Signal Processing 112 (2015) 43–52

Contents lists available at ScienceDirect

Signal Processing

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

Sparse auto-encoder based feature learning for human body detection in depth image

Song-Zhi Su [a,b], Zhi-Hui Liu [a,b], Su-Ping Xu [a,b], Shao-Zi Li [a,b], Rongrong Ji [a,b,*]

[a] School of Information Science and Technology, Xiamen University, Xiamen 361005, China
[b] Fujian Key Laboratory of the Brain-like Intelligent Systems, Xiamen University, Xiamen 361005, China

ARTICLE INFO

Article history:
Received 18 March 2014
Received in revised form
3 November 2014
Accepted 4 November 2014
Available online 20 November 2014

Keywords:
Human detection
Depth image
Feature learning
Sparse auto-encoder

ABSTRACT

Human body detection in depth image is an active research topic in computer vision. But depth feature extraction is still an open problem. In this paper, a novel feature learning method based on sparse auto-encoder (SAE) is proposed for human body detection in depth image. The proposed learning based feature enables capturing the intrinsic human body structure. To further reduce the computation cost of SAE, both convolution neural network and pooling are introduced to reduce the training complexity. In addition, upon learning SAE based depth feature, we further pursuit the detector efficiency. A beyond sliding window localization strategy is proposed based on the fact that the depth values of object surface are almost the same. The proposed localization strategy first uses the histogram of depth to generate candidate detection window center, and then exploits the relationship between human body height and depth to determine the detection window size. Thus, it can avoid the time-consuming sliding window search, and further enables fast human body localization. Experiments on SZU Depth Pedestrian dataset verify the effectiveness of our proposed method.

1. Introduction

Human body detection is a fundamental step for understanding human behaviors in camera(s) with applications to intelligent video surveillance, assistant vehicle driving, and automatic action recognition. Therefore, it has become an active research filed in the computer vision community. However, the appearance of human body is distracted by variations caused by different illuminations, poses, viewing angles, and partially occlusion. It is therefore fairly challenging to build a human body detector in real scenes.

RGB camera was the main imaging device in the early stage of human body detection. Extensive progresses on detection algorithms based on RGB or gray images [1–4] have been obtained. However, RGB or gray images are still suffering from photographing variations as stated above, which largely decrease the performance of detection algorithms. Fortunately, with the popularity of depth sensors like RealSense and Kinect, it is now feasible to obtain the distance between the surface of objects and the camera in dynamic environments. Moreover, those depth images are insensitive to illumination variations and shadows. As a result, research on detection algorithms over depth image has attracted ever increasing attentions.

In this paper we pursuit the detection efficiency. Currently, sliding window (SW) is the mainstream approach in object or human body detection, e.g. as in the PASCAL Visual Object Classes (VOC) challenge [5], the majority of the entries used a "sliding window" approach to the detection task. Essentially, detection or localization is a binary classification problem to determine whether there is an object in

* Corresponding author at: School of Information Science and Technology, Xiamen University, Xiamen 361005, China.
E-mail address: rrj@xmu.edu.cn (R. Ji).

http://dx.doi.org/10.1016/j.sigpro.2014.11.003
0165-1684/

a given scanned window, to which end classifiers such as Neuro-network (NN), Support Vector Machines (SVMs), Random Forest (RF), and AdaBoost are widely investigated. In terms of feature part, descriptors such as Histogram of Oriented Gradient (HOG) [6], Local Binary Pattern (LBP) [7], Integral Channel Feature [8], and Haar-like Feature [9] are widely used.

To carry out human body detection in depth image, the first task is to design discriminative descriptor. In the literature, quite few feature extraction algorithms are developed exclusive for depth image. Most of the proposed depth descriptors are similar to those in RGB domain. For example, Spinello and Arras [10] and Wu et al. [11] proposed Histograms of Oriented Depths (HOD) and Histogram of Depth Difference (HDD), respectively, both of which are very similar to HOG. Yu et al. [12] proposed a Simplified Local Ternary Patterns (SLTP) descriptor, which was an improvement of Local Ternary Patterns (LTP). Ikemura and Fujiyoshi [13] proposed a descriptor called Relational Depth Similarity Feature (RDSF), which is based on statistic features on depth values. However, the above features were handcraft designed. Therefore to some extent, these features only encode part of the human body information while neglecting the intrinsic structure of the human body, which are therefore not discriminative enough in complex scenes.

