Neurocomputing 184 (2016) 66–77

Contents lists available at ScienceDirect

Neurocomputing

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

Learning discriminative shape statistics distribution features for pedestrian detection

Jifeng Shen [a,*], Xin Zuo [c], Wankou Yang [b], Hualong Yu [c], Guohai Liu [a]

[a] School of electronic and informatics engineering, Jiangsu University, Zhenjiang, Jiangsu 212013, China
[b] School of automation, Southeast University, Nanjing, Jiangsu 210096, China
[c] School of Computer Science and Engineering, Jiangsu University of Science, Zhenjiang, Jiangsu 212003, China

ARTICLE INFO

Article history:
Received 22 January 2015
Received in revised form
1 August 2015
Accepted 2 August 2015
Available online 2 December 2015

Keywords:
Multiple channel feature
Shape Statistic Distribution feature
Feature learning

ABSTRACT

Discriminative feature plays an important role in object detection system and traditional tactics heavily depends on the hand-designed feature. Recent study shows that feature can be learned from data, and this idea opens a new way to deal with many computer vision problems. In this paper, we propose a novel feature which learns discriminative information based on data distribution and shape statistics for pedestrian detection. It makes use of data distribution which is rich of discriminative information from positive and negative samples, and also utilizes shape statistics which comes from average human template. The proposed method exploits the distribution in multiple channel image spaces, and learns an optimal hyper plane to separate pedestrians from background patches in specific area with shape statistics. It maintains the merit of simplicity in computation and also obtains powerful discriminative ability. Two versions of Shape Statistic Distribution features are proposed which are derived from Informed Haar-like feature, but more discriminative than the original one. Experimental results based on INRIA, ETH and Caltech-USA datasets show that our proposed methods can achieve state-of-art performance. Furthermore the running speed of our detector can reach at 22 fps for 480×640 images.

1. Introduction

In recent years, real-time pedestrian detection has attracted lots of attention from both the researchers from academic field and technique engineers in the industry field. It is widely used in numerous fields such as video surveillance, sports athlete evaluation and autonomous-driven vehicles. Pedestrian detection in unlimited environment is a challenge task, for the sake of articulation, occlusion, illumination and view-changing, etc.

The notable landmark for pedestrian detection is due to Dalal & Trigg's work, who proposed the Histogram of Gradient (HOG) features [1] which is also a great success in generic object detection. The simple application of HOG feature trained with the SVM classifier can achieve a detection rate of 45% @ FPPI = 0.1. Although this method is very effective to detect nearly upright pedestrians, the running speed of detector is quite slow which takes more than 500 ms to scan a 320×240 image. This drawback baffles its application in many fields which demand real-time running speed. After that many researchers have proposed improved versions [2–7] based on the HOG features, but still cannot meet the need for the application in the industry. The first breakthrough is due to Piotr Dollar, who has designed a detector which is based on integral channel features [8] and trained with boosting algorithm [9]. This detector can achieve a real-time running speed of 20 fps for a 640×480 image. Moreover, other researchers make use of the GPU hardware and other cues [10,11] in video to further improve the running speed at a high speed of 100 fps [12]. Although it has gained a great success in this field, most of the features are hand designed which may not optimal for building an optimal detector. Recent studies [13,15] show that learning features based on deep learning for object detection is a novel idea which is very effective to solve many computer vision problems, but this method heavily depends on the high-end computers and also very slow to train. More recently, informed Haar-like feature [16] is proposed which generates a set of template with prior information of human parts instead of exhaustive sampling like traditional Haar-like features [17], and is very effective to calculate. In this paper, we will not try to cover all the methods published in the recent years, since there's a survey [18] which gives a detailed description of most of the state-of-arts methods. In this paper, we will only try to refer to those close to our proposed method.

