

Accepted Manuscript

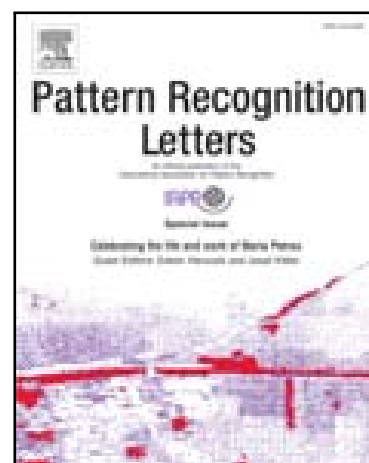
Detecting and tracking people in real time with RGB-D camera

Jun Liu, Ye Liu, Guyue Zhang, Peiru Zhu, Yan Qiu Chen

PII: S0167-8655(14)00302-X
DOI: [10.1016/j.patrec.2014.09.013](https://doi.org/10.1016/j.patrec.2014.09.013)
Reference: PATREC 6096

To appear in: *Pattern Recognition Letters*

Received date: 14 May 2014
Accepted date: 24 September 2014



Please cite this article as: Jun Liu, Ye Liu, Guyue Zhang, Peiru Zhu, Yan Qiu Chen, Detecting and tracking people in real time with RGB-D camera, *Pattern Recognition Letters* (2014), doi: [10.1016/j.patrec.2014.09.013](https://doi.org/10.1016/j.patrec.2014.09.013)

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Please save a copy of this file, complete and upload as the “Confirmation of Authorship” file.

As corresponding author I, Yan Qiu Chen, hereby confirm on behalf of all authors that:

1. This manuscript, or a large part of it, has not been published, was not, and is not being submitted to any other journal.
2. If presented at or submitted to or published at a conference(s), the conference(s) is (are) identified and substantial justification for re-publication is presented below. A copy of conference paper(s) is(are) uploaded with the manuscript.
3. If the manuscript appears as a preprint anywhere on the web, e.g. arXiv, etc., it is identified below. The preprint should include a statement that the paper is under consideration at Pattern Recognition Letters.
4. All text and graphics, except for those marked with sources, are original works of the authors, and all necessary permissions for publication were secured prior to submission of the manuscript.
5. All authors each made a significant contribution to the research reported and have read and approved the submitted manuscript.

Signature Yan Qiu Chen Date 24, Aug, 2014

List any pre-prints:

Relevant Conference publication(s) (submitted, accepted, or published):

J. Liu, Y. Liu, Y. Cui, and Y. Q. Chen, Real-time human detection and tracking in complex environments using single GRBD camera, in IEEE International Conference on Image Processing (ICIP), Sep. 2013. (Published)

Justification for re-publication:

This paper is an extension of the conference paper presented at ICIP 2013, and we have made a lot of improvements in this manuscript:

1. Additional technical details:
 - (a) More implementation details are included.
 - (b) The approach to locating the ground plane are added.
 - (c) A discussion about adapting our method to multi-camera system is added.
2. More experiments:
 - (a) The proposed system is evaluated with more RGB-D video sequences captured by the authors.
 - (b) The experiments performed on two other public datasets (including a mobile camera dataset and an office dataset) are also added to evaluate the robustness and general applicability of the system more systematically.
 - (c) Contribution analysis of each detection feature proposed by us is added.
 - (d) The accuracy of the first detection phase in our method is also quantitatively analyzed.
 - (e) The effect of distance on detection accuracy is evaluated.
 - (f) The performance of the proposed method is compared against a more recent approach.
 - (g) More training samples are utilized to improve the performance of the detector.
3. A clearer explanation of the motivation of the proposed method:

The motivation and contribution of our method are explained more clearly in the introduction section and method section.

As we have made several extensions and significant improvements, we consider this paper appropriate for publication in Pattern Recognition Letters.

Research Highlights (Required)

To create your highlights, please type the highlights against each \item command.

It should be short collection of bullet points that convey the core findings of the article. It should include 3 to 5 bullet points (maximum 85 characters, including spaces, per bullet point.)

