Signal Processing: Image Communication 53 (2017) 110–122

Contents lists available at ScienceDirect

Signal Processing: Image Communication

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

Robust object proposals re-ranking for object detection in autonomous driving using convolutional neural networks

Cuong Cao Pham, Jae Wook Jeon*

College of Information and Communication Engineering, Sungkyunkwan University, Suwon, Republic of Korea

ARTICLE INFO

Keywords:
Object proposals
Autonomous driving
Object detection
Convolutional neural networks
Stereo vision.

ABSTRACT

Object proposals have recently emerged as an essential cornerstone for object detection. The current state-of-the-art object detectors employ object proposals to detect objects within a modest set of candidate bounding box proposals instead of exhaustively searching across an image using the sliding window approach. However, achieving high recall and good localization with few proposals is still a challenging problem. The challenge becomes even more difficult in the context of autonomous driving, in which small objects, occlusion, shadows, and reflections usually occur. In this paper, we present a robust object proposals re-ranking algorithm that effectivity re-ranks candidates generated from a customized class-independent 3DOP (3D Object Proposals) method using a two-stream convolutional neural network (CNN). The goal is to ensure that those proposals that accurately cover the desired objects are amongst the few top-ranked candidates. The proposed algorithm, which we call DeepStereoOP, exploits not only RGB images as in the conventional CNN architecture, but also depth features including disparity map and distance to the ground. Experiments show that the proposed algorithm outperforms all existing object proposal algorithms on the challenging KITTI benchmark in terms of both recall and localization. Furthermore, the combination of DeepStereoOP and Fast R-CNN achieves one of the best detection results of all three KITTI object classes.

1. Introduction

Developing autonomous driving systems that can assist drivers in making decisions is one of the most active and challenging research areas [1]. The goal is to improve safety, reduce traffic accidents, and move closer towards fully autonomous cars and intelligent transportation systems. Among the solutions that have been developed over the past few years, the computer vision-based approach offers the most cost effective solution as it uses cameras rather than other types of more costly sensors. In this paper, we focus on object detection in autonomous driving.

State-of-the-art object detectors employ the exhaustive "sliding window" paradigm [2–4], in which a large number of bounding boxes generated from various scales and ratios is used for classification. This approach has been widely used as a standard object detection framework for many years. Recently, the use of more sophisticated and powerful classifiers [5–11] has improved detection accuracy. However, given the significant increase of computation time per window, the use of the "sliding window" paradigm has become infeasible. An alternative approach with the use of "object proposals" has been successfully introduced to gain both computational efficiency and high detection accuracy [12–15]. The key idea is to generate a moderate set of candidate bounding box proposals that are likely to contain objects and use this set for further classification instead of searching for objects at every image location and scale. This approach facilitates the ease of detection and can also improve accuracy by pruning away false positives before classification.

Since its remarkable discovery [12–15], the development of object proposals has quickly evolved since many methods have been innovated and improved. The goal is to introduce an algorithm that is able to achieve high recall and good localization. Extensive survey and evaluation can be found in [16,17]. In this paper, we briefly outline the existing work and review current methods not covered in [16,17]. Existing methods can be categorized into two main approaches according to their strategy used for generating proposals: grouping and scoring [17]. Grouping approaches typically generate multiple segments that are likely to contain objects by merging similar small regions based on diverse cues. On the other hand, scoring approaches tend to be faster by first initializing a set of bounding box proposals, and then scoring each proposal using an objectness function.

High recall and good localization are the most important properties of an object proposals algorithm [18,19,17]. A robust generator must be

* Corresponding author.
E-mail addresses: cuongpc@skku.edu (C.C. Pham), jwjeon@yurim.skku.ac.kr (J.W. Jeon).

http://dx.doi.org/10.1016/j.image.2017.02.007
Received 12 November 2016; Received in revised form 26 January 2017; Accepted 16 February 2017
Available online 20 February 2017

able to obtain high recall across various intersection over union (IoU) thresholds using a modest number of proposals, ranging from hundreds to a few thousands per image. While each existing algorithm has its own strengths and weaknesses, all algorithms have two common limitations. First, the proposals they produce are not well-localized since the recall drops significantly as the IoU threshold increases, especially for scoring approaches. Second, they are not able to preserve recall, which is relatively high for a large number of proposals but low for a small number of proposals.

Regarding autonomous driving, in which KITTI [1] represents the state-of-the-art benchmark, none of the existing algorithms work well until recently [19]. The existing algorithms not only require a very large number of proposals in order to achieve reasonable recall, but the recall also drops dramatically. Apart from the aforementioned limitations, the lack of success of existing algorithms can also be explained by the fact that KITTI images are more challenging, since they contain small objects, occluded objects, reflections, and shadows [1,19].

Chen et al. [19] first tackle object proposals for autonomous driving by introducing the class-specific 3DOP method, which is able to obtain high recall across various IoU thresholds. Notably, the generated proposals are also well localized since the recall drops gradually and slowly. While 3DOP is the best object proposal generator to date, its disadvantages are twofold. First, it must be run seperately with regard to each object class in order to obtain high recall, which is not efficient. This class-specific property leads to an increase in the processing time of the classification stage, in which the number of needed proposals is linearly increased with respect to the number of object classes. For example, Chen et al. [19] used a total of 6000 proposals for three object classes, while our class-independent approach achieves slightly better accuracy with only 2000 proposals. In fact, generating object proposals is usually referred to as a class-independent task [12,14,17], which measures the likelihood of a window containing an object without considering its class. Second, we observe that even though the results of 3DOP with around 10,000 proposals achieve very high recall, its top-ranked proposals with fewer candidates do not preserve recall effectively. This is because the Markov Random Field (MRF) ranking model of 3DOP is not very robust and because 3DOP only uses depth features, while visual RGB features are not exploited.

