Pattern Recognition 69 (2017) 26–38

Contents lists available at ScienceDirect

Pattern Recognition

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

Learning discriminated and correlated patches for multi-view object detection using sparse coding

Zehuan Yuan[a], Tong Lu[a,*], Chew Lim Tan[b]

[a] National Key Laboratory for Novel Software Technology, Nanjing University, China
[b] School of Computing, National University of Singapore, Singapore

ARTICLE INFO

Article history:
Received 20 July 2015
Revised 11 February 2017
Accepted 25 March 2017
Available online 30 March 2017

Keywords:
Multi-view object detection
Hough decision tree
Transfer vote
Sparse coding

ABSTRACT

Multi-view object detection is an open and challenging problem due to its inherent intra-class variability among discrete viewpoints. This paper aims to perform multi-view object detection by learning discriminated and correlated patches firstly and then making inference based on them. In the training stage, discriminative patches are discovered for each view by a Hough decision tree corresponding to leaf nodes with high distinctiveness and stable spatial distributions in the tree. Then discriminated patches across different views are linked to establish the correlations between any two neighboring views. During multi-view detection, intra-view direct votes and inter-view transfer votes are integrated to obtain voted Hough images through a probabilistic approach with each view having one Hough image, and Mean-Shift estimation is finally employed to detect object instances and infer image viewpoint. The experiments performed on two benchmark multi-view 3D Object Category datasets and PASCAL VOC'06 Car dataset illustrate the effectiveness of the proposed framework.

1. Introduction

Multi-view object detection is always receiving an increasing attention in the computer vision and pattern recognition community [1–4]. It focuses on how to accurately analyze viewpoint variations of object instances of the same category in natural scene images downloaded from the Internet or captured by cameras and how to utilize viewpoint to facilitate object detection. By intuition, the goal of multi-view object detection is to combine view prediction and object detection in the same model, and theoretically the exploration on multi-view analysis is an indispensable step for improving the accuracies of the existing object detection methods due to the fact that the explicit viewpoint prediction helps eliminate the intra-class visual variations caused by varied views. Therefore, multi-view object detection has become an important issue especially for developing various real-life systems such as pedestrian detection [5], automatic driving [6], intelligent robots [7] and industrial automation [8], where people always hope to find consistent object information even though there exist viewpoint changes.

In the past years, patch-based models (e.g., poselets [9] and meaningful parts [10,11]) have been widely used for object detection, and flexible shape models (e.g., the typical star model in DPM [12]) have also been adopted to combine the discovered parts. These object detection strategies have been proved useful to solve the variability problem of object appearance in some sense. However, they are not flexible enough for fine-grained multi-view object detection since most of them only construct a unified discriminative patch set (part filters) regardless of viewpoint variations. This is probably because discovering and relating different discriminative patches in all views under the full perspective projection are always computational costly and often too much complicated. As a result, sometimes they face difficulties in estimating an unknown viewpoint of an object instance, which in turn makes it hard to characterize appearance variabilities associated with viewpoint changes for accurate object detection.

In this paper, we propose a novel framework for multi-view object detection. It is inspired by the following two observations. Our first observation is that there often exist view-specific discriminative patches for object appearances in different views. View-specific discriminative patches here refer to those image patches that can not only distinguish the object from the background but also are capable of distinguishing different object views. For illustration, Fig. 1 shows several object instances from different views, where we can see *headlamps* are discriminative for identifying View-1 as the front view, while *taillights* are helpful in identifying View-2 as the back view. It means that the detection of *headlamp* patches from an input image will be useful in identifying the car as the front view. Likewise, the detection of *taillights*

* Corresponding author.
E-mail address: lutong@nju.edu.cn (T. Lu).

