Signal Processing 91 (2011) 773–781

Contents lists available at ScienceDirect

Signal Processing

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

Efficient HOG human detection

Yanwei Pang [a], Yuan Yuan [b,*], Xuelong Li [b], Jing Pan [c]

[a] School of Electronic Information Engineering, Tianjin University, Tianjin 300072, P. R. China
[b] Center for OPTical IMagery Analysis and Learning (OPTIMAL), State Key Laboratory of Transient Optics and Photonics, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an 710119, Shaanxi, P. R. China
[c] School of Electronic Engineering, Tianjin University of Technology and Education, Tianjin 300222, P. R. China

ARTICLE INFO

Article history:
Received 21 August 2009
Received in revised form
11 July 2010
Accepted 23 August 2010
Available online 16 September 2010

Keywords:
Image and video processing
Human detection
HOG
Fast algorithm

ABSTRACT

While Histograms of Oriented Gradients (HOG) plus Support Vector Machine (SVM) (HOG+SVM) is the most successful human detection algorithm, it is time-consuming. This paper proposes two ways to deal with this problem. One way is to reuse the features in blocks to construct the HOG features for intersecting detection windows. Another way is to utilize sub-cell based interpolation to efficiently compute the HOG features for each block. The combination of the two ways results in significant increase in detecting humans—more than five times better. To evaluate the proposed method, we have established a top-view human database. Experimental results on the top-view database and the well-known INRIA data set have demonstrated the effectiveness and efficiency of the proposed method.

1. Introduction

Object detection is an important step of high-level computer vision. Reliable object detection is essential to image understanding and video analysis. Faces and human bodies are among the most important objects in images and videos. Therefore, face detection and human detection have attracted considerable attention in applications of video surveillance [24,4], biometrics [25], smart rooms, driving assistance systems, social security [2], and event analysis. In this paper, we focus on human detection.

Detecting humans in images is challenging due to the variable appearance, illumination, and background [13,1,4]. Generally speaking, detecting humans in a static image (i.e. a single image or a video frame) is more challenging than in an image sequence. Detecting humans in an image sequence can usually be viewed as moving object detection. So motion information can be used [5,6]. Background can be modeled and used for foreground detection [3,6]. However, in a static image there is no motion clue and the background cannot be modeled. Usually, the core of detecting objects in a static image is effectively modeling the intrinsic characteristics of the objects. This paper limits its scope to human detection in static image.

Feature extraction and classifier designing are two key steps for reliable human detection in a static image. Haar-like rectangle features have been used for detecting humans [15] and faces [16]. To improve the representative and discriminating capacities, the original Haar-like rectangles have been extended to rotated features [17], diagonal features [17], and center-surrounded features [14]. The Haar-like rectangle features encode the intensity contrast between neighboring regions. Such features are suitable for face detection. All frontal faces have similar facial components and the facial components have fixed neighboring relationship. Importantly, the intensity contrast between neighboring facial parts is relatively stable. But the situation in human bodies is quite different. The intensity contrast between regions of a human body depends on the appearance of the humanwear, which varies randomly. So the Haar-like feature is not a discriminating feature for human detection. The human detection performance using merely Haar-like rectangle features is far from acceptable.

* Corresponding author.
E-mail address: yuany@opt.ac.cn (Y. Yuan).

0165-1684/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.sigpro.2010.08.010

Scale invariance feature transformation (SIFT) [19] is an alternative feature for human detection [20]. Phung and Bouzerdoum [21] proposed a novel feature called edge-density (ED) for human detection. Another feature taking advantages of edge information is edge orientation histogram (EOH) [22]. Region covariance matrix (RCM) [18,7] is among the state-of-the-art features for human detection. RCM is a matrix of covariance of some image statistics computed inside an image region. It is a matrix-form feature instead of the usual vector-form feature. In this paper we concentrate on the most successful and popular vector-form feature: histograms of oriented gradients (HOG) [9,10,26]. It is inspired by SIFT but different from SIFT. HOG can be regarded as a dense version of SIFT. It is shown that the HOG features concentrate on the contrast of silhouette contours against the background. Finally, it is noted that different types of features can be combined to enhance detection performance [13]. For example, Wang et al. [27] proposed to combine HOG features and local binary pattern (LBP) features in an elegant framework. But feature fusion/combination is beyond the scope of the paper.

The second step of human detection is designing classifier. Large generalization ability and less classifying complexity are two important criteria for selecting classifiers. Linear support vector machine (SVM) [12] and AdaBoost [23] are two widely-used classifiers satisfying the criteria. In this paper, we place emphases on the HOG feature and the SVM classifier and our contributions lie in efficiently computing of HOG features.