In this paper, we present a robust object proposals re-ranking algorithm for autonomous driving using convolutional neural networks (CNN) [20]. Considering the robustness of 3DOP in generating well-localized proposals and the success of deep learning in the last few years, we aim to present a learning algorithm that is able to overcome the aforementioned limitations of 3DOP. The goal is to introduce a class-independent algorithm that is able to achieve high recall and good localization with few candidates. Specifically, we propose a lightweight two-stream CNN that exploits both RGB features and depth features to re-rank proposals, which are generated from a customized class-independent 3DOP. Here, the depth features include disparity map and distance to the ground, which can be computed from a stereo image pair. We call our algorithm DeepStereoOP. Fig. 1 shows the block diagram of the proposed approach. The experiments show the effectiveness of DeepStereoOP, achieving the highest recall across all IoU thresholds and occlusion levels. Ultimately, the combination of DeepStereoOP and the state-of-the-art Fast R-CNN object detector [11,19] achieves one of the best detection results of all three KITTI object classes.

The remainder of this paper is structured as follows. Section 2 presents related work including existing object proposal algorithms, recent improvements, and top-performing object detectors. Section 3 presents the proposed approach, while Section 3.1 presents the experimental results with KITTI dataset to compare our approach to those in the literature. Section 3.2 concludes the paper.

2. Related work

Object detection has undergone a fundamental shift from the traditional sliding window paradigm to the object proposals approach. Therefore, instead of searching for objects at every image location and scale, the classifier just focuses on a set of candidate bounding boxes generated from an object proposals algorithm. Notable pioneer works that shape this research include Objectness [12], CPMC [13], Endres [14], and SelectiveSearch [15]. Subsequent to such pioneering work, many novel algorithms ranging from objectness-based scoring to similarity-based grouping as well as supervised learning and other improvements have also been introduced.

Regarding scoring approaches, Alexe et al. [12,21] first proposed Objectness by sampling an initial set of proposals according to a saliency map, and ranking them using several objectness cues such as saliency, contrast, edge density, and superpixel straddling. Rahtu et al. [22] and Zhang et al. [23] then extended Alexe et al.'s work using structured output ranking and cascaded ranking SVMs, respectively. Cheng et al. [24] introduced BING, which uses a simple linear classifier with the learned normed gradient features to rank proposals generated from sliding windows. Although BING is very fast, its localization is cursory due to the weak discrimination of simple gradient features [25]. Zhang et al. [26] then introduced BING++ to overcome this weakness. Recently, Zitnick et al. [18] proposed EdgeBoxes, which rapidly scores millions of windows by measuring the relationship between the contours completely enclosed within the box and those overlapping the box's boundary. Similar to BING, EdgeBoxes suffers from localization bias [17]. Lu et al. [27] proposed ContourBox to reject proposals that do not have explicit closed contours. Kuo et al. [28] introduced DeepBox, which re-ranks proposals generated from EdgeBoxes using CNN, so that it can achieve better recall with fewer proposals. However, the localization bias issue remains. In summary, the main limitation of these scoring approaches is their strong localization bias; while they achieve high recall at a low IoU threshold, the high recall is barely maintained as the IoU threshold increases. Chen et al. [29] proposed a refinement approach to reduce this bias using multi-thresholding straddling expansion, which erodes and dilates bounding boxes based on superpixels tightness. However, the localization bias issue was not completely resolved.

Compared to scoring methods, grouping methods typically obtain

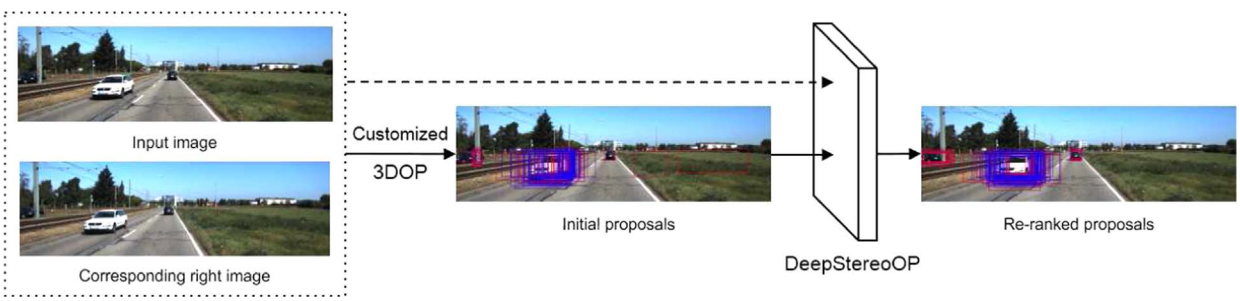

Fig. 1. Block diagram of the proposed object proposals re-ranking approach. Initial proposals with around 10,000 candidates are first generated using a customized class-independent 3DOP, and are then re-ranked using our proposed DeepStereoOP network.