Image and Vision Computing 45 (2016) 11–21

Contents lists available at ScienceDirect

Image and Vision Computing

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

A novel low false alarm rate pedestrian detection framework based on single depth images☆

Xiaohui Zhao [a,b], Yicheng Jiang [a,*], Tania Stathaki [b,**]

[a] Research Institute of Electronic Engineering Technology, Harbin Institute of Technology, Mailbox 338, Harbin 150001, China
[b] Communications and Signal Processing Research Group, Department of Electrical and Electronic Engineering, Imperial College London, Exhibition Road, London SW7 2AZ, United Kingdom

ARTICLE INFO

Article history:
Received 20 October 2014
Received in revised form 9 October 2015
Accepted 6 November 2015
Available online 3 December 2015

Keywords:
Pedestrian detection
Histogram of Oriented Gradients
Shape context
Chamfer matching

ABSTRACT

Pedestrian detection is an important image understanding problem with many potential applications. There has been little success in creating an algorithm which exhibits a high detection rate while keeping the false alarm in a relatively low rate. This paper presents a method designed to resolve this problem. The proposed method uses the Kinect or any similar type of sensors which facilitate the extraction of a distinct foreground. Then potential regions, which are candidates for the presence of human(s), are detected by employing the widely used Histogram of Oriented Gradients (HOG) technique, which performs well in terms of good detection rates but suffers from significantly high false alarm rates. Our method applies a sequence of operations to eliminate the false alarms produced by the HOG detector based on investigating the fine details of local shape information. Local shape information can be identified by efficient utilization of the edge points which, in this work, are used to formulate the so called Shape Context (SC) model. The proposed detection framework is divided in four sequential stages, with each stage aiming at refining the detection results of the previous stage. In addition, our approach employs a pre-evaluation stage to pre-screen and restrict further detection results. Extensive experimental results on the dataset created by the authors, involves 673 images collected from 11 different scenes, demonstrate that the proposed method eliminates a large percentage of the false alarms produced by the HOG pedestrian detector.

1. Introduction

Pedestrian detection is one of the most popular areas in computer vision over the recent years [1–3]. A pedestrian is defined as a human standing in an almost upright position. There are two main classes of methods for pedestrian detection, namely, the generative methods and the discriminative methods. Both methods aim at distinguishing between human and non-human classes of objects [4]. The main concept behind generative methods is to find one or several models (e.g., shape models [5], texture models [6], and others) for classification, while the main concept of discriminative methods is to find the best descriptive features (e.g., Haar wavelet features [7], Histogram of Oriented Gradients (HOG) [8], Scale-invariant Feature Transform (SIFT) [9]) or find the optimal parameter values for their classifier (e.g., Neural Networks (NN) [5,10], Support Vector Machines (SVM) [11] and Adaboost [12]) [7]. The majority of techniques which were originally developed for pedestrian detection perform well only in scenarios where the entire human silhouette is present in the image. Recently, researchers have started to consider using techniques which investigate the characteristics of the image at a local fashion [11,13]. The main advantage of this type of approaches is that the object of interest may be recognized even in the case of occlusion by other objects or when part of it is not visible [14,15].

Inspired by these methods, we propose a pedestrian detection scheme which investigates both global and local shape information and take advantage of recent methodologies developed by experts in this field, to achieve a more accurate detection result. Specifically, to detect pedestrians in an image, firstly, possible candidate regions are estimated by the HOG detector. These regions are kept for subsequent investigation while the remaining regions where no humans were detected using HOG are discarded from the investigation. Taking into consideration the high false alarm rate of HOG detector, candidate regions are further investigated using alternative models. In other words, the proposed framework aims at eliminating false alarms produced by the HOG pedestrian detector while preserving the true detections. There are several challenges in pedestrian detection, e.g., posture variations, heavily cluttered background, occlusion and others. To decrease the influence of these problems, in the proposed work, multiple models rather than one single model are involved to improve evaluation effectiveness. Furthermore, the Kinect depth sensor

☆ This paper has been recommended for acceptance by Massimo Piccardi, Ph.D.

* Corresponding author. Tel.: +86 13936440899.
** Corresponding author. Tel.: +44 207 594 6229.
E-mail addresses: xh.zhao@outlook.com (X. Zhao), jiangyc@hit.edu.cn (Y. Jiang), t.stathaki@imperial.ac.uk (T. Stathaki).

http://dx.doi.org/10.1016/j.imavis.2015.11.002
0262-8856/

is also employed to facilitate the elimination of irrelevant signals which affect the detection results, such as background signals, weak edges and texture. We explore the proposed method on a dataset created by the authors that involved 673 patches collected from 11 different scenes with 470 human existing patches and 203 human-like patches.

We discuss the method in Section 2, describe the experiments and performance evaluation results in Section 3 and give conclusions in Section 4.

2. Methodology

2.1. Overview of the method

This section gives an overview of the proposed pedestrian detection scheme. To start with, in this work we use both RGB and Kinect sensors. RGB sensors provide detailed information which facilitates the extraction of refined edges. HOG detector requires a rich and detailed edge map in order to perform sufficiently well and for that reason, in this work the HOG detector is implemented on the outputs of the RGB sensors. However, in order to eliminate false detections produced by HOG, we examine local Shape Context (SC) models [16]. We believe that in order to focus on local features that correspond solely to the object of interest irrelevant signals which affect the detection results, such as background signals, weak edges and texture must be eliminated. Detailed edge maps should not provide a positive contribution in our attempt to deduce the false alarms produced by HOG. Therefore, SC models are implemented on the outputs of the depth (Kinect) sensors only. More specifically, we only examine regions with the depth sensor outputs which have been detected by HOG in the corresponding RGB outputs. We assume that RGB and depth sensors are perfectly registered.

The basic flowchart of the proposed detection scheme is shown in Fig. 1. As already mentioned above, input images are image patches taken from the depth sensor output, selected on the basis that the corresponding patches of the images taken from the RGB sensor have been detected as suspicious regions for the presence of a pedestrian. The first stage of the detection framework involves a sequence of pre-processing operations. Initially, the image that is obtained from the depth sensor is transformed into a black and white (binary) image. After that we obtain the edge map of the binarized depth sensor response by employing the widely used Canny edge detector. We observe that the edge maps extracted from binarized depth images often have a number of line artifacts toward the bottom of the image patches. As far as true detections are concerned, these line artifacts are due to the presence of the ground under the human silhouette and the fact that both ground and silhouette might have the same distance from the camera. Furthermore, there are weak edges and texture within the entire image which appear like blobs and other types of irrelevant micro-structures when we obtain the edge maps from the depth sensor, all of which should be erased in the pre-processing stage. The detailed pre-processing method is introduced in Section 2.2. After the pre-processing of the depth sensor output, the SC representation of each edge pixel is extracted, using the technique described in Section 2.3. In order to use SC we must compare the SC of an edge pixel located within a region of interest with the SC of an edge pixel located within a prototype region. In other words, in contrast with the HOG model which is a training-based model, the SC model is a prototype-based model. In this work we use a database of several representative prototype images of pedestrians at slightly varying poses. The aim is to compare the SC of edge pixels of an image region obtained from an image of interest with the SC of edge pixels of an image region obtained from a prototype image. Comparison of two SC models is realized using specific metrics that have been developed by researchers. Two edge points are "matched" if they possess "similar" shape context. The final stage is introduced in Section 2.4, where matched edge points are clustered into several clusters according to their spatial distribution, and then these clusters are evaluated according to their characteristics either kept or rejected in order to complete the detection scheme.

2.2. Summary of pre-processing and pre-evaluation stages

In this section we assume that the regions detected by HOG after this is applied on the RGB sensor outputs are available. We locate these regions in the corresponding depth images in order to extract the SC models. Before the SC extraction, the edge map extracted by the Canny detector is obtained. It is important to stress out the fact the edge map is not obtained directly from the original depth image but from a binarized version of it constructed in a fashion that will be described in the subsequent section. As already mentioned, it is observed that in the binarized depth image some line artifacts exist toward the bottom of the image. By the term "line artifact", we are referring to a horizontal line (row) with an excessive number of white pixels. There are two possible detected scenarios that may produce line artifacts. A possible scenario relates to true detections in the depth image where the depth values within human feet are almost the same as the depth values of the ground around human feet. In that case the line artifacts are concentrated toward the bottom of the detected patch. Another possible scenario relates to false detections in the depth image where it is observed that there are line artifacts evenly distributed across the falsely detected patch, and not necessarily present toward the bottom of the detected patch. Furthermore, it is noticed that some relatively small blob artifacts are present and randomly distributed across the binary detected patches. Both line artifacts and blob artifacts can significantly undermine the quality of the extracted edge map. A pre-processing method based on the structural properties of these artifacts is applied in order to eliminate them, which is described in detail in Section 2.2.1 below. Then, a pre-evaluation process is used to eliminate false detections. These detections can be identified by investigation of simple features, e.g., the number of white pixels in the binary image patch and the degree of similarity of the two edge maps, i.e., that of a detected patch in the complex image under investigation and that of a perfect model of a human body (prototype). The idea behind the pre-evaluation process is that some images without human that are occasionally found should be eliminated at the early stage of the detection scheme without further processing in order to save computation time. The details of the pre-evaluation process are introduced in Section 2.2.2.

2.2.1. Description of the pre-processing stage algorithm

Firstly, we aim at converting the detected depth image patches into a binary image patches. This is different from the standard binarized edge map where a threshold is set and pixels with edge response greater or smaller than this threshold are replaced with white or black respectively. Our binary image consists of two areas, namely, the object of interest (human-like object) and the background. It is obvious to assume that the depth values of pixels which are located within the object of interest show a small variation around their mean. Therefore, the object of

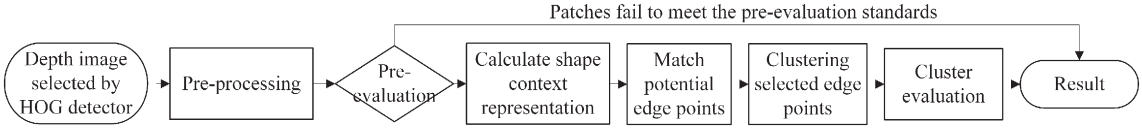

Fig. 1. The overview of the proposed detection scheme.