Neurocomputing 188 (2016) 139–150

Contents lists available at ScienceDirect

Neurocomputing

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

Joint components based pedestrian detection in crowded scenes using extended feature descriptors

Van-Dung Hoang [a], Kang-Hyun Jo [b,*]

[a] Quang Binh University, Vietnam
[b] University of Ulsan, Ulsan, Republic of Korea

ARTICLE INFO

Article history:
Received 24 July 2014
Received in revised form
19 October 2014
Accepted 2 November 2014
Available online 15 December 2015

Keywords:
Joint components
Haar-like feature
Histograms of oriented gradient
Boosting machines
Support vector machine
Multiple scales of block based HOG features

ABSTRACT

This paper presents a framework for human detection based on the joint component model using extended feature descriptors. This framework provides two contributions for handling the partially occluded problem of pedestrian detection in crowded environments. First, it presents feature descriptors based on extension of two well-known feature descriptors for accommodation in partially occluded pedestrian detection. The feature descriptors, which use multiple scale blocks-based histograms of oriented gradients (MHOG) and parallelogram based Haar-like feature (PHF), are proposed for improving the accuracy of the detection system. As a result of using MHOG, an extensive feature space allows for the obtaining of highly discriminated features, which supports robust detection. On the other hand, the PHF is adaptive for the shape of human limbs in detection of pedestrians. This contribution also presents special data structures to store image intensities and image gradients for using an integral image method, which is helpful for fast computation for both feature descriptors. Second, the joint component model for human detection is proposed for the training and detecting of pedestrians in crowded scenes. The proposed detection method based on fusion of the boosting technique and the support vector machine (SVM) is presented. The SVM is known as one of the most efficient learning models for classification. The advantage of boosting is a strong classifier based on a combination set of weak classifiers. However, the performance of boosting depends on the kernel of element classifier. The method based on the extended feature descriptors and the joint component model is helpful for constructing an efficient classification. The experimental results demonstrate that this method for pedestrian detection outperforms current methods in crowded situations under a variety of outdoor environments.

1. Introduction

In the present day, human detection systems based on vision sensors have been considered key tasks for a variety of applications, which have potential impact in many modern intelligence and surveillance systems, that involve knowledge integration and management and in autonomous systems relating to image retrieval, automatic personal assistance, and intelligent transportation. However, there are still many challenges in the detection procedures such as various articulate poses, appearances, illumination conditions and complex backgrounds of outdoor scenes, as well as partially occluded human in crowded scenes. In recent years, several methods for object detection have been proposed to deal with these problems, such as those described in [1–8]. A summary of the state-of-the-art research in the field of human detection was presented by Dollar et al. [9]. The related contributions can be roughly divided into two categories. The first group has focused on detecting humans under general circumstances. The standard approach investigated Haar-like features using the SVM classification for object detection [10,11]. However, the performance of Haar-like features is limited to human detection applications [1] because it is sensitive to a variety of human appearances, skeletons, complex backgrounds of scenes, and dynamic illumination in outdoor environments. To deal with these problems, other authors proposed the HOG descriptor [3,12–14]. Some authors proposed the typical methods based on neuron network for object classification and recognition problem, as presented in [15–19]. The standard human detection approach based on the HOG feature descriptor and the SVM classification machine were presented by Dalal and Triggs [3]. The result of that approach is acceptable under various conditions including illumination, distortion and the noise of the outdoor environments.

* Correspondence to: School of Electrical Engineering, University of Ulsan, Ulsan, Republic of Korea. Tel.: +82 52 259 1664.
E-mail addresses: dunghv@qbu.edu.vn (V.-D. Hoang), acejo@ulsan.ac.kr, hvzung@gmail.com (K.-H. Jo).

http://dx.doi.org/10.1016/j.neucom.2014.11.104
0925-2312/

The second group has investigated the detection of humans in special situations involving partial occlusion. Some researchers proposed special feature descriptors for representing objects, such as [20,21]. To deal with partially occluded people in crowded scenes, Wang et al. [20] proposed a new feature descriptor based on combining the HOG and the Local Binary Pattern (LBP) using the SVM classification machine for detecting the partially occluded human in crowded scenes. In that system, the authors accumulated both the HOG and the LBP features to construct feature vectors, which are fed into the linear SVM in both the training and the detection stages. The experimental results indicated that the system was capable of handling partial occlusion. However, combining both feature descriptors is complicated by high computational cost. In another approach, Felzenszwalb et al. [5] presented a full learning system for recognizing objects. In that paper, the object detection system uses the deformable part model (DPM), which is trained by using a discriminative approach based on bounding boxes of objects within images. Schwartz et al. [22] proposed a method for integrating whole body detection with face detection to reduce the false positive rate. However, the camera orientation is not always opposite pointing at the human face, so the face does not appear clearly in image for detecting. The authors in [23] proposed a method for the detection and tracking of occluded humans based on learning human parts in some special situations. In that contribution, the authors presented the approach based on the articulations of the body in appearance changing using SVM classification for detecting and tracking. The subset of parts is selected based on maximizing the classification probability for improving the accuracy of the human detection system in crowded scenes. The tracking stage is proposed for addressing the problem of occlusion by learning based on corresponding parts, which helpfully detects and predicts partial occlusions while maintaining the performance of the system. Tang et al. [24] proposed a method using the DPM for detecting occluded people in crowded scenes with two phrases. The first stage simultaneously detects humans that overlap with each other based on a double-person detector. The second stage detects humans using joint detectors based on training using the special double-people dataset. To improve the accuracy, the training data is built under special situations for a single-person and double-person. However, the precision of the system depends on the training dataset and has a high computational cost.