- A PEI representation is proposed to alleviate overlapping in original image domain while preserving information of all pixels.
- A human plausible candidates locating technique is proposed to quickly reduce search space of the detector.
- Two novel features are proposed to characterize human shape and appearance in 3D space.
- A single-pass, progressive refinement framework enables the system to achieve high accuracy at real time.



Pattern Recognition Letters
journal homepage: www.elsevier.com

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,**}

^aSchool of Computer Science, Fudan University, Shanghai 201203, China

^bCollege 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 physique-wise 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 Ltd. All rights reserved.

1. Introduction

Detecting and tracking people in video sequences is a key research problem that has wide application in security surveillance (Zhou and Hoang, 2005), people collective behavior study, etc, and has attracted a great deal of research attention (Fablet and Black, 2002; Fleuret et al., 2008; Zhang et al., 2008; Liu et al., 2009; Wang et al., 2012). 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 (Perez et al., 2002; Moore, 2003; Zhu et al., 2006) 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 methods involve extracting spatially global features (Felzenszwalb, 2001) and some others use statistical learning with local features and boosting, such as EOH (Levi and Weiss, 2004), HOG (Dalal and Triggs, 2005) and edgelet

(Wu and Nevatia, 2007). 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 (Yous et al., 2008; Ess et al., 2009; Enzweiler et al., 2010; Haritaoglu et al., 2013) using depth data (or disparity image) produced by stereo rigs are proposed. The method proposed by Mitzel and Leibe (2012) 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. (2007) 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

^{**}Corresponding author.

e-mail: ljchangyu@hotmail.com (Jun Liu), yeliu@fudan.edu.cn (Ye Liu), guyuezhong13@fudan.edu.cn (Guyue Zhang), peiruzhu@gmail.com (Peiru Zhu), chenyq@fudan.edu.cn (Yan Qiu Chen)

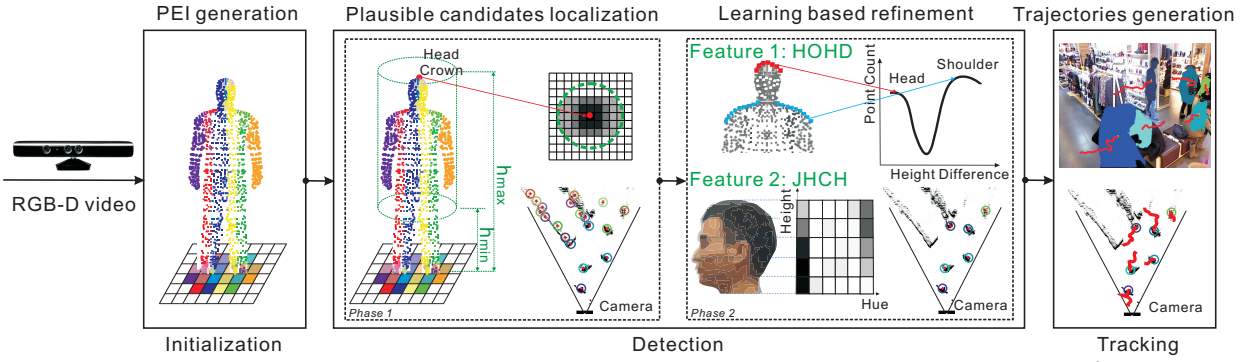


Fig. 1. Major stages of proposed system.

processing is difficult to achieve (Ikemura and Fujiyoshi, 2010).

More recently, depth cameras such as Kinect (Han et al., 2013), 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 (Luber et al., 2011; Zhang et al., 2013) 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 (Dan et al., 2012; Bevilacqua et al., 2006) or horizontally at the same level as humans (Salas and Tomasi, 2011; Han et al., 2012). 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 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.

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 (Liu et al., 2013).

The contributions of this paper include: (1) A PEI representation is proposed to alleviate overlapping in the original image domain while preserving information of all pixels. (2) A human plausible candidates locating technique is proposed to quickly reduce search space to avoid the need by the subsequent pro-

cedures to scan the entire point cloud. (3) Two novel features, HOHD and JHCH, are proposed to effectively and efficiently characterize human shape and appearance in 3D space. (4) A single-pass, progressive refinement framework enables the system to achieve real-time performance as increasingly fewer candidate positions need to be examined by latter stages.

