-the-shelf object proposal algorithm known as the EdgeBoxes [44], which generates object proposals based on the edge information present in the image. After the extraction of object proposals $B = \{b_1, b_2, \ldots, b_N\}$ from the image I, we score each proposal based on the probability that the region contains an object. Here, we extract $N = 1000$ object proposals. Edgeboxes algorithm ranks each proposal based on objectness score s_{obj}, which are computed from the edge contour information within the proposal. We combine this objectness score of each proposal with the saliency score in order to compute the overall score of each proposal. In order to do this, we compute the saliency map S of the input image I using the saliency algorithm proposed by Margolin et al. [23]. From the saliency map S, we compute the average saliency score for each object proposal s_{sal}. After this, the overall score s_i for a proposal is computed as $s_i = s_{obj} \times s_{sal}$, where $i = 1$ to N. As a result, proposals that have high objectness score and those that cover the salient region of the image will have high overall score.