This paper focuses on detecting occluded pedestrians in crowded scenes, offers two major contributions. First, two extensions of feature descriptors are proposed to improve the accuracy of the systems: The multiple scale block-based histograms of oriented gradients feature (MHOG) and the parallelogram-based Haar-like feature (PHF). The MHOG result in extensive feature space, which allows for the selection of highly discriminative features to obtain high accuracy classification. To that end, the set of HOG features within each block (HOGB) is used to feed into the SVM classifier. The boosting technique is used to select high discriminative HOGBs, which are used in the classification stage. The MHOG is utilized for representing the full body, as well as features of the head and torso. Otherwise, the PHF is adaptive for the limb shapes of human. The fast computing feature descriptors method based on the "integral image" technique for each feature type is also proposed. The second contribution provides an efficient detection approach using an interpolation method for combining human components, which supports the detection of the partially occluded people in crowded scenes. The candidate regions are filtered using full body detection so that it achieves a high true positive rate and a high false positive rate. This condition is able to ensure that the candidate regions cover all people in the image, although it includes a high misdetection rate because almost all partially occluded humans in crowed scenes are caused by people overlapping each other. Finally, component interpolation is applied for estimating the human region, with the SVM classification used to detect each component. The boosting technique is used as a global system for interpolating components to construct the components based full human. The component detector is used as a weak classifier inside of the boosting technique.

2. Feature description

This section proposed two feature descriptors: the MHOG based on the histogram of oriented gradients feature descriptor and the PHF based on the Haar-like feature descriptor. The PHF feature is proposed for describing human arm and leg components. The MHOG feature is used for the description of the full body, as well as the head and torso components.

2.1. PHF feature descriptor

This subsection presents a new feature descriptor, which extends the Haar-like feature descriptor [25]. The PHF feature represents the difference of intensities in adjacent parallelogram regions [26], as depicted in Fig. 1. The original Haar-like feature used the "integral image" method based on a cumulative sum of intensities within rectangular regions, which supports fast computing a feature with only eight accesses to regions of any size. However, the feature descriptor is restricted within rectangle regions, so it is not adaptive to various poses of the human components. The PHF descriptor is based on a modified Haar-like feature, which significantly enriches the basic set suitable for the detection of skewed components of human shape, e.g., legs and arms part.

The most expensive step of the feature process is computing the sum of intensities within each region. This operation is repeated many times for different scale levels. This task focuses on reducing the computational time of the computing features. As mentioned above, the PHF is based on the basic sum of image intensities and the feature is computed in multiple scale areas. Therefore, the fast computing method is proposed to avoid repeated computation. First, the cumulative sum of intensities is computed only one time for the overall image and is stored in a table, known as the summed table of parallelogram region (*TP*). The feature is computed based on the difference of sum intensities within the white region and gray region. The table *TP* is used to compute the sum image intensities within a parallelogram region (*SP*) only requiring four accessed operations for any sized region. Notice that each kind of feature descriptor should be computed in different ways. The details of the process are presented as follows.

The *TP* of the first feature type in Fig. 1(a) is denoted as $TP^{(1)}$. The cumulative sum of image intensities is calculated by starting the top-left corner to the bottom-right of the image, as illustrated in Fig. 2(a), as described in following formulation:

$$TP^{(1)}(x,y) = \sum_{j=1}^{y} \sum_{i=1}^{x+y-j} I(i,j) \quad (1)$$

where $I(i, j)$ is the intensity of the image at pixel (i, j).

Taking advantage of the *TP* characteristics, the late elements of table *TP* are computed based on previous elements plus the current image intensity. The expanded Eq. (1) is:

$$TP^{(1)}(x,y) = TP^{(1)}(x-1,y) + TP^{(1)}(x+1,y-1) - TP^{(1)}(x,y-1) + I(x,y) \quad (2)$$

where $TP^{(1)}(x,0) = 0, \quad TP^{(1)}(0,y) = TP^{(1)}(1,y-1)$ for all x, y.

Similarly, the $TP^{(2)}$ table of the second feature descriptor type is computed in a similar manner with some modifications. It is the