Detecting and tracking people in real time with RGB-D camera

Jun Liu a, Ye Liu b, Guyue Zhang a, Peiru Zhu 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

Abstract

We propose a novel approach to automatic detection and tracking of people taking different poses in cluttered and dynamic environments using a single RGB-D camera. The original RGB-D pixels are transformed to a novel point ensemble image (PEI), and we demonstrate that human detection and tracking in 3D space can be performed very effectively with this new representation. The detector in the first phase quickly locates human physiquestwise plausible candidates, which are then further carefully filtered in a supervised learning and classification second phase. Joint statistics of color and height are computed for data association to generate final 3D motion trajectories of tracked individuals. Qualitative and quantitative experimental results obtained on the publicly available office dataset, mobile camera dataset and the real-world clothing store dataset we created show very promising results.

© 2014 Elsevier B.V. All rights reserved.

1. Introduction

Detecting and tracking people in video sequences is a key research problem that has wide application in security surveillance [33], people collective behavior study, etc., and has attracted a great deal of research attention [8,10,19,26,32]. The problem remains largely open due to many serious challenges, such as occlusion, change of appearance, complex and dynamic background, etc.

The vast majority of early work [22,23,34] used conventional video cameras that lack depth information. The task of detecting and tracking people in such image sequences has proven very challenging although sustained research over many years has created a range of smart methods that work quite well on some benchmark videos. Several other use statistical learning with local features and boosting, such as EOH [17], HOG [4] and edgelet [27]. Although many papers report that these methods can lead to good detection and tracking results on their benchmarks, their performance will dramatically deteriorate in more challenging real-world applications, since features obtained from conventional cameras encounter numerous difficulties in characterizing humans if the background is dynamic and complex. Other challenges such as vast variation in human poses, self-occlusions and cross-occlusions make the problem even more complicated.

To counter these challenges, several approaches [6,7,13,30] using depth data (or disparity image) produced by stereo rigs are proposed. The method proposed by Mitzel and Leibe [21] uses a learned upper-body depth template and can work well in crowded scenarios, but it requires multi-scale down-sampling (sampling rate will restrict the accuracy and operation efficiency). The system of Bahadori et al. [1] exploits the correspondence between image blobs and world blobs, which can alleviate occlusions in original images but still faces difficulties in blob segmentation when people are highly crowded. Though these methods show advantages over those using conventional cameras which lack 3D information, the quality of depth data is limited as it is generated by stereo matching algorithms which encounter difficulties in handling textureless regions and depth discontinuities. In addition, computation cost of stereopsis without dedicated hardware is high, thus real-time processing is difficult to achieve [15].

More recently, depth cameras such as Kinect [12], Xtion and TOF cameras have become widely available at affordable prices, and compared to stereo rigs, the quality of depth maps has been greatly improved. Studies using these sensors [20,31] have demonstrated the great value (both in accuracy and efficiency) of depth camera in coping with severe occlusions among humans and complex background. In existing works, depth cameras are often placed either vertically overhead [2,5] or horizontally at the same level as humans [11,25]. Occlusion is significantly reduced in the former case, details of the human body however cannot be well captured. Advantages and shortcomings for the latter case are just the opposite: human body can be observed more completely but severe occlusions may happen frequently. An oblique view may be a good trade-off between completeness and occlusion. Even with this setup, it is still challenging to detect partially

* This paper has been recommended for acceptance by G. Sanniti di Baja.