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 Point Ensemble Image (PEI) representation to overcome segmentation difficulties encountered by directly using RGB-D image. We have also developed a multi-stage 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 Joint Histogram of Color and Height (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 establish a new coordinate system with the ground as its XY plane and a vertical upper vector as the Z axis, and discretize the ground plane using a square grid. The point cloud, mapped to the new coordinate system, is projected to XY plane and generates the plan-view on the floor. With this virtual view, segmentation can be dramatically simplified in many cases.

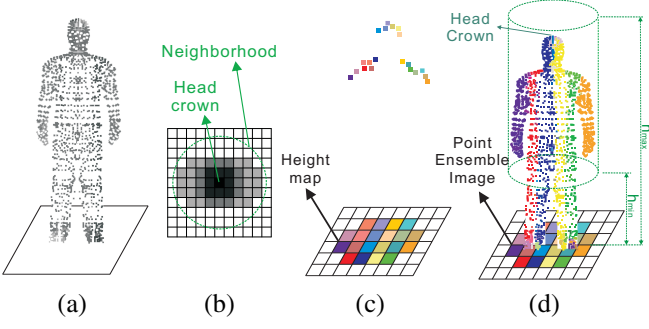


Fig. 2. (a) 3D point cloud of a person. (b) The point cloud in plan-view. (c) In height map representation, only the upper portion of the body is kept well. (d) In the proposed PEI, all of the pixels (points) from original RGB-D image (3D point cloud) are maintained. To differentiate, the cells in the 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.

The square grid for discretizing the plan-view are defined as $g^{i,j} : \begin{cases} (i-1)\varepsilon \leq x < i\varepsilon \\ (j-1)\varepsilon \leq y < j\varepsilon \end{cases}$, where $g^{i,j}$ is the (i, j) indexed cell of the grid, (x, y) is the point coordinates on XY plane, and ε is the grid resolution (a smaller ε value may help represent more shape details, but may increase calculation time. Besides, the limitations of the camera's depth accuracy should be considered. Typically, 1-4 cm is suitable).

In a height map representation (Bahadori et al., 2007), 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 (Yous et al., 2008). We propose a novel Point Ensemble Image (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:

$$E^{i,j} = \{p \mid p \in P, (x^p, y^p) \in g^{i,j}\}, \quad (1)$$

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 (Xia et al., 2011; Dan et al., 2012) 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 physique-wise 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 physique-wise 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 $\omega/2$ (ω 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:

$$\mathcal{N} = \{x \mid x \in c', c' \in E, |c' - c| < \omega/2\}, \quad (2)$$

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 $\mathcal{N}$, which are collected from PEI. As we just focus on body's upper part here, only a subset of $\mathcal{N}$ 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;

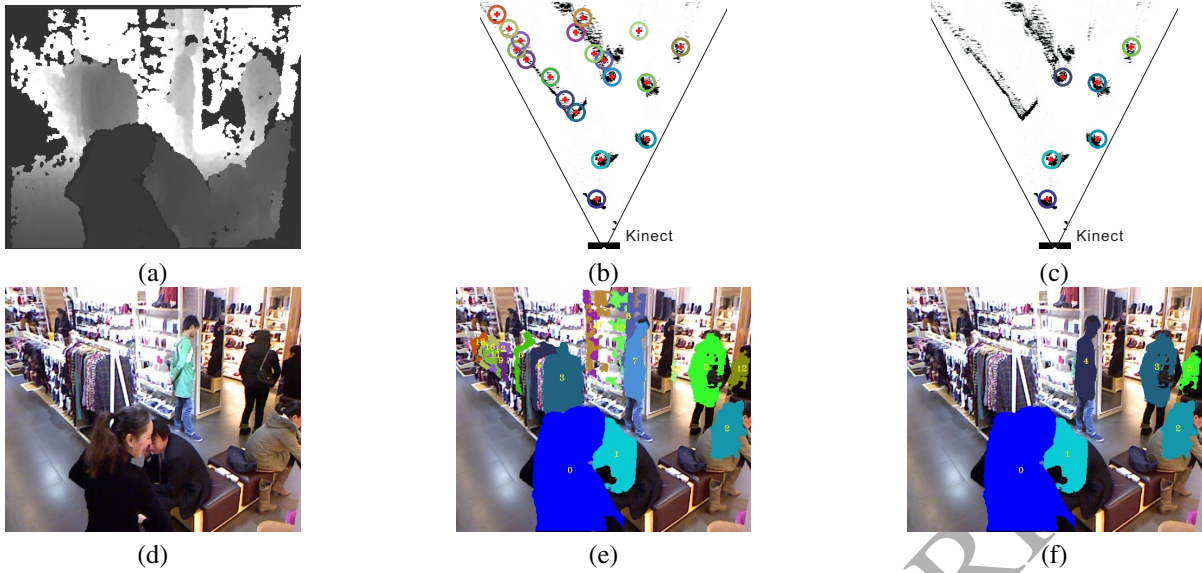


Fig. 3. Detection process. The left column shows the original depth and RGB images. The middle and right columns respectively illustrate the detection results of Phase 1 and Phase 2. Figures (b) and (c) show the detection results from plan-view perspective, and the circle centers indicate the detected head crown positions. In (e) and (f), the points in the cylindrical neighborhood of a detected head crown in 3D space are considered to belong to the same object and are given the same color mark.

b) The height difference between the point and the head crown is less than a threshold (η , 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 (Xia et al., 2011), thus this feature is effective on noisy and even fragmentary data.

Joint Histogram of Color and Height (JHCH):

Color distribution (Perez et al., 2002) and height (or depth) pattern (Choi et al., 2013) have proved to be effective in human detection and tracking, and simple linear combination of the two has been exploited (Dan et al., 2012). In order to utilize the correlated information between them to achieve higher performance, we propose a Joint Histogram of Color and Height (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. The white and black pixels are considered specially and are recorded in extra bins. Head crown's neighboring points whose height values are less than λ (the approximate length of human head) smaller than that of head crown are collected to build the 2D JHCH.

By describing color and height statistically, JHCH is able to work well even in dealing with people in different hairstyles and head poses.

2.3. People tracking

The objects surviving the detection stage are passed to the tracking module. As the proposed human detection method can

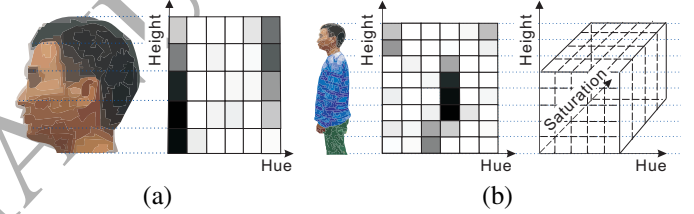


Fig. 4. (a) 2D JHCH of human head. With the increment of height, the probability of the occurrence of skin color decreases, while the color of hair may be the opposite. This can be considered as a distinct feature of human head. (b) 2D and 3D JHCHs of the body. JHCH, which encodes information about the positions of colors on the object, is an effective descriptor of a human. In this figure, probability of each bin is expressed with a grayscale value (a darker bin indicates a higher probability).

produce relatively clean responses, human tracking becomes more manageable. A simple yet effective and also efficient tracking scheme is adopted. We use a basic Kalman Filter model and a greedy data association technique in which each track is assigned a response in current frame which is most similar to it and closest to the predicted position of Kalman Filter. The challenge is to define a similarity function that is able to better reflect the differences among different humans.

Generally, different people may have different color-height distributions. To automatically track unknown number of people, the JHCH descriptor presented in Section 2.2 is utilized here again. JHCH is superior to color histogram or height model, or a simple combination of both in distinguishing a person from others, in that JHCH better explores the relation between color and height (it records the positions of colors on the object, thus provides more discrimination information). 2D JHCH pattern is sufficient and effective in describing the appearance of human head, but in characterizing entire human bodies, the effectiveness will be limited due to the great vari-

eties of clothes color. So we adopt the 3D hue-saturation-height histogram to characterize a person, as shown in Fig. 4(b). We divide the hue range 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 $H_t = \{H_t^k\}_{k=1,2,\dots,r}$, 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:

$$w(p_i) = \exp\{\alpha \cdot h(p_i)\}, \quad (3)$$

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

$$H_t^k = f \cdot \sum_{p_i \in N} \left\{ w(p_i) \cdot \delta[\mathcal{B}(p_i) - k] \right\}, \quad (4)$$

where δ is Kronecker delta function, $\mathcal{B}(p_i)$ indicates the bin of JHCH that p_i should be assigned to, N is the neighboring point set collected from PEI, and $f = 1 / \sum_{p_i \in N} w(p_i)$ is the normalization factor ensuring $\sum_{k=1}^r H_t^k = 1$.

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

$$\zeta(a, b) = \gamma \cdot \rho(H_a, H_b) + (1 - \gamma) \cdot \psi(L_a, L_b), \quad (5)$$

where L_i is the spatial location of object i ($i = a, b$), $\rho(H_a, H_b)$ is the Bhattacharyya similarity (Kailath, 1967) between the JHCHs of objects a and b , $\psi(L_a, L_b)$ represents the spatial location similarity between the two objects, and γ is a weighting factor to control the relative importance of the appearance similarity and 3D location similarity. $\rho(H_a, H_b)$ and $\psi(L_a, L_b)$ can be calculated as:

$$\rho(H_a, H_b) = \exp\left\{ -\mu \cdot \left(1 - \sum_{k=1}^r \sqrt{H_a^k H_b^k} \right)^2 \right\}, \quad (6)$$

$$\psi(L_a, L_b) = \exp\left\{ -\nu \cdot D(L_a, L_b) \right\}, \quad (7)$$

where $D(L_a, L_b)$ is the Euclidean distance between a and b , and both μ and ν are weighting factors. The similarity value $\zeta(a, b)$ is between 0 and 1, and $\zeta(a, b) = 1$ indicates that the two objects are perfectly matched with each other.

With the above similarity function ζ , the tracks and the detection responses in current frame are compared. The appearance similarity ρ is calculated with the JHCHs of the tracked object and the detection response in current frame. The spatial location similarity ψ 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. (2013) in an office environment. This publicly available dataset contains 17 video sequences of 2-3 minutes 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. (2013) 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 minutes 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 (Rusu et al., 2009) 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 (Yang and Forstner, 2010). 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 dataset that is 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

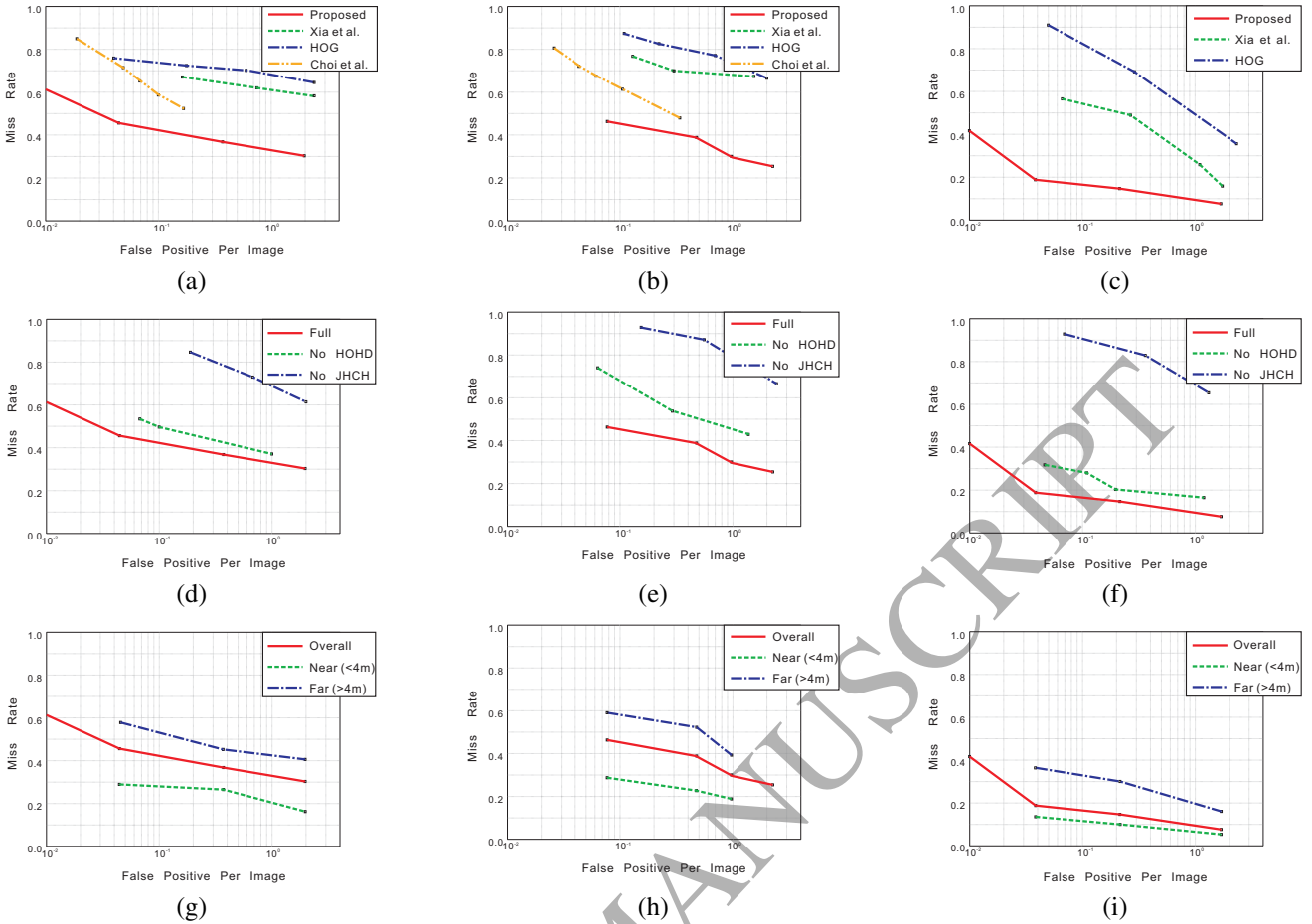


Fig. 5. FPPI vs. miss rate curves. Left column: results on office dataset. Middle column: results on mobile camera dataset. Right column: results on clothing store dataset. The top row shows comparison results with other approaches, and the middle row gives the contribution analysis of each detection feature. The bottom row depicts the detection miss rates for different distance ranges. In (d) - (f), the “Full” curve means using both features and other curves show the results when a specific feature is removed.

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 29691 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 24240 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 ω , λ , h_{max} and h_{min} .

3.2. Results

Detection performance: We compare the proposed system against a conventional HOG detector (Dalal and Triggs, 2005), a depth-based detector proposed by Xia et al. (2011), and a recent detector proposed by Choi et al. (2013). Fig. 5(a)-(c) shows the False Positive Per Image (FPPI) 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., 2013) 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.

Table 1. Miss rate and FPPI in first detection phase

Dataset	Office	Mobile camera	Clothing store
Miss rate	0.06	0.08	0.05
FPPI	27	10	12

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 ($< 4m$) and long distance ($> 4m$), as depicted in Fig 5(g)-(i). The results show that the proposed method yields higher detection accuracy for more near people.

Tracking performance: We consider two types of tracking errors (Harville, 2005), 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 42985 frames from office dataset, the first 18529

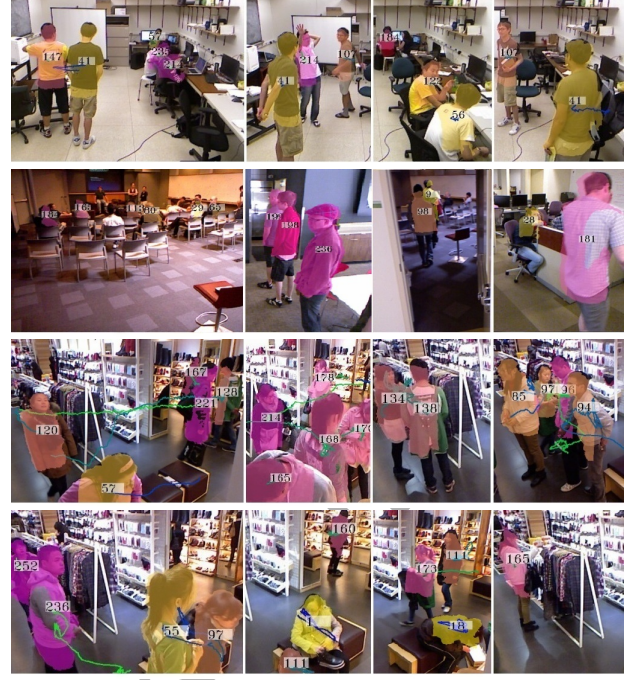


Fig. 6. Examples of detection and tracking results on office dataset (top row), mobile camera dataset (second row) and clothing store dataset (bottom two rows).

frames from mobile camera dataset, and the first 29802 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.

Table 2. Tracking error counts for three association methods

	Office 42985 frames			Mobile camera 18529 frames			Clothing store 29802 frames		
	co	ho	ch	co	ho	ch	co	ho	ch
#TL	87	93	65	14	21	10	15	14	13
#IDS	7	22	0	6	11	0	6	15	1
#Total	94	115	65	20	32	10	21	29	14

TL: track lost. IDS: ID switch.

co: color only. ho: height only. ch: JHCH.

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

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 filed (Luber et al., 2011), and meanwhile alleviates mutual interference caused

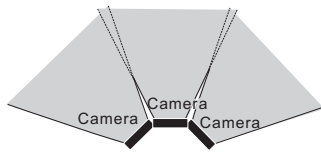


Fig. 7. Setup of multi-camera system offering a joint view field.

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 be 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 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.

References

Bahadori, S., Iocchi, L., Leone, G., Nardi, D., 2007. Real-time people localization and tracking through fixed stereo vision. *Applied Intelligence* 26, 83–97.

Bevilacqua, A., Stefano, L.D., Azzari, P., 2006. People tracking using a time-of-flight depth sensor, in: *IEEE International Conference on Video and Signal Based Surveillance*, pp. 89–93.

Choi, W., Pantofaru, C., Savarese, S., 2013. A general framework for tracking multiple people from a moving camera. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35, 1577–1591.

Dalal, N., Triggs, B., 2005. Histograms of oriented gradients for human detection, in: *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 886–893.

Dan, B.K., Kim, Y.S., Suryanto, Jung, J.Y., Ko, S.J., 2012. Robust people counting system based on sensor fusion. *IEEE Transactions on Consumer Electronics* 58, 1013–1021.

Enzweiler, M., Eigenstetter, A., Schiele, B., Gavrila, D.M., 2010. Multi-cue pedestrian classification with partial occlusion handling, in: *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 990–997.

Ess, A., Leibe, B., Schindler, K., Gool, L.V., 2009. Robust multiperson tracking from a mobile platform. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31, 1831–1846.

Fablet, R., Black, M.J., 2002. Automatic detection and tracking of human motion with a view-based representation, in: *European Conference on Computer Vision*, pp. 476–491.

Felzenszwalb, P., 2001. Learning models for object recognition, in: *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1056–1062.

Fleuret, F., Berclaz, J., Lengagne, R., Fua, P., 2008. Multicamera people tracking with a probabilistic occupancy map. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 30, 267–282.

Han, J., Pauwels, E.J., de Zeeuw, P.M., de With, P.H.N., 2012. Employing a rgb-d sensor for real-time tracking of humans across multiple re-entries in a smart environment. *IEEE Transactions on Consumer Electronics* 58, 255–263.