Histograms of Oriented Gradients (HOG) plus Support Vector Machine (SVM) [12] is one of the most successful human detection algorithms [9,10,11,27]. The HOG+SVM algorithm [9,10] employs sliding window principle to detect humans in an image. It scans the image at different scales and at each scale examines all the subimages. In each subimage, a 3780-dimensional HOG feature vector is extracted and SVM classifier [12,8] is then used to make a binary decision: human or non-human. Such a detection process is very slow. To overcome the drawback, Zhu et al. [11] proposed to use AdaBoost algorithm to select a small subset of the 3780 HOG features. However, in most cases, the classification accuracy is lower than that of the original HOG+SVM [9,10].

In this paper, we propose to speed up the HOG+SVM algorithm without sacrificing the classification accuracy. The contributions of the papers are: (1) by properly setting the scanning step, the block-based HOG features can be reused for all intersecting detection widows, which significantly reduces the computational cost. (2) We develop a sub-cell based interpolation algorithm to accelerate the calculation of the HOG features in one block, which removes unnecessary (at least unimportant) interpolation while the necessary interpolation is remained; and (3) to deal with occlusion problem; we propose to capture images in top view and we have established a top-view human database.

The rest of the paper is organized as follows: in Section 2, we describe traditional HOG+SVM based human detection algorithm. In Section 3, we describe the proposed method. The experimental results are presented in Section 4. Finally, we provide a brief summary in Section 5.

2. HOG+SVM based human detection

The success of the HOG+SVM human detection algorithm [9,10] lies in its discriminative HOG features and margin-based linear SVM classifier. The HOG+SVM algorithm concentrates on the contrast of silhouette contours against the background [9,10]. Different humans may have different appearances of wears but their contours are similar. Therefore the contours are discriminative for distinguishing humans from non-humans. It is worth noting that the contours are not directly detected. It is the normal vector of the separating hyperplane obtained by SVM that places large weights on the HOG features along the human contours. The HOG+SVM algorithm is outlined as follows:

Input: The scaled input image at the current scale. The size of sliding detection window is 64×128. The sliding step d (d=8 for example).
Output: The locations of the subimages of size 64×128, which are declared to contain humans.
Step 1: For each pixel of the whole image, compute the magnitude $|\nabla f(x,y)|$ and orientation $\theta(x,y)$ of the gradient $\nabla f(x,y)$.
Step 2: From top to bottom and left to right, scan the whole image with a 64×128 window. Extract the 3780 HOG features from the subimage covered by the detection (scanning) window and then apply the leaned SVM classifier on the high-dimensional HOG feature vector to classify the subimage as human or non-human.

The computation of HOG and the principle of SVM are described in the following two sections, respectively.

2.1. HOG

The first step of HOG extraction is to compute the magnitude $|\nabla f(x,y)|$ and orientation (angle) $\theta(x,y)$ of the gradient $\nabla f(x,y)$. The second step of HOG extraction is to derive the orientation histogram from the orientations and magnitudes. The size of the detection window is 64×128, which was experimentally determined for front view humans (see [10,9]). The subimage covered by the detection window (e.g. the dashed rectangles in Fig. 1(a)) is divided into 7×15 overlapping blocks. Each block consists of 4 cells and each cell has 8×8 pixels (see Fig. 1(b)). In each cell the orientation histogram has 9 bins, which correspond to orientations $i\times\pi/9$, i=0,1,…,8 (see Fig. 1(c)). Thus each block contains $4\times9=36$ features and each 64×128 subimage contains $7\times15\times36=3780$ features.

Trilinear interpolation can reduce the aliasing effect [9] and therefore is employed to compute HOG features. Trilinear interpolation smoothly distributes the gradient to four cells of a block. In Fig. 2(a), the 4 cells of a block are identified by their centers, (x_i,y_i), i=1,2,3,4. The 4 histograms corresponding to the 4 cells are represented by $h(x_i,y_i,\theta)$ where i=1,2,3,4 and $\theta\in\{0\times\pi/9,1\times\pi/9,\ldots,8\times\pi/9\}$. For a given gradient $\nabla f(x,y)$, its orientation $\theta(x,y)$ lies in the range $[i\times\pi/9(i+1)\times\pi/9]$ with i being a proper integer. We describe the range by $\theta_1=i\times\pi/9$ and $\theta_2=(i+1)\times\pi/9$ (see Fig. 2(b)). Thus $\nabla f(x,y)$ contributes to both $h(x,y,\theta_1)$ and