REAL-TIME HUMAN DETECTION AND TRACKING IN COMPLