J. Vis. Commun. Image R. 31 (2015) 177–185

Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

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

An ultra-fast human detection method for color-depth camera ☆

Jun Liu [a], Guyue Zhang [a], Ye Liu [b], Luchao Tian [a], Yan Qiu Chen [a,*]

[a] School of Computer Science, Fudan University, Shanghai 201203, China
[b] College of Automation, Nanjing University of Posts and Telecommunications, Nanjing 210023, China

ARTICLE INFO

Article history:
Received 26 September 2014
Accepted 23 June 2015
Available online 29 June 2015

Keywords:
Human detection
Color-depth image
Depth camera
RGB-D sensor
Kinect
Real-time
Cascade-structured
Cluttered and dynamic environments

ABSTRACT

Real-time human detection is important for a wide range of applications. The task is highly challenging due to occlusions, complex backgrounds, and variation of human poses. We propose a cascade-structured approach to real-time human detection in cluttered and dynamic environments with both color and depth data seamlessly incorporated. The first stage efficiently exploits depth data which generates a set of physically plausible yet over-detected candidates. These candidates are then purified by another two filters: a knowledge based human upper portion locator and a data-driven learning based filter. Experimental results show high detection accuracy achieved by the proposed method at 80–140 fps on a single CPU core (without GPU acceleration).

1. Introduction

Robust and efficient human detection has been a very active research area for several decades [1–6]. The importance of this problem originates from its wide application in security surveillance, human–machine interaction, autonomous vehicle, robotics, wearable computer, etc. This task proves highly challenging due to many factors, such as occlusions, complex background, variations in illumination conditions and human poses. Moreover, the detector should be fast enough to be practically useful for real-time applications.

Recent years have seen considerable progress in this area and a large number of methods have been proposed [7,8]. Most of these methods use conventional video cameras that do not provide depth data. Detectors based on appearance features of human subjects, such as EOH [9], HOG [10,11] and LBP [12,13], are reported to be effective by several papers, but may dramatically deteriorate in more challenging real-world tasks where detection accuracy and computation efficiency are both critical.

An effective way to counter the challenges is to employ depth information [14–17]. The depth image produced by stereo rigs or depth cameras significantly simplifies the task of real-time people detection, but there are still difficulties in dealing with occluded subjects in horizontal front view images. A top-down view setup [18] is a good choice for decreasing overlapping and facilitating segmentation, but the visual field is limited and details of human subject cannot be well observed. A smart methodology taking advantage of top-down view while maintaining sufficient visual details is adopted by [19] and our previous work [20]. These methods take a front (or oblique) view, and project the 3D point cloud of the scene onto the ground plane to obtain a virtual plan view. With this virtual view, occlusion is dramatically reduced. Though a variety of interesting results are achieved, the need of additional information of ground plane equation makes these methods difficult to adapt to more complicated environments when the camera is moving (in autonomous vehicle, robotics, and wearable computer applications) or the observed objects are on the stairs. There are approaches [21,22] utilizing neither top-down view nor virtual plan view, but the requirements of both high detection accuracy and very fast speed in some challenging applications still make the task largely open. The method proposed by Xia et al. [23] uses a 2D head contour and a 3D hemisphere model to match heads in depth image, but it requires multi-scale sampling which limits the efficiency. The system of Choi et al. [24] adopts a depth based template and has difficulties in locating people that are not fully observed.

This paper addresses the problem of fast human detection in color-depth images captured from horizontal (or slightly oblique) view. To achieve both high accuracy and high efficiency, a multi-stage detection framework is developed. The initial retriever localizes the plausible candidates in about 2 ms. Then the resultant

☆ This paper has been recommended for acceptance by Yehoshua Zeevi.

* Corresponding author.
E-mail addresses: ljchangyu@hotmail.com (J. Liu), guyuezhang13@fudan.edu.cn (G. Zhang), yeliu@fudan.edu.cn (Y. Liu), hualitlc@163.com (L. Tian), chenyq@fudan.edu.cn (Y.Q. Chen).

http://dx.doi.org/10.1016/j.jvcir.2015.06.014
1047-3203/

responses are progressively filtered by two effective filters which in order are a novel knowledge based upper body locator performing well under occlusions and depth data loss, and the Joint Histogram of Color and Height (JHCH) based classifier proposed in previous work [25]. These two filters respectively exploit shape and appearance information of human body, and can both refine the results produced by their previous stages in 1–5 ms. This cascade structure allows very fast detection and yields good performance in our evaluation experiments, in which both human activities and backgrounds are quite complicated.

The main contributions of this work include:

(1) An extremely fast technique to locate positions that are plausibly humans is proposed to quickly reduce searching space and eliminate the need by subsequent procedures to scan the entire image.
(2) A novel knowledge based human locator which can deal with partial occlusion and incomplete depth data is proposed.
(3) A stepwise filtering framework enables the system to perform very quickly as increasingly fewer candidate positions need to be examined by latter stages.