http://dx.doi.org/10.1016/j.patcog.2017.03.033
0031-3203/

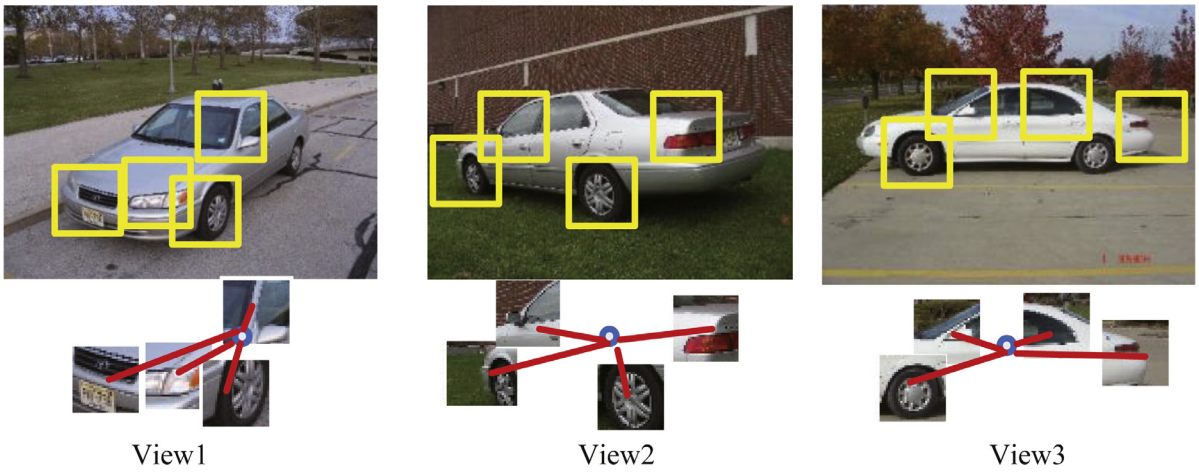

Fig. 1. Different view examples. Discriminative patches in each view are marked by yellow rectangles. The second line shows the corresponding spatial configurations of these extracted discriminative patches relative to object centers (blue circles). We can observe that each view essentially has its own view-specific discriminative patches and conventional spatial configurations. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

patches probably indicates the input image is more likely from the back view. Our second observation is that the spatial configuration, namely, the specific spatial distribution of a discriminative patch, often helps identify its corresponding view. For example, it can be found from Fig. 1 that the spatial distributions of *wheel* patches from all the three views are visually different, which in turn indicate their respective views. In this sense, the learning of view-discriminative patches and their view-specific spatial distribution will facilitate view estimation. On the other hand, object detection in each view can also benefit from these discovered patches as they are discriminative and the spatial configuration is relatively stable in a single view.

Specifically, we first learn view-specific discriminative visual patches both in appearance and the spatial distribution for each view independently based on sparse coding representation and Hough-tree learning. That is, a Hough decision tree is constructed for each view from a group of sampled positive and negative training patches. Unlike the traditional decision trees, for every internal node, the best split criteria is selected to divide the node into two child nodes with both higher class certainty and more stable spatial offsets to object centers. This implies that the patches in child nodes are purer and more discriminative than those in their parents because patches with consistent spatial locations come from the same part usually. Thus when we reach a pure leaf node, the patches in it correspond to the most discriminative patches of that view, which can not only be used to separate an object from background, but also have stable spatial offsets to indicate object centers.

Next, we associate Hough trees of different views at the leaf level as a linked multi-view Hough forest by means of matching similar patches to each other. In multi-view object detection, a patch in one view often has duplicates with similar appearances and locations in images of neighboring views. Therefore, object detection from one view can not only be inferred by its view-specific visual discriminative patches but also can benefit from the occurrences of correlated patches in other views, particularly when there exist many noisy patches in the pick-out discriminative patches for the given view. We thus use patch correspondences to correlate leaf nodes in the Hough trees of neighboring views. Actually, correlating Hough trees is necessary and important for multi-view detection. In most scenarios, training images are unbalanced over different views and thus there might be few positive samples for some rare views. However, correlated neighboring views can assist object detection in these views to alleviate overfitting in establishing its corresponding Hough tree.

Finally, after constructing a linked Hough forest with each leaf node containing view-characteristic discriminative and cross-view correlated patches, we use a variant of Hough vote to perform both object detection and view prediction simultaneously. Specifically, for any test image with an unknown viewpoint, we adopt a probabilistic mechanism composed of *intra-view voting* and *transfer inter-view voting* to infer the existence and location of an object instance based on these learned discriminated and correlated patches. That is, following each Hough tree a test image patch can be finally matched to a particular leaf node, which will contribute Hough votes (intra-view and inter-view votes) on the likelihood and potential locations of the specific object instance under different views. The overall probabilities both from intra-view votes and transfer inter-view votes are accordingly accumulated over all test image patches in the Hough space for each view. Object detection and view prediction will be obtained by finding the maximal response using Mean-Shift estimation.

The two main contributions of our method are as follows. First, we learn view-characteristic discriminated patches for each view using Hough decision trees and sparse coding representation, where visual appearances and spatial configurations of object instances of different views are considered jointly. Second, to establish the relations across different views, cross-view discriminative patches are related to assist object detection and view estimation. The experiments performed on two benchmark multi-view 3D Object Category datasets and PASCAL VOC'06 car dataset illustrate the effectiveness of the proposed framework.

The rest of the paper is organized as follows. Section 2 introduces the related research on multi-view object detection. In Section 3, we introduce how to discover view-specific discriminative patches and further correlate them based on sparse coding representation and Hough-tree learning. Section 4 details the process of multi-view object detection using these patches, and Section 5 presents our experimental results. Finally, Section 6 concludes the proposed method.

2. Related work

Depending on whether an explicit 3D object model is generated or used, the existing research on multi-view object detection can be roughly categorized into two classes: *two-dimensional methods* and *three-dimensional methods*.

For *two-dimensional object class detection*, an intuitive and straight-forward way is to build view-aware feature clusters which are then treated as object parts and combined spatially to detect