Pattern Recognition 60 (2016) 227–238

Contents lists available at ScienceDirect

Pattern Recognition

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

Pedestrian detection aided by fusion of binocular information

Zhiguo Zhang [a], Wenbing Tao [a,*], Kun Sun [a], Wenbin Hu [b], Li Yao [c]

[a] National Key Laboratory of Science and Technology on Multi-spectral Information Processing, School of Automation, Huazhong University of Science and Technology, Wuhan 430074, China
[b] School of Computer, Wuhan University, Wuhan 430072, China
[c] School of Computer Science and Engineering, Southeast University, Nanjing 210096, China

ARTICLE INFO

Article history:
Received 27 January 2015
Received in revised form
5 April 2016
Accepted 14 May 2016
Available online 24 May 2016

Keywords:
Pedestrian detection
Binocular vision
Matching
Fusion
Double-pedestrian detection

ABSTRACT

In this paper, a pedestrian detection framework aided by the fusion of information between binocular vision is proposed. In this framework, we follow the intuition that a pedestrian has consistent appearance when observed from different viewpoints. A baseline detector is used on both the left and right images with a conservative threshold in order to preserve a larger candidate set. Then adjacency constrained search based on the disparity map is applied to find the optimal matching pairs between the left and right candidate sets. After that, a mixture model of two-pedestrian detector is designed to capture the unique visual cues which are formed by two nearby pedestrians but cannot be captured by single-pedestrian detectors. Finally, an information fusion module is established to model the relationship between the single- and double-pedestrian detectors as well as to refine the final detection decisions. Compared with single image pedestrian detection, our detection framework has the potential of aggregating information from multiple images to improve the detection on individual image. Thirteen state-of-the-art pedestrian representations are investigated on the widely used ETH dataset. Experimental results show that our framework improves all these approaches.

1. Introduction

Pedestrian detection is one of the most important topics in object detection and has been widely applied in many fields such as intelligent video surveillance, autonomous robotic navigation, and automotive safety. Meanwhile, it also confronts many challenges, such as variable appearance and shape, highly cluttered backgrounds, illumination, occlusions, viewpoint changes and so on. In recent years, many available approaches for pedestrian detection have been proposed. The present pedestrian detection methods can be roughly divided into two types: monocular-based and binocular-based.

The conventional monocular-based pedestrian detection approaches usually train pedestrian detectors from a set of manually labeled training images and then put the detectors into use on individual test images. During the process, various distinctive features can be extracted as the input of the classifier from the training images.

Viola and Jones (VJ [1]) applied multiscale Haar wavelets selected by a variant of AdaBoost [2] to describe objects and introduced integral image for fast feature computation. It is one of the notable landmarks in the process of object detection. Gradient-based feature is the most widely used feature to discriminate the pedestrian from the background. The histogram of gradients feature (HOG) combined with a linear SVM was proposed by Dalal and Triggs (HOG [3]). It makes a breakthrough of pedestrian detection and forms a prominent type. Many other methods either extend the features to combine with other information like color or texture features or replace the classifier with other substitutes. Wang et al. (HOGLBP [4]) integrated HOG features with cell-structured Local Binary Patterns (LBP [5]) to handle occlusion. Dollár et al. (ChnFtrs [6]) computed Haar-like features over grayscale channel, gradient magnitude channel, LUV color channels, and gradient magnitude quantized by orientation and proposed a uniform framework for integrating multiple feature types. And then, they (FPDW [7]) sped up the ChnFtrs [6] by approximating features from nearby scales. Wojek and Schiele (MultiFtr [8]) combined the shapelet features [9], Haar-like features [1] and HOG features [3] and obtained good performance. Walk et al. extended MultiFtr [8] by combining color self-similarity features (MultiFtr+CSS [10]) and motion features (MultiFtr+Motion [10]). Considerable effort has also been devoted to improving the learning framework. The histogram intersection kernel for SVMs (HikSVM) was proposed by Maji et al. (HikSVM [11]), which made it possible to use the nonlinear SVM in sliding window-based detection and sped up the computation. Other features, like shape

* Corresponding author.
E-mail address: wenbingtao@hust.edu.cn (W. Tao).

http://dx.doi.org/10.1016/j.patcog.2016.05.006
0031-3203/

and motion features, are also important cues for pedestrian detection. Gavrila and Philomin [12] matched image edges to a set of shape templates with the Hausdorff distance transform rapidly. Sabzmeydani and Mori (Shapelet [9]) used multiple local gradient-based shape descriptors to classify different objects. Lin and Davis (PoseInv [13]) described a pedestrian's shape with different body parts (head, upper body, and legs) and used HOG features along the shape's outline to express the body. For motion features, Dalal and Triggs [14] integrated optical flow with the HOG features which modeled the motion based on an optical flow field's internal differences. In recent years, part-based methods (LatSvm [15,16]) have achieved great performance for pedestrian detection. They achieve classification by combining the results of the root and different part filters. Many of their variants provided leading performance in recent PASCAL VOC (Pattern Analysis, Statical Modeling and Computational Learning, Visual Object Classes) competitions.