** Corresponding author. Tel.: +86 21 65643842; fax: +86 21 65643842.
E-mail addresses: ljchangyu@hotmail.com (J. Liu), yeiliu@fudan.edu.cn (Y. Liu), guyuezhang13@fudan.edu.cn (G. Zhang), peiruzhu@gmail.com (P. Zhu), chenyq@fudan.edu.cn (Y.Q. Chen).

http://dx.doi.org/10.1016/j.patrec.2014.09.013
0167-8655/© 2014 Elsevier B.V. All rights reserved.
We propose in this paper a real-time system to detect and track people in oblique-view RGB-D videos. The proposed method is three-stage cascade structured. In the first stage, RGB-D data is transformed to point ensemble image (PEI) from plan-view perspective to facilitate the subsequent two stages. In the second stage, an unsupervised detector retrieves positions that are human body plausible very quickly, and then these positions are further refined by a classifier using two new features: histogram of height difference (HOHD) and joint histogram of color and height (JHCH). Finally, data association is carried out to the detection responses to generate the 3D trajectories. Experimental results show that the proposed cascade structure offers very fast detection and tracking, and produces significant performance advantages over existing methods. A preliminary account of this work has been presented at [18].

The remainder of this paper will describe the details of the proposed method in the following Section 2, and will present the experimental results in Section 3. Conclusions will be drawn in Section 4.

2. Method

Our motivation is to develop a human detection and tracking system using a single RGB-D camera to achieve both real-time operation and high accuracy. In order to achieve this objective, we have developed a novel PEI representation to overcome segmentation difficulties encountered by directly using RGB-D image. We have also developed a multistage incremental filtering detection method so that implausible candidates are very quickly discarded by the beginning procedures to improve computational efficiency and the remaining much fewer candidates are scrutinized by the latter procedures to obtain high accuracy. We propose a novel JHCH for associating the detection results, and together with Kalman Filter we obtain the 3D motion trajectories of the humans. Fig. 1 shows an overview of the proposed system.

2.1. PEI representation

Due to frequent mutual occlusions among humans from an oblique view, segmentation in original image domain often suffers from under-segmentation and over-segmentation. These two troubles can be greatly alleviated by performing a plan-view transformation to the original data, as shown in Fig. 2.

With depth information, pixels can be back-projected into 3D space to construct a 3D point cloud in camera’s coordinates. We can observed humans while keeping a very low false alarm rate in complex environments. Moreover, the method should be fast enough in order to be useful in real-time applications.

Fig. 2. (a) 3D point cloud of a person. (b) The point cloud in plan-view. (c) In height map representation, only the upper portion of the body is kept well. (d) In the proposed PEI, all of the pixels (points) from original RGB-D image (3D point cloud) are maintained. To differentiate, the cells in the plan-view are color coded in (c) and (d), and each 3D point which will be recorded in the projected cell is labeled with the corresponding color of that cell. In (b) and (d), the green dotted lines indicate the neighborhood area of the human head crown in PEI. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)
are defined as amplified in many cases. The square grid for discretizing the plan-view floor. With this virtual view, segmentation can be dramatically simplified in many cases. The square grid for discretizing the plan-view system, is projected to using a square grid. The point cloud, mapped to the new coordinate system, is projected to color in this figure legend, the reader is referred to the web version of this article.)

In a height map representation [1], each cell of the plan-view image keeps the information of the highest point (the one with the maximum Z value) which are projected into that cell, as shown in Fig. 2(c). Although interesting results have been achieved, much of the detail information is lost [30]. We propose a novel PEI, as illustrated in Fig. 2(d), to take advantage of plan-view while preserving the information of all 3D points.

In PEI (denoted by E), each cell records an ensemble of points projected into that cell and can be formulated as:

\[ E^i_j = \{ p \mid p \in P, (x^p, y^p) \in g^{i,j} \} \]

where \( P \) is the point cloud in the new coordinate system, and \( p = (x^p, y^p, z^p) \) is a 3D point in \( P \).

2.2. Human detection

Background subtraction is widely used in human detection, but in real-world situations, such as supermarkets and clothing stores, it may not work well since background is often dynamic. There are several depth camera based human detection methods [5,28] without background subtraction. Their performances will in many situations be limited as the shape prior of head and body which is crucial in these methods can be easily corrupted in the case of mutual occlusions and incomplete depth data.

We do not carry out detection by imposing strong shape prior on targets, instead we do it progressively. Our two-phase filtering method is able to leverage the advantages of depth and RGB data jointly. In the first phase, most of the false positives that are human physiquestwise implausible can be highly quickly rejected, then the responses are further purified by a learning based classifier in the second phase (illustrated in Fig. 3).

2.2.1. Phase 1: Detecting human physiquestwise plausible candidate locations

In the preliminary detection phase, we try to find the 3D points which are higher than all other points within its neighborhood (local height maxima). For a point \( p \) in the point cloud, we draw a cylindrical neighborhood with radius \( \omega/2 \) (\( \omega \) is the average width of human torso), as shown in Fig. 2(b)-(d). The height values of the points inside the cylinder are compared with that of \( p \). This computation is very efficient thanks to the point ensemble image representation. The set of neighboring points of \( p \) can be calculated on PEI (denoted by \( E \)) as:

\[ N^e_c = \{ x \mid x \in c', c' \in E, |c' - c| < \omega/2 \} \]

where \( c \) is the cell of \( E \) that \( p \) is projected into, and \( c' \) denotes the nearby cells of \( c \) (including \( c \)).

All the points with the maximum height value in their neighborhoods are regarded as potential crowns of human heads (the whole neighboring point set represents the human body). To improve processing speed and also reliability, only the points whose heights are between a min-height threshold \( h_{min} \) and a max-height threshold \( h_{max} \) are taken into account, which filters out a large amount of unfeasible points, such as those near floor and ceiling.

Even in highly complex and crowded situations, this stage makes sure that true head crown locations are often contained in the resultant responses. Even though there are false positives, the space to be further searched has been substantially reduced.

2.2.2. Phase 2: Learning based refinement

We then try to determine whether the objects picked up by the previous phase are indeed humans. We train a classifier with two types of features encoding shape and appearance information respectively. Both of them are extracted in the neighboring point set.
collected from PEI. The features are based on the upper portion of human body as the lower part suffers from more frequent occlusions and is more deformable. The descriptions of the features are as follows:

**Histogram of height difference (HOHD).** The 3D shape of human’s upper part is very distinct from other objects. This can be observed clearly in plan-view, as shown in Fig. 2(b). In plan-view, the height of the head crown is larger than other points on the head and significantly larger than the shoulder area points. Besides, the height difference between head crown and shoulder is often in a small and learnable range. This kind of height variation can be described with height differences in a statistical manner.

For a local height maximum (the head crown), its height value is subtracted by those of the points in its cylindrical neighborhood, which are collected from PEI. As we just focus on body’s upper part here, only a subset of \( N \) is selected, in which the points satisfy the following two criteria:

(a) The point has the largest height value in the cell it is projected into;

(b) The height difference between the point and the head crown is less than a threshold \( \eta \), a quarter of average human height.

Given the statistical data of height difference, we construct a normalized histogram with 15 bins. By constructing a histogram, we avoid making hard assumptions on human shape like [28], thus this feature is effective on noisy and even fragmentary data.

**Joint histogram of color and height (JHCH).** Color distribution [23] and height (or depth) pattern [3] have proved to be effective in human detection and tracking, and simple linear combination of the two has been exploited [5]. In order to utilize the correlated information between them to achieve higher performance, we propose a JHCH with 5 height intervals to characterize the appearance information of the human head. As shown in Fig. 4(a), the color statistics of the head are basically collected from hair and face or hair alone (if human is observed from the back) and are located at different height levels. Since the hue component of HSV color space is less sensitive to illumination change, it is used as the color model of our histograms. In our work, the hue range is divided into 9 intervals, saturation range into 5 intervals, and height range into 5 intervals. Only those points that are in the cylindrical neighborhood of the head crown are meaningful to represent the person and are selected to build 3D JHCH.

We use \( r \) to denote the number of bins in 3D JHCH (9 \( \times \) 5 \( \times \) 5), and sequentially enumerate these bins from left to right, top to bottom, and front to back, then the JHCH of target \( t \) can be denoted as \( H_t = \{ H_{t}^{k} \}_{k=1}^{9} \), where \( k \) is the bin index.

As different body parts may have different image stability levels (the upper part is more reliable as it is less likely to be occluded), different weights are assigned to the 3D points at different height levels by employing a weighting function as follows:

\[
w(p_i) = \exp(\alpha \cdot h(p_i)),
\]

where \( h(p_i) \) is the height value of point \( p_i \), and \( \alpha \) is a weight. So each bin of JHCH can be calculated as:

\[
H_{i}^{k} = f \cdot \sum_{p \in N} \{ w(p_i) \cdot \delta(\beta(p_i) - k) \}.
\]

where \( \delta \) is Kronecker delta function, \( \beta(p_i) \) indicates the bin of JHCH that \( p_i \) should be assigned to, \( N \) is the neighboring point set collected from PEI, and \( f = 1/ \sum_{p \in N} w(p_i) \) is the normalization factor ensuring \( \sum_{k=1}^{9} H_{i}^{k} = 1 \).

Then we can define a similarity function between targets \( a \) and \( b \) as:

\[
\zeta(a, b) = \gamma \cdot \rho(H_a, H_b) + (1 - \gamma) \cdot \psi(L_a, L_b).
\]

where \( L_i \) is the spatial location of object \( i \) (i.e., \( a, b \), \( \rho(H_a, H_b) \) is the Bhattacharyya similarity [16] between the JHCHs of objects \( a \) and \( b \), \( \psi(L_a, L_b) \) represents the spatial location similarity between the two objects, and \( \gamma \) is a weighting factor to control the relative importance of the appearance similarity and 3D location similarity.
\[ \psi_l(a, b) = \exp \left( -\mu \sum_{k=1}^{r} \sqrt{H_{a,k}^2 + H_{b,k}^2} \right) \]  
\[ \psi_l(a, b) = -v \cdot D_l(a, b) \]

where \( D_l(a, b) \) is the Euclidean distance between \( a \) and \( b \), and both \( \mu \) and \( v \) are weighting factors. The similarity value \( \psi(a, b) \) is between 0 and 1, and \( \psi(a, b) = 1 \) indicates that the two objects are perfectly matched with each other.

With the above similarity function \( \psi \), the tracks and the detection responses in current frame are compared. The appearance similarity \( \rho \) is calculated with the JHCHs of the tracked object and the detection response in current frame. The spatial location similarity \( \psi \) is derived with the track's predicted position produced by Kalman Filter and the location of current detection response.

In our work, if a track has no similar response in current frame, the track is updated with the predicted state produced by Kalman Filter. The responses which have no matching tracks will be regarded as new objects, and tracks which have no matching responses for a predefined period of time will be terminated.

3. Experiments and discussions

In order to evaluate the effectiveness and efficiency of the proposed system, we have tested it on three challenging RGB-D datasets captured with Kinect at 640 × 480 resolution.

Office dataset: We used the dataset captured by Choi et al. [3] in an office environment. This publicly available dataset contains 17 video sequences of 2–3 min. long each. People in this dataset take various poses, such as standing and sitting on a chair, and are subject to a large degree of occlusions.

Mobile camera dataset: The mobile camera dataset provided by Choi et al. [3] has also been used for our evaluation. This dataset containing 18 sequences was collected with a Kinect mounted on a mobile platform (a PR2 robot). The robot was driven around in a building. This dataset includes various illumination conditions and cluttered backgrounds.

Clothing store dataset: We have also created our own dataset captured in a clothing store (consisting of two video sequences of about 45 min. long each) to make the evaluation more comprehensive. The challenges of this dataset include complicated and dynamic background, and people taking a variety of poses, walking in groups and frequently interacting with each other. This dataset is available at http://www.cv.fudan.edu.cn/humandetection.htm.