Along with the recent advance on sparse coding and deep learning, learning based feature is receiving ever-increasing research attention. For instance, Ren and Ramanan [14] proposed a learning based feature based on sparse coding. Dollar et al. [8] proposed a learning based integral channel feature, with encouraging performance on pedestrian detection. Sermanet et al. [15] also proposed an unsupervised multi-stage feature learning method for pedestrian detection. However, these features work on the RGB image domain only. And in the literature, there are few works on learning based depth features. Recently, sparse auto-encoder (SAE) becomes popular, and can automatically learn features from unlabelled data. SAE has got satisfying performance on many applications, such as image classification, voice recognition, and hand gesture recognition. In this paper, we introduce SAE to learn depth image feature for human body detection. SAE also brings another merit on the efficiency, as its objective function can be solved via fast backward propagation.

Upon learning SAE based depth feature, we further pursuit the detector efficiency: during the human body localization, it would be time-consuming to use sliding window to scan all the windows in the image scale-space. Researchers have proposed methods to reduce the search space in RGB images, such as brand-and-bound [16] and jumping windows [17]. In some special cases, supplementary information of scenes can be exploited to reduce the search space. For example, under the circumstance that the camera is fixed, moving targets can be extracted by background subtraction, and then the detection can be only performed on the foreground regions. Other researchers also use the scene geometry information to reduce the search space [18]. In the driver assistant system, regions on the road can be quickly identified by calibrating the cameras [19], and then the potential objects are detected only on those regions. In depth images, since depth values on the object surface are aggregated together, potential positions of the objects can be directly predicted before detection. This inspires us to propose a method to accelerate detection speed.

Overall, a detection framework beyond sliding window was proposed aiming at fast human body detection in depth image. Our contributions are three-fold: (1) in terms of discriminative depth feature extraction, we introduce SAE to learn depth feature automatically for human body detection. (2) Histogram of depth is exploited to generate candidate detection windows, thus avoiding the time-consuming exhaustively sliding window search. (3) We model the relationship between human body height and depth values, which is then used to determine the object size, thus avoiding the multiple scale scanning that is typical in the sliding window based detector.

This paper is outlined as follows. We summarize the state-of-the-art human body detection in RGB images and depth images in Section 2; Section 3 gives the overview of the proposed human body detection systems; the SAE based feature learning method is described in Section 4; experimental results are presented in the Section 5; the conclusions and future work are presented in Section 6.

2. Related work

2.1. Human body detection in RGB images

Roughly speaking, there exist three human body models, i.e., Holistic model, part-based model, and patch-based model.

Holistic models: The Holistic model, which is also called the monolithic model, extracts human body feature as a whole rather than concatenating part-based features. The extracted feature (also called the descriptor) in a scanning window is then fed into a binary classifier to decide whether the window contains human body. Various human body descriptors have been proposed, such as Edge Templates [20], Haar-like feature [9,21], HoG [6], Integral Channel Features [8], and Local Binary Pattern [7]. These descriptors are based on color, texture, or gradients. Some learning based mid-level descriptors are also proposed, for example, Edgelet [22] and Shapelet [23]. As for the classifier used for sliding window detection framework, neural networks (NN), support vector machines (SVM), and Adaboost are widely used. Many variants of these classifiers are proposed recently. For instance, Tuzel et al. [24] modified the traditional boosting framework on Riemannian manifolds, making it work well with their proposed covariance matrices descriptor. Maji et al. [25] proposed an approximation to histogram intersection kernel for SVM (HIKSVM), allowing a nonlinear SVM to be used in sliding-window detection framework rapidly. Chen and Chen [26] proposed a cascade of Real Adaboost classifier with meta-stages, where both the stage-wise classification information and the inter-stage information are exploited to improve detection performance.

Part-based models: Instead of the holistic model, part-based model accounts for a wider variety of human poses. This is achieved by modeling the geometrical relation