On the other hand, it is well known that the binocular images can provide more information such as disparity and depth than a single image. This information is unique to binocular images. The binocular information is commonly used in separate modules of pedestrian detection as a supplementary information [17] since the stereo information is a reliable and helpful cue in general situation. The idea of using binocular information to help pedestrian detection has been proposed for a long time. In [18], a binocular vision based pedestrian detection system was proposed. 3D-point map is extracted based on Canny features and filtered according to a neighborhood criterion. Candidates are located using a subtractive clustering attention mechanism and Support Vector Machines (SVM) classifier is used to classify each candidate. Then Kalman filter is used to enhance the performance of the pedestrian detection system. In [19], a new method of ROIs (Regions of Interest) extraction using two-stage segmentation was proposed. The pedestrian detector is used to detect hypotheses with a limited sliding windows process in the ROIs instead of taking ROIs as inputs directly. Mammeri et al. [20] proposed a keypoint-based binocular distance measurement for the pedestrian detection system. The HOG-SVM pedestrian detection method is used to detect pedestrians in both left and right images. A crossover re-detection mechanism is proposed to enhance the robustness of the detection method. Then, Speeded Up Robust Features (SURF) and Oriented FAST and Rotated BRIEF (ORB) are applied to extract keypoints and match them. Most of the binocular vision based methods focus on the current frame. Different from that, our method not only focus on the current frame but also the counterpart in the other sequence and the fusion of the two sequences.

Depth- or disparity-based ROIs generation is one of the major applications of stereo pairs. Zhao and Thorpe [21] obtained foreground regions by clustering in the disparity space. Labayrade et al. [22] proposed the "V-disparity" technique based on the disparity map which provides a suitable representation of the geometric content of the road scene. Inspired by the "V-disparity", the method of "stixel world" was proposed by [23] to obtain the regions of obstacles on the ground. The stixel world is a very simplified world model which assumes that the ground is locally flat and all the objects can be described as flat sticks raising vertically above the ground. Benenson et al. [24] improved the detection speed by transferring the computation from test time to training time and utilizing the stixel world extracted from stereo images. The speed can reach 135 fps when processing rectified binocular images with GPU acceleration. The methods based on binocular images are widely used in autonomous driving. Franke and Heinrich [25] merged the stereo images with interframe motion and optical flow [26]. They extracted depth information with time correlation and motion analysis in order to permit early detection of moving objects. A pedestrian detection framework using stereo information was also proposed in [27]. Our paper is different from [27] in three aspects: (1) the main focus is pedestrian detection on a sequence or video in [27] while our paper only requires two rectified stereo images. (2) the approach in [27] only uses a baseline detector for detecting while this paper uses 1-pedestrian detector and 2-pedestrian detector to obtain hypotheses. (3) the approach in [27] refines the confidence of the detection in two steps following three basic rules while this paper incorporates the information with a weighting function.

The consistency of appearance between a pair of binocular images is another useful characteristic. It is often employed as a supplementary information for pedestrian detection. We follow the key intuition that the same pedestrian has consistent appearance when observed from different viewpoints. This information can be used as a supplement for the final decisions. In addition, it is well known that significant partial occlusion is challenging for pedestrian detection and most of the current detectors fail to robustly detect the pedestrian. For example, the DPM method (Deformable Part Models [16]) will fail at about 20% [28] of occlusion. The higher percentage the occlusion is, the worse the performance will be. We noticed that when a pedestrian is occluded in one image, it is not always occluded in another image observed from a different viewpoint. If we build the matching relationship between the two instances and combine these two instances into one, the percentage of occlusion will be reduced. Sometimes, the pedestrians cannot be detected well in an image while they can be easily detected by other ways (e.g. detect two pedestrians in the meantime).

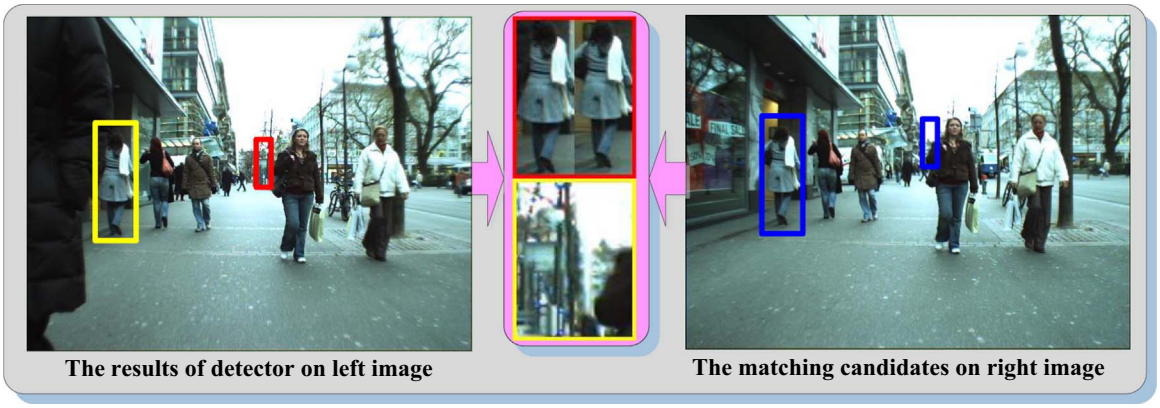

Fig. 1. Illustration of our idea with example. The yellow box in the left image cannot be well detected. The red box in the left image is a false alarm. We find the matching region on the right image and combine the two matching regions into one image. We use the information unique to 2-pedestrian image to refine the single pedestrian detection. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)