3.1. Implementation details

The ground plane location needed for constructing PEI is obtained by applying RANSAC [24] to the point cloud, which is an iterative method to estimate plane parameters from a set of 3D points that may contain outliers. RANSAC selects three points randomly at first, and calculates the parameters of the corresponding plane. It then searches for all points which belong to this plane with a given threshold in the point cloud. Afterwards, these procedures are repeated for a predefined times. At each iteration, the obtained result is compared with the last saved one, and if the new result is better, then the saved result will be replaced by the new one [29]. To locate the ground plane accurately and quickly, we combine RANSAC with the prior knowledge that the ground plane often has relatively large area and is often at the bottom of the point cloud. For fixed camera datasets (office and clothing store datasets), the ground plane locating procedure is performed once at the very beginning, while for mobile camera dataset, the ground plane is detected every frame. To improve efficiency, the parameters of the ground plane obtained in the previous frames are used as an initial estimation of the subsequent frames based on the knowledge that the camera moves smoothly in consecutive frames.

Due to lack of publicly available datasets that are comprehensive and appropriate for learning with our features (HOHD and JHCH), several RGB-D sequences were captured in various environments, including laboratory, passageway, lounge and store, by the authors. In all, 3556 frames from the sequences are selected as sample frames for training. The first detection phase (physiquewise plausible candidates localization) is performed on these frames, then a total of 29,691 positions are automatically produced and recorded as plausible candidates. These candidates are then manually labeled as human or non-human targets. Features (HOHD and JHCH) are extracted on these candidates for learning. A total of 5451 positive (human) samples and 24,240 negative (nonhuman) samples are labeled. The training samples contain more than 100 different people taking various poses, such as standing, walking, sitting, bowing, etc. The support vector machine (SVM) with a linear kernel is used as our classifier, and the feature vector is the linear combination of the two normalized histograms (HOHD and JHCH) with equal weights. In our training process, the human samples whose heads are wholly invisible in the images are removed from the training set, as both HOHD and JHCH are based on human upper body (including the head), and these samples without information of head will affect the performance of the classifier.

The samples are also utilized to obtain parameters for the proposed system. In our implementation, we calculate the average torso width, average head size, maximum and minimum height values of the labeled human samples to get the parameters \( \omega, \lambda, h_{\text{max}} \) and \( h_{\text{min}} \).

3.2. Results

3.2.1. Detection performance

We compare the proposed system against a conventional HOG detector [4], a depth-based detector proposed by Xia et al. [28], and a recent detector proposed by Choi et al. [3]. Fig. 5(a)–(c) shows the false positive per image (FPP) vs. miss rate curves for the four approaches. As the source code of Choi et al.’s approach is not available, its performance is only evaluated on office and mobile camera datasets (using authors’ reported performance). Four images per second from office and mobile camera datasets and one image every three seconds from clothing store dataset are used to evaluate the performance. In the evaluation frames, the upper bodies of people are hand-annotated with bounding boxes, and 3D locations are inferred from the depth images and bounding boxes. Two evaluation criteria from Choi et al. [3] are used to determine a positive detection. The first is based on overlapping degree between the detected regions and the ground truth bounding boxes. The second is based on 3D distance.

The experimental results show that the proposed detector significantly outperforms the other three methods. The performance of the HOG detector for detecting persons in RGB image is limited due to the high complexity of background and people's poses. Xia et al.’s approach utilizes a 2D head contour and a 3D head surface model to match heads in depth images. As in our test scenes, partial occlusion of head occurs frequently, consequently, its recall rate is limited. In the approach proposed by Choi et al., multi-cue (including depth template) detection is utilized, but the 3D information is not fully exploited and the depth-based upper body template encounters difficulties when depth data of human body is partially acquired. The proposed two-phase human detector combines RGB and depth data and employs the 3D shape and appearance information jointly, and performs well in case partial occlusion occurs.

Next, we evaluate the contribution of the first detection phase in our proposed detector, with results in Table 1. The results show that this phase produces a low miss rate yet greatly reduces the search space for the detector, as only several false positives are contained in the responses.
The missing detection in this phase is mainly caused by depth artifacts (depth data loss). All three datasets contain some frames in which people are very far from the camera, and depth data of these people is almost totally missing. People in this case are missed by the first detection phase.

Individual contribution of the two features used in the proposed detector is also quantitatively evaluated. We turn off one feature at a time and compare the detection results on the three datasets. As illustrated in Fig. 5(d)–(f), turning off the JHCH feature causes higher performance decrease than the HOHD feature. This can be explained as the JHCH is a joint combination of appearance and depth information, and provides more discrimination information, while HOHD is just collected from the depth data.

Finally, we evaluate the effect of object distance from camera on detection accuracy, as for more distant objects, the depth data is more noisy and incomplete. We respectively analyze the miss rates for short distance (< 4 m) and long distance (> 4 m), as depicted in Fig. 5(g)–(i). The results show that the proposed method yields higher detection accuracy for more near people.

3.2.2. Tracking performance

We consider two types of tracking errors [14], i.e., track lost error (fail to re-associate a track after occlusion) and ID switch error (swapping identities between two tracked persons), while false positives and false negatives have been covered in the evaluation of detection performance. As the robot is moving in the mobile camera dataset, people are tracked in local coordinate system of the robot in this case.

We compare the performance of the 3D JHCH for data association with that of independent color or height histogram by using the 42,985 frames from office dataset, the first 18,529 frames from mobile camera dataset, and the first 29,802 frames from clothing store dataset. The comparison results of tracking error counts are shown in Table 2. We can see that by constructing 3D joint histogram of...
color and height, the tracking errors (especially the ID switch errors) is significantly reduced.

Fig. 6 shows several visualized experimental results of the proposed system on the three datasets.

### 3.2.3. Computational efficiency

The proposed system runs on a desktop PC with a quad-core Intel i5-2500 CPU, 8 GBs RAM. With this configuration, the proposed human detection and tracking system (C++ implementation) achieves real-time performance (30–50 fps, without GPU acceleration).

### 3.3. Discussions about adapting to multi-camera system

In the above experiments, the proposed approach is evaluated with single camera setup. We also consider adapting our method to multi-camera system which can achieve a broader view field. We consider setting up the system as illustrated in Fig. 7. Such a setup provides a joint view field [20], and meanwhile alleviates mutual interference caused by overlapping infrared patterns projected by different RGB-D cameras (Kinects).

As people are detected and tracked in 3D point cloud in our method, the fusion of point clouds produced by multiple calibrated cameras can be considered as a preliminary stage, then human subjects can been detected and tracked in the resultant fused point cloud.

### 4. Conclusions

We have presented in this paper a human detection and tracking system that uses a single RGB-D camera and can cope with complex and dynamic environments. The experimental results show that the proposed method can effectively detect and track people in various poses and appearances by integrating the RGB and depth data, and can provide accurate body contours when 3D points are re-projected.

---

Table 2

<table>
<thead>
<tr>
<th>Office 42985 frames</th>
<th>Mobile camera 18529 frames</th>
<th>Clothing store 29802 frames</th>
</tr>
</thead>
<tbody>
<tr>
<td>co</td>
<td>ho</td>
<td>ch</td>
</tr>
<tr>
<td>#TL 87</td>
<td>93 65</td>
<td>14 21 10</td>
</tr>
<tr>
<td>#IDS 7</td>
<td>22 0</td>
<td>6 11 0</td>
</tr>
<tr>
<td>#Total 94</td>
<td>115 65</td>
<td>20 32 10</td>
</tr>
</tbody>
</table>

TL: track lost; IDS: ID switch; co: color only; ho: height only; ch: JHCH.
to original image domain. The three stages of the system accomplish the tasks incrementally, and the PEI representation greatly facilitates the computation in the detection and tracking stages.

Acknowledgments

This research work is supported by National Natural Science Foundation of China (Grant No. 61175036). We also would like to thank Ying Cui, Xi En Cheng and Shuohong Wang for many helpful discussions.

Supplementary Material

Supplementary material associated with this article can be found, in the online version, at 10.1016/j.patrec.2014.09.013.

References