In this paper, we propose a new feature family named Shape Statistics' Distribution (*SSD*) feature, which aims to learn

* Corresponding author.
E-mail address: shenjifeng1980@hotmail.com (J. Shen).

http://dx.doi.org/10.1016/j.neucom.2015.08.107
0925-2312/

Fig. 1. Different channels of the input image (first row corresponds to the original, magnitude, gradient angle (6), sobel edge image; Second row corresponds to the LBP, orientation, LUV, HSV and canny image).

discriminative information based on the shape statistics from average human template, and the distribution in the feature space which discriminates positive samples from negative ones. Our method is motivated by the usage of both integral channel feature [8] (*ICF*) and aggregated channel features [19] (*ACF*) which utilized multiple heterogeneous channels derived from the original image, such as color, texture and shape cues. The key contribution of our paper is to design a novel semi-handcraft feature for pedestrian detection, which not only utilize the multiple channel information, but also make use of the distribution from the training data and shape statistics. As most of the state-of-art methods focus on hand-craft feature or automatic learning feature from data (such as Deep Learning) for object detection. Our proposed method focus on learning semi-handcraft feature which is simple to implement, efficient to run, and does not depend on the high-end computer with high memory capacity and GPU and obtains state-of-art performance. Moreover, our method is much easier to implement, faster to train in ordinary computer and also can achieve real-time running speed.

The rest of the paper is organized as follows: Section 2 discusses the related work which our work based on; Section 3 and Section 4 introduces our proposed SSD features with different filter learning algorithm; Section 5 covers the experimental results and analysis and Section 6 concludes this paper.

2. Related work

2.1. Integral channel features

The ICF [11] models the feature C of image I as a channel generation function Ω, so each channel image of the original I can be represent as $C_i = \Omega_i(I)$, where $\Omega = \{\Omega_i\}, i = 1, 2, \ldots, n$, C_i is the ith channel for image I, Ω_i is the ith channel generation function. The channel generation function can be linear, such as the gray-level image of original color image I, or nonlinear transform, such as the gradient of image. Each channel represents a different feature space which derived from original image. Fig. 1 shows 18 different channel images can be generated from the original image, which reflects different information source. For example, the gradient image with different angles can reflect the orientation information in the image which is similar to the Gabor filter, the canny image can represent the edge

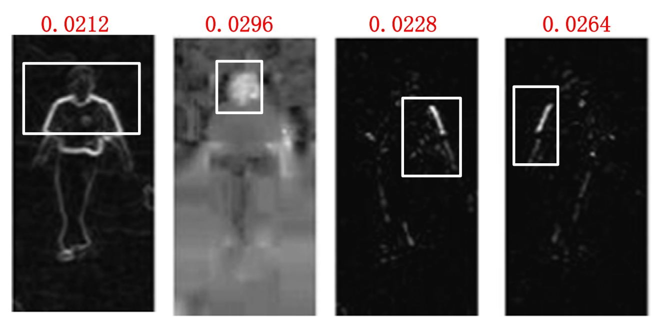

Fig. 2. Integral channel features (magnitude, chrome of LUV, third and fifth gradient images. The decimal represent the random weight of each feature).

information of human and different color space can reflect the color consistency in clothes and human parts, etc. The different channel is heterogeneous with each other, so extracting multiple channels is a mean to obtain different information from a single image.

After all the channels features are generated, randomly sampling a set of rectangles in different channel images in order to get informative feature such that they can differentiate positive samples from negative samples. Suppose $R_i(C)$ is the cumulative value in the ith rectangle area of channel image C, the ICF value can be represent as $\sum_i^c w_i R_i(C_i)$, where w_i is the weight for the ith rectangle. Then a fast version of Adaboost algorithm [20] is utilized to do feature selection, and finally construct a strong classifier which comprise of a set of weak classifiers based on random sampled rectangles. In each round of Adaboost algorithm, it generates a large pool of rectangles, such as 30,000 areas, and finds the most discriminative one as current weak classifier. Fig. 2 shows the first ICF selected by the Adaboost algorithm.

2.2. Aggregate channel feature

The ACF is a simplified version of ICF, which shrinks the channel images from size $128 \times 64 \times 10$ to size $32 \times 16 \times 10$ with a factor of 0.25. In the shrank channel images, the simple zero-order information is used to represent features and fed into the Adaboost algorithm in order to train a high performance classifier. The difference between ICF and ACF is that, the former one need to sample sub-regions with different size and average the values in this sub-region as the features, but the later one ignores this step and use