Pattern Recognition Letters 000 (2016) 1–6

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

Learning arbitrary-shape object detector from bounding-box annotation by searching region-graph☆

Liantao Wang[a], Jianfeng Lu[a,*], Xiangyu Li[a], Zhan Huan[b], Jiuzhen Liang[b], Shuyue Chen[b]

[a] School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China
[b] School of Information Science and Engineering, Changzhou University, Changzhou 213164, China

ARTICLE INFO

Article history:
Received 22 December 2015
Available online xxx

Keywords:
Object localization
Arbitrary-shape
Region annotation
Region-graph

ABSTRACT

Arbitrary-shape is argued more precise than bounding-box for object detection. However, an arbitrary-shape detector usually requires pixel-level human annotation, which is very expensive and hardly afforded for any real-world application. On the other hand, bounding-box is much easier than pixel-wise segmentation in human labeling. In this paper we aim to realize the arbitrary-shape detection from bounding-box human annotation. To this end, we propose *location positiveness*, which encodes the information of bounding-box annotation to help obtain region annotation. In addition, we propose two graph-based methods to embed the *location positiveness*, which enable more accurate model trained from simpler annotation. Experimental results validate the performance of our method.

1. Introduction

Object detection can be solved by searching for the best-scoring sub-image, when a proper metric to measure the degree of a region belonging to an object is designed. The most popular one of such metrics is based on classification score, where a reliable classifier is trained with the annotated training data, and the sub-image that has the maximum response to the classifier in a new image is considered as the object location. Armed with this idea, any classifier can be applied to object detection as long as the best-scoring region is found, and a naive strategy for looking for the region is sliding window. However, this exhaustive strategy is too slow to be used in real-world applications, instead, approximate solution is usually sought by searching some promising areas according to certain prior knowledge. By exploiting the linearity of the scalar product, Lampert et al. [11] rewrite the decision function of a SVM classifier as a sum over per-feature contributions, which enables the usage of branch-and-bound strategy, and consequently speeds up the search process significantly.

Although the branch-and-bound strategy is efficient, the detection form is still limited to bounding-box, which is imprecise for non-boxy objects. Therefore, some variants such as composite boxes [27] and twisted windows [8] are proposed to improve the precision of detection. Fundamentally, Vijayanarasimhan and Grauman [22] propose efficient region search (ERS) to realize arbitrary-shape object detection. With properly annotated images, an additive classifier (e.g. linear SVM with histogram feature representation) is trained. Given a new image, it is first over-segmented into regions. By considering each region as a vertex and exploring the adjacency of the regions, a vertex-weighted region-graph is constructed, where the weights are the response of the regions to the classifier. It has been shown that the problem of obtaining the best-scoring contiguous set of regions in this graph is equivalent to the maximum-weight connected subgraph (MWCS) problem, and in turn is transformed into the prize-collecting Steiner tree (PCST) problem, which is efficiently solvable in practice with a branch-and-cut algorithm. The strategy is further applied to activity detection [3].

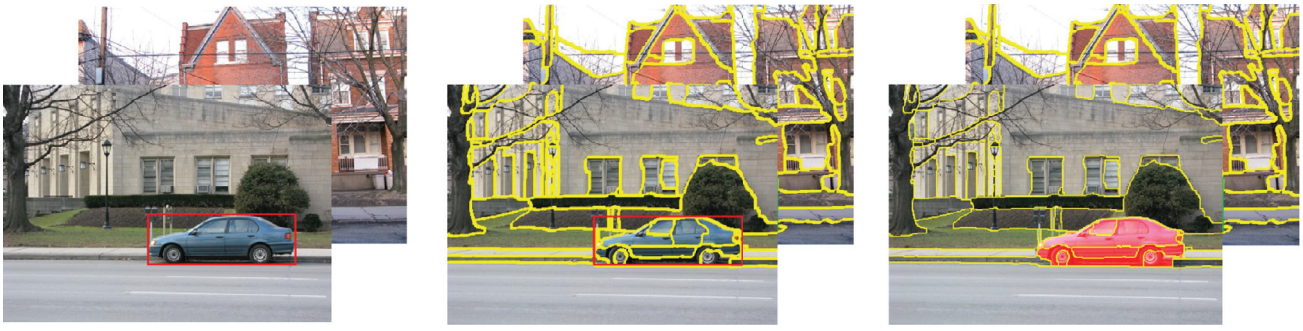
Although these object search strategies are efficient and some of them can realize arbitrary-shape detection, there are still some limitations.

(1) To obtain arbitrary-shape object detector, ERS [3,22] requires pixel-wise annotation as training data. Nevertheless, pixel-wise annotation is much more expensive than bounding-box.
(2) Many efficient object search methods including ERS proceed based on classification system, however, the region with the maximum classification score does not always correspond to a perfect object location prediction, which leads to imprecise localization.

