Signal Processing: Image Communication 56 (2017) 20–27

Contents lists available at ScienceDirect

Signal Processing: Image Communication

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

Improving object proposals with top-down cues

Wei Li[a,*], Hongliang Li[a], Bing Luo[a], Hengcan Shi[a], Qingbo Wu[a], King Ngi Ngan[a,b]

[a] *School of Electronic Engineering, University of Electronic Science and Technology of China, Chengdu, China*

[b] *Department of Electronic Engineering, The Chinese University of Hong Kong, ShaTin, Hong Kong*

ARTICLE INFO

Keywords:
Object proposals
Object detection
Object recognition

ABSTRACT

The generation of object proposals plays an important role in object detection. Most existing methods produce object proposals by using bottom-up cues, such as closed contour or superpixel. In this paper, we propose a novel method to improve the ranking of object proposals by combining bottom-up cues with top-down information of objectivity. Firstly, we utilize the bottom-up method to generate initial object proposals of the given test image. Then we retrieve its top-k similar images from training images set. Considering both appearance and spatial similarity between initial object proposals and the ground truth bounding boxes of these top-k similar images, we obtain the top-down guided scores of initial object proposals. Finally, the refined score of each initial object proposal is modeled as a fusion of the bottom-up score and the top-down score. Experiments show that our method achieves better performance compared with the state-of-art on the Pascal VOC2007 dataset.

1. Introduction

Object detection is an important but very challenging research field in computer vision. “Sliding window” is the common paradigm to address this problem, which is to obtain the detector scores by exhaustively searching all the bounding boxes over all positions, scales and aspect ratios [1,2]. However, this method is time-consuming and computationally expensive.

Instead of handling with exhaustive sliding windows, the paradigm of “recognition using regions” was proposed [3], which firstly extracts object proposals, then utilizes discriminative classifier to determine which proposal contains an object. Object proposals refer to a series of bounding boxes or segment regions which have high probability to contain objects in one image. This paradigm not only avoids exhaustive scanning windows across images search, but also achieves better performance by using the more sophisticated learning model, such as R-CNN [4]. This paradigm was successfully used in current framework of object detection [4–7].

A lot of methods have been proposed to generate object proposals [8–18], which can be roughly divided into two paradigms: superpixels-based methods and windows-based methods. The main idea of the superpixels-based methods is to design different strategies of superpixel merging or image over-segmentation [11,10,12–15]. For the windows-based methods, after generating lots of candidate windows, the general process is to score each candidate window according to how likely it contains an object and removes the low score windows [8,9,17,18]. In order to ensure high recall, these methods produce a certain number of proposals and give a confidence score for each proposal by bottom-up cues and they have made great progress in object-proposals generation.

However, as mentioned in [19,20], purely bottom-up processes have difficulty in generating perfect object proposals and they conclude that the top-down framework will likely play a more core role. DeepBox [21] demonstrates that there are more other cues to the possibility of one proposal containing objects than bottom-up grouping or saliency. DeepBox proposes the four layer convolutional neural networks to rerank proposals from a bottom-up method. Indeed, the bottom-up cues are not sufficient to indicate objectness. For instance, some background elements like murals, railing, windows, sky also have a closed contour, so they may score higher than the objects. We believe the top-down object information will improve the ranking of object proposals while decrease the ranking of non-object proposals.

In addition, the power of the object proposals is critically dependent on the efficiency and accuracy. Other than speed, a higher localization accuracy is normally required in some occasions, e.g., the task of offline video analysis, action detection. In practical term, our goal is to take a further step towards improving the accuracy. The aim of this work is to boost the bounding box detection AP performance across a wide range of Intersection-over-Union thresholds.

In this paper, we propose a method to improve the ranking of object proposals by combining bottom-up initial bounding boxes with top-down generic object information. Given the input test image, we first retrieve its top-K similar images from training dataset by a KNN-based similarity search and use bottom-up cues to generate initial bounding boxes. It is motivated from the observation that these top-K similar

[*] Corresponding author.
E-mail address: weili.cv@gmail.com (W. Li).

http://dx.doi.org/10.1016/j.image.2017.04.006
Received 11 August 2016; Received in revised form 7 February 2017; Accepted 10 April 2017
Available online 13 April 2017

images may contain the objects with the same category, so they can provide some useful information to describe objects further. Then, the ranking of initial bounding boxes can be aided by using the annotated object information of these similar images. We regard the annotated object information as the top-down cues because they indicate the high-level semantic property of the image, such as the appearance and location of the object. At last, the top-down score can be computed based on similarity between initial bounding boxes and ground truth bounding boxes of these top-K similar images. We design the score function to evaluate the similarity considering both appearance and spatial consistency. Our contributions can be summarized as two-fold:

- We improve the ranking of object proposals by combining bottom-up cues and top-down generic object information. Experimental results are demonstrated to show the efficiency of our proposed method.
- We propose a simple and effective matching strategy to mining useful top-down cues for improving performance through both appearance and spatial similarity. Finally, the refined confidence score of each bounding box is achieved by fusing the bottom-up and top-down score.