Han, J., Shao, L., Xu, D., Shotton, J., 2013. Enhanced computer vision with microsoft kinect sensor: A review. *IEEE Transactions on Cybernetics* 43, 1318–1334.

Haritaoglu, I., Flickner, M., Beymer, D., 2013. Video-crm: Understanding customer behaviors in stores, in: *SPIE Proceedings, Video Surveillance and Transportation Imaging Applications*.

Harville, M., 2005. Stereo person tracking with short and long term plan-view appearance models of shape and color, in: *IEEE Conference on Advanced Video and Signal Based Surveillance*, pp. 522–527.

Ikemura, S., Fujiyoshi, H., 2010. Real-time human detection using relational depth similarity features, in: *Asian Conference on Computer Vision*, pp. 25–38.

Kailath, T., 1967. The divergence and bhattacharyya distance measures in signal selection. *IEEE Transactions on Communication Technology* 15, 52–60.

Levi, K., Weiss, Y., 2004. Learning object detection from a small number of examples: the importance of good features, in: *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 53–60.

Liu, J., Liu, Y., Cui, Y., Chen, Y.Q., 2013. Real-time human detection and tracking in complex environments using single rgb-d camera, in: *IEEE International Conference on Image Processing*, pp. 3088–3092.

Liu, Y., Chen, X., Yao, H., Cui, X., Liu, C., Gao, W., 2009. Contour-motion feature (cmf): A space-time approach for robust pedestrian detection. *Pattern Recognition Letters* 30, 148–156.

Luber, M., Spinello, L., Arras, K.O., 2011. People tracking in rgb-d data with on-line boosted target models, in: *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 3844–3849.

Mitzel, D., Leibe, B., 2012. Close-range human detection for head-mounted cameras, in: *British Machine Vision Conference*, pp. 1–11.

Moore, D., 2003. A real-world system for human motion detection and tracking. Ph.D Thesis, California Institute of Technology.

Perez, P., Hue, C., Vermaak, J., Gangnet, M., 2002. Color-based probabilistic tracking, in: *European Conference on Computer Vision*, pp. 661–675.

Rusu, R.B., Sundareshan, A., Morisset, B., Hauser, K., Agrawal, M., Latombe, J.C., Beetz, M., 2009. Leaving flatland: Efficient real-time three-dimensional perception and motion planning. *Journal of Field Robotics* 26, 841–862.

Salas, J., Tomasi, C., 2011. People detection using color and depth images, in: *Mexican Conference on Pattern Recognition*, pp. 127–135.

Wang, J., Chen, D., Chen, H., Yang, J., 2012. On pedestrian detection and tracking in infrared videos. *Pattern Recognition Letters* 33, 775–785.

Wu, B., Nevatia, R., 2007. Detection and tracking of multiple, partially occluded humans by bayesian combination of edgelet based part detectors. *International Journal of Computer Vision* 75, 247–266.

Xia, L., Chen, C.C., Aggarwal, J.K., 2011. Human detection using depth information by kinect, in: *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 15–22.

Yang, M.Y., Forstner, W., 2010. Plane detection in point cloud data. Technical Report, Institute of Geodesy and Geoinformation, University of Bonn.

Yous, S., Laga, H., Chihara, K., 2008. People detection and tracking with world-z map from a single stereo camera, in: *The Eighth International Workshop on Visual Surveillance*.

Zhang, H., Reardon, C., Parker, L.E., 2013. Real-time multiple human perception with color-depth cameras on a mobile robot. *IEEE Transactions on Cybernetics* 43, 1429–1441.

Zhang, Z., Venetianer, P.L., Lipton, A.J., 2008. A robust human detection and tracking system using a human-model-based camera calibration, in: *The Eighth International Workshop on Visual Surveillance*.

Zhou, J., Hoang, J., 2005. Real time robust human detection and tracking system, in: *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 149–159.

Zhu, Q., Yeh, M.C., Cheng, K.T., Avidan, S., 2006. Fast human detection using a cascade of histograms of oriented gradients, in: *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1491–1498.