Neurocomputing 212 (2016) 58–64

Contents lists available at ScienceDirect

Neurocomputing

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

Object localization with mid-level part detectors

Xiaoqin Kuang, Nong Sang*, Changxin Gao

Science and Technology on Multi-spectral Information Processing Laboratory, School of Automation, Huazhong University of Science and Technology, Wuhan, China

A R T I C L E I N F O

Article history:
Received 16 November 2015
Received in revised form
19 January 2016
Accepted 15 February 2016
Available online 28 June 2016

Keywords:
Object detection
Part models
Cluster
LDA
Feature selection

A B S T R A C T

We present a method for object detection combining the effectiveness of a set of mid-level parts. These parts are learned weak-supervised from object bounding box annotations. The approach based part models can handle the detection of objects across changes in viewpoint, intraclass variability and object deformation. The objects are localized by the detected parts with learned information of location and scale. We evaluate the detection method on the standard PASCAL VOC 2007 dataset. Our system is competitive with the state of art in localizing the object.

1. Introduction

Object localization and detection has been a popular topic in computer vision and pattern recognition for many years. The classifier and feature are the most important focuses. Various classifiers such as SVMs [1] and boosting [2,3] have been commonly used in detection. In addition, various types of image features have been considered such as SIFT [4], HOG [5], and CNNs [6] which is popular in recent times. It is still a challenging problem because the variation of intra-class, illumination, viewpoint and occlusion exists in real-world images. Many methods have been proposed to represent in the last few decades.

Dalal and Triggs [5] train holistic classifier for human detection using the fixed HOG template of entire person. The monolithic model cannot hold the variation within category. Later, multiple models have been applied. Each of the Exemplar-SVMs [7] is trained by a single positive instance and millions of negatives using HOG detector for the global object which would not be expected generalize enough. What is more, many samples are highly similar so it is not efficient to train detector with every exemplar. Method with multi-component models [8] picks seed object and aligns the rest objects to the seed as the component and trains individual model for each one, and then learns a second classifier that operates at the category level by aggregating responses from multiple components. But it needs keypoint and mask annotation when aligning.

In the nature world, some objects are highly articulated such as humans, cats and so on, and some other objects that are made artificially have rigid structure such as aeroplanes, motorbikes, cars, etc. However, the common feature of them is that there exist shared parts among the objects. It is natural to recognize humans though we have only seen the head, shoulder or torso. Additionally, we know that wheels are necessary for cars and aeroplanes have aerofoils. Thus part-based model would help to improve the detection in more general situations, as well as the significant partial occlusion.

Torralba et al. [9] represent the sharing features with boosting for detecting multiclass and multiview object. The method selects generic edge-like features shared across the classes which could be considered as low-level templates and not discriminative for specific object. The Implicit Shape Model (ISM) [10] learns a codebook of local appearance and contains information of where the local structures may appear on objects of the target category. The patches constituting the codebook are obtained by firstly applying interest point detector to get a set of image regions. In the Hough Forest detection [11], the patches are sampled densely from positive and negative training images. It constructs a random forest and each node stores the statistics of class and spatial information. Every leaf node plays the role of codebook to cast probability vote for position for test images. The more densely pathes are sampled, the more accurate the detection is. But among the sampled patches, many are slightly different or not distinctive enough to vote efficiently. The popular method of Deformable Parts Model (DPM) [12] proposes a detection system of mixtures of multi-scale deformable part models allowing for small deformation and multiple postures. While the numbers of components and parts are predefined to a fix number not referring to

* Corresponding author.
E-mail address: nsang@hust.edu.cn (N. Sang).

http://dx.doi.org/10.1016/j.neucom.2016.02.080
0925-2312/© 2016 Elsevier B.V. All rights reserved.

the data. The parts are initialized to cover high-energy regions of each root filter. The paper [13] consists of a canonical appearance template together with a dictionary of deformations as flow fields instead of deformation cost for each possible placement in DPM. The number of candidate deformation is much fewer than that for a DPM.

However, the patches captured in these methods suffer from the restriction that the amount of semantic information is strictly small, which often produces insufficient discriminative representation. Therefore, some recent works focus on characterizing objects based on mid-level semantic concepts. Many progress has been made in this area.

The “poselet” is a new notion of part given in [14]. They describe some particular parts of the human pose under some given viewpoint, which are tightly clustered in both 3D joint configuration space and 2D image appearance using annotation of 3D pose information. Ysuf Aytar and Andrew Zisserman afford an immediate detection method [15] in which they propose a new image representation based on mid-level discriminative patches. They use a sparse weighted combination of classifier patches to approximate the query HOG template by another HOG template. And the images' reranking is based on the spatial consistency of the maximum response of each patch. In [16], the authors obtain the part candidate proposal by sampling a random positive example, scale, aspect ratio, and location within the object bounding box. The parts are measured with the *Average Max Precision* (AMP) so as to select a subset of ones that are discriminative and complementary. They use the “Bag of Parts” model scored over the regions of object proposals. Gkioxari et al. [17] combine the poselets and DPMs to train part model for person, and select a fixed number of k-poselet detectors according to the AMP. The authors in [18] propose a method that selects a small subset containing non-redundant discriminative ones from the large pool of part filters automatically.

The conference paper of [19] detects object via classifying the candidate regions with features of maxpooling by part models, while the results depend on high recall from the coarse detection by multi-components. Additional experiment shows that it is better to find rigid object.

In this paper, we aim to automatically discover mid-level parts and select a subset of discriminant and non-redundant ones. The bottom-up localization takes advantage of the effective part detectors and affords an approach for general objects. What is more, the selecting process is simple and practical.

The remainder of the paper is structured as follows. In Section 2, we describe details of discovering semantic part detectors, and Section 3 is the procedure of how to locate object with the learned part. The performance of our method on detection datasets is evaluated in Section 4 and concludes of this paper are given in Section 5.

2. Discovering parts

In this section, we describe the procedure of discovering the mid-level parts.

2.1. Part candidates

Given a set of training data, our goal is to learn a compact collection of part detectors for the category. We sample densely to get all possible sub-windows and collect hundreds of thousands of parts. Patches are represented by augmented HOG descriptors [12] with the same dimensions.

Clustering is applied to generate “seed” for training the initial detectors. We run k-means to cluster patches using whitened HOG (WHO) features. The WHO features are considered to be better for clustering and classification, since the whitening removes the correlations and leaves the discriminative gradients [20]. We compute the covariance matrix Σ and mean feature μ_0 on HOG features with all background samples in advance. Then we cluster the positive objects using whitened features which are transformed from the HOG feature x to $\hat{x} = \Sigma^{-1/2}(x - \mu_0)$. As mentioned in [21], the number of clusters is set quite high since they do not trust k-means to generalize well, so we set k quite high ($k = S/5$, where S is the number of patches sampled for clustering) similar to them to ensure the consistency of each cluster. Clusters with less than 4 patches are regarded as poor seeds and removed.

Naturally, clusters are probably rough due to this unsupervised clustering. So the training scheme is followed which aims at collecting patches purer and more consistent.

We apply LDA to train detector for each cluster since it has the similar performance with SVM but accelerates the computation [20]. In order to improve the consistency, we use the cross-validation to refine detectors. The LDA model is a linear classifier over feature x with weights given by $w = \Sigma^{-1}(x_{mean} - \mu_0)$, and x_{mean} is the mean feature of the patch cluster.

We divide the training set D into two equal, non-overlapping subsets (D_1, D_2) as train-set and validation-set. Given the initial clusters obtained from train-set D_1, we train a discriminative classifier for each of them, using patches in images of other category as negative samples. We run the trained detectors in validation-set D_2 to discover the corresponding patches, and then new clusters are formed by the detections scored top m. We set $m=6$ to keep the purity of each cluster. After this training and detection, we switch the role of D_1 and D_2 and repeat the process until convergence. During the iteration, the detections with small number of patches are eliminated since they occur rarely to characterize appearance of the object.

2.2. Part selecting

The training procedure has produced a candidate set of part detectors as $D^c = \{d_i\}$. We know that there exists redundancy in this candidate set, thus the next task is to select a subset of discriminative detectors from them as $D^s \subset D^c$.

For an image patch I in the dataset, it will be quantized to feature vector f by the candidate detectors D^c as

$$f^c = [\varphi(I; d_1), \ldots, \varphi(I; d_i), \ldots, \varphi(I; d_{|D^c|})], \quad d_i \in D^c$$

When $\varphi(\cdot)$ is one dimension as $\varphi(\cdot) \in R$, the feature vector for patch I is

$$f^c = [\varphi_1, \ldots, \varphi_i, \ldots, \varphi_{|D^c|})]$$

If the patch is quantized with the selected part detectors, the feature vector is

$$f^s = [\varphi(I; d_1), \ldots, \varphi(I; d_i), \ldots, \varphi(I; d_{|D^s|})], \quad d_i \in D^s$$

We use the max-pooling technique to represent the patch to f, thus each dimension is the maximum firing score by one part detector at all scales and locations.

As $D^s \subset D^c$, then $f^s \subset f^c$. The task of part selection is equal to the feature selection. The discriminativeness of part detectors can be measured by their ability of classifying the object and background.

The training dataset T will be represented as feature vector set $F^c = \{f_i^c\}_{i=1}^{|T|}$. Let $\Phi^c(f^c)$ be the discriminant classifier trained on F^c, then the classified ability of part detector d_i is defined by

$$e(d_i) = \frac{\partial \Phi^c(f^c)}{\partial \varphi_i} \tag{1}$$

$e(d_i)$ describes the contribution of detector d_i for classification. The lager it is, the greater effect it has. When we take the linear