The rest of this paper is organized as follows. Section 2 introduces our approach to detecting human beings in color-depth images. Section 3 presents the experimental results and discussions. Finally, Section 4 draws conclusions.

2. Method

The motivation behind our design is to use the first stage to very quickly scan the depth image to detect locations that may potentially be head-tops. We try to ensure all true head-top locations are included in the responses while discarding as many unfeasible locations as possible. The latter stages are high accuracy filters to try to keep all but only true head-tops. This design philosophy can achieve both high speed and high accuracy. An overview of the proposed detection method is shown in Fig. 1.

2.1. Fast scan to find plausible locations

We propose a novel ultra-fast assessing-while-scanning technique, in which a very simple and high-speed operation is performed on every pixel to assess whether it has potential to be a head-top position.

Head-top is generally the highest point of the human body and has depth discontinuities against nearby background pixels in depth image. This observation leads us to find an effective cue that the depth value differences between the head-top pixel and the neighboring pixels from the row above it are always large even in highly crowded and complicated situations. This criterion can be used for discarding positions that are implausible to be head-tops.

We use a triple (r_p, c_p, d_p) to denote respectively the row index, column index, and depth value[1] of pixel p in depth image. For every pixel p being scanned, we construct an upper row pixel set:

$$\wp = \{x | r_x = r_p - 1, |c_x - c_p| \leqslant \gamma_p\}, \tag{1}$$

where x is a pixel from the above row of p (the above row is illustrated by the dotted line in Fig. 3), and γ_p is the radius of head in depth image. We use ω to denote the head radius in world coordinates, then the radius of head projected into image can be calculated as $\gamma_p = \frac{\kappa}{d_p} \cdot \omega$, where κ is a constant factor which can be obtained with camera's intrinsic parameters [26].

Consequently, for a true head-top pixel p, the pixel set $\wp$ satisfies:

$$\forall x \in \wp, \quad |d_x - d_p| > \omega. \tag{2}$$

This criterion is very efficient for locating plausible candidates, since only a few pixels are examined for each position. Even in highly crowded environments, this stage makes sure that people's head-top positions are included in the resultant responses. Although there are false positives, the space to be further searched has been significantly reduced. As shown in Fig. 1(c), only 29 positions are detected as candidates. Detected candidates with the same row index, adjacent column indexes and small depth gradients ($< \omega$) are considered to belong to the same person head, and only one of them is kept. The resultant positions may not be the exactly physical head-top of human subject when the image is captured obliquely (or tiltedly), but the overall performance of our detection method does not drop if the oblique angle is small ($< 30^\circ$).

2.2. Knowledge based human locating

Head-shoulder template matching [19,27,23] is widely used for detecting human beings, but it often fails in handling partially observed subjects. We propose a novel knowledge based human locator using shape information of human body which can deal with partial occlusion and depth data loss to filter the results produced by the previous stage.

The idea is to design a Ring-wedge Mask (RWM) [28] to segment the underlying nearby area of a head-top position p. As illustrated in Fig. 2, we use 3 rings and 18 wedges, and divide the potential human upper area into 54 subareas, then features can be extracted from these subareas. The rings have the same center (with row at $r_p + \gamma_p$ and column at c_p) and different outer radiuses. The radiuses can be formulated as $R_i = f_i \cdot \gamma_p$, where $i = 1, 2, 3$, and $f_1 < 1 < f_2 < f_3$.

As shown in Fig. 2, the innermost 18 subareas (denoted by Group I) and lower outermost 5 subareas (denoted by Group II) are all located in human body and their depth values are close to d_p (depth value of head-top pixel), while situation of the upper outermost 9 subareas (denoted by Group III) is the opposite. These three subarea groups forming a distinguishable human upper portion pattern can be used for human localization. However, in real-world environments, this pattern is often corrupted due to variation of human directions, depth data loss, and partial occlusions. In order to make this pattern practically useful, we evaluate the possible situations of the pixels that are close to the head-top pixel and try to categorize them.

Generally, the nearby area pixels of the person's head-top in depth image can be divided into 4 categories (as depicted in Fig. 3):

Categ. 1 (detectee): Pixels that are part of human detectee.
Categ. 2 (overlapper): Pixels that do not belong to this subject but occlude it;
Categ. 3 (background): Pixels that can be regarded as the background of the person;
Categ. 4 (depth loss): Pixels that encounter depth data loss.

The categorization accuracy of a pixel is sensitive to the category calculation method. Interestingly, a very simple scheme turns out to be good enough. A very high speed yet quite effective scheme is used to identify the category index of the pixel x by comparing its depth value with that of the head-top:

[1] Depth value records the Euclidean distance from the camera plane to the object.