This paper is organized as follows: The related work is briefly described in Section 2. Section 3 details on the proposed method. Experimental results are provided in Section 4 to demonstrate the effectiveness of our method, and Section 5 summarizes the proposed method and the following work.

2. Related work

In the past few years, machine learning algorithms [22,23] are widely used in image classification and object detection. A number of methods have been presented to generate object proposals from an image. Most of them can be classified into two categories, i.e., superpixels-based methods [11,10,12–15] and windows-based methods [8,9,17,18]. Recently, deep learning methods have been applied to generate object proposals [24–26,21]. We only give a brief review on these methods. A good overview can be found in [19,20].

2.1. Superpixels-based methods

The superpixels-based methods either merge superpixels or segment for multiple seed regions. SS [13,14] first generates superpixel [27], then merges them according to pre-defined similarity distance in a greedy manner. RP [28] generates object proposals based on a randomized Prim's algorithm with the connectivity graph of an image's superpixels. CPMC [11,12] initializes the seeds randomly and uses graph cut to segment the image for many times, then designs some features for ranking object proposals. In addition, some methods combine the shape or contour information to extract object proposals. MCG [16] combines multiscale segmentation into object proposals by exploring combinatorial space. The grouping-based methods can generate proposals with high quality, but with high computation cost because the process of merging superpixels or aggregating hierarchical segmentation is time-consuming.

2.2. Windows-based methods

Objects tend to have a closed contour edge. The windows-based methods design different cues to score candidate boxes. Some typical methods are as follows: Objectness [8,9] first selects initial set of proposals from salient image locations, then estimates the proposal score by various cues, such as color, edges, location, size and superpixel straddling. In BING [17], Cheng et al. train a linear classifier to calculate the proposal score based on binarized gradient features while EB [18] calculates the proposal score by the number of edges in the box and minuses those that are members of contours that overlap the box's boundary. Although the windows-based methods can achieve faster computation then the superpixels-based methods, they always face the localization bias owing to the initial random sampling and lack effective cues to describe objects. With the Intersection-over-Union (IoU) increasing, recall will decline faster. In order to solve the localization bias, Chen [29] et al. propose an effective and fast approach to improve the quality of object proposals with Multi-Thresholding Straddling Expansion (MTSE). This approach can be viewed as a box refinement post-processing. In our work, we introduce top-down object information to aid bottom-up cues.

Given an input image, most deep-based approaches directly produce object proposals with confidence scores [24–26,21]. Multibox [24] directly predicts a set of class-agnostic bounding boxes by a saliency-inspired neural network model. In [25], Pedro et al. designs a discriminative convolutional network to generate segmentation proposals directly from image pixels. SSPB [26] first exploits deep CNN-SPP features, the EdgeBoxes score and the BEV descriptor, then extracts proposals by sparsity-inducing group normalized SVM. Our approach also utilizes deep-based feature, but we use it to retrieve the top-K similar images of the test image and the bounding box pair. DeepBox [21] uses a lightweight four-layer convolutional neural networks (CNNs) to rerank proposals from a bottom-up method. Our work is related to DeepBox in ranking object proposals. Although our method also utilizes the bottom-up method to produce initial bounding boxes, the biggest difference is that we score each bounding box by combining the bottom-up with top-down cues. Deep-based approaches improve the quality of proposals by directly using the neural network model, while our approach utilizes top-down cues by establishing a matching way.

3. The proposed method

In this section, we describe our method for ranking object proposals. The output of our model is to refine the ranking of the initial bounding boxes given by bottom-up cues. The confidence scores of bounding boxes are obtained by fusing the bottom-up scores and the top-down scores. The first term is obtained by bottom-up cues and the second term is derived from measuring the top-down guided similarity of a set of bounding boxes. The entire flowchart of our proposed method is illustrated in Fig. 1.

3.1. Bottom-up score

In the current literature, most models focus on the bottom-up mechanism for the object proposals generation and good results can be achieved by exploiting various cues, such as color contrast, superpixels straddling, or closed contour. Our proposed approach is compatible and can be applied to any bottom-up method for generating the initial bounding boxes, such as the windows-based methods and superpixels-based methods.

We define the initial bounding boxes $\{b_i\}_{i=1}^{T}$ of the test image represented as I, where T represents the total number of the initial bounding boxes. The initial score of the i-th bounding box b_i can be represented as $\boldsymbol{S}_{bu}(b_i)$. $\boldsymbol{S}_{bu}(b_i)$ can be computed from the bottom-up mechanism by exploiting various cues, such as color contrast, superpixels straddling, or closed contour, ie,

$$\boldsymbol{S}_{bu}(b_i) = \boldsymbol{R}(f(b_i)) \tag{1}$$

where $\boldsymbol{R}$ represents the defined ranking function and $f(b_i)$ represents the regional feature of the i-th bounding box b_i. Motivated by their promising performance and efficient computation, in our experiment we generate initial bounding boxes with ranking by using Edge Boxes [18] or Multiscale Combinatorial Grouping [16,29].