Through unsupervised over-segmentation, images are transformed into regions, and regions can be considered as mid-level

☆ This paper has been recommended for acceptance by Cheng-Lin Liu.
* Corresponding author. Tel.: +86 25 84313997; fax: +86 2584315960.
E-mail address: lujf@njust.edu.cn (J. Lu).

http://dx.doi.org/10.1016/j.patrec.2016.06.022
0167-8655/



(a) Bounding-box annotation (b) Over-segmentation (c) Region annotation

Fig. 1. Illustration of our work. Given a set of images where objects are annotated with bounding-boxes in (a), we first over-segment the images into regions as shown in (b), and explore the region distribution aided by the bounding-boxes. Finally we employ graph-based methods to annotate the regions in (c).

components to constitute object and background (pixels can be considered as low-level components). Note that a small image segment has more information than a single pixel, and at the same time it is more likely to belong to a single object compared to a rectangular patch. Inspired by this idea, we perform object detection on region-graph, and try to overcome the above limitations.

If the region annotations are available, we can estimate the class-conditional probabilities using strategies such as histogram or kernel density estimation. Alternatively, we can estimate the ratio of class-conditional probabilities of a region using density ratio estimation methods [9,18]. We will show either of them corresponds to an implementation of object detection based on region-graph. So the key issue is whether we have region annotations. An intuitive solution may be pixel-wise segmentation, with which one can compare a region and determine its category. However, pixel-wise segmentation is much more expensive, and hardly affordable even for a database with medium size. Another much easier annotation is bounding-box, which trades off convenience against precision. In this paper, we will show how to obtain region annotation from ambiguous bounding-box annotation using the region-graph methods. Illustration of our work is shown in Fig. 1.

The contributions of this paper are as follows:

(1) We propose *location positiveness*, which encodes the information of bounding box annotation to help obtain region annotation.
(2) We propose two graph-based methods to embed the *location positiveness*, which enables more accurate model training from simpler annotation.
(3) We propose a method to compute region weight for ERS by introducing density ratio estimation strategy. Different from weighting methods based on classification score, the weight in our method is derived from the region distribution, and it is more suited for object localization.

The rest of the paper is organized as follows: we first review the works related to our method in Section 2. Section 3 describes the region-graph and *location positiveness*. Then we detail two graph-based methods for region annotation aided by bounding-box in Sections 4 and 5, respectively. The experimental results are shown in Section 6. Finally we conclude the paper with Section 7.

2. Related works

2.1. Graph-based methods

Graph-based methods have been a powerful tool in computer vision tasks, such as image segmentation [6,21], video segmentation [7,25], image retrieval [19,24], feature representation [12]. They have been employed for arbitrary-shape object detection too. Based on unsupervised over-segmentation, branch-and-cut is introduced to search the maximum-weight connected subgraph as arbitrary-shape object detection for images [22,30] and videos [3]. Wang et al. [23] exploit the localization property of a maximal entropy random walk on a weighted graph to rank the feature-points in an image to reveal the likelihood of their belonging to an object. More recently Yan et al. [26] take object detection as a multi-label superpixel labeling problem by graph-cut that minimizes an energy function.

2.2. Weighting region-graph

As most of efficient object search methods [11,27,28], ERS [22] decomposes the region score into feature-weights based on linear SVM classifier and histogram feature representation. Therefore, the weight in the region-graph is the sum of the feature-weights within the region. This kind of weighting is also used in other methods [3,29,30]. However, the weight is actually a classification score, and the final located region with the maximum classification score does not necessarily correspond to the perfect localization.

Yuan et al. [28] used mutual information to cope with intra-pattern variations. Based on naive Bayes assumption, the weight for a region (actually the mutual information between a region and a concept) is expressed as the sum of feature-scores, which is a function of two probability densities. Then the kernel density estimation is employed to estimate two densities separately. Though the weight computed from feature-point distribution performs well for object localization, kernel density estimation is not reliable in high-dimensional problems since division by an estimated quantity can magnify the estimation error [18]. We instead use a least-squares framework [9,10] to directly estimate the density ratio to weight the region-graph.

3. Region-graph and location positiveness

3.1. Region-graph

For an image $\mathcal{I}$, we first over-segment it into n mutually exclusive regions. The image is then represented by a collection of regions $\mathcal{I} = \{v_1, v_2, \ldots, v_n\}$. Any region v_i is described by a feature vector $\mathrm{x}_i \in \mathbb{R}^d$, and is associated with a label $y_i \in \{0, 1\}$, with 0 for non-object and 1 for object.

Considering each region as a vertex and certain neighborhood system, an image is represented by a region-graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$, where the vertex set $\mathcal{V} = \{v_1, v_2, \ldots, v_n\}$, and the edge set $\mathcal{E} \in \{v_1, v_2, \ldots, v_n\}^2$ is determined by a certain adjacency metric. The

Please cite this article as: L. Wang et al., Learning arbitrary-shape object detector from bounding-box annotation by searching region-graph, Pattern Recognition Letters (2016), http://dx.doi.org/10.1016/j.patrec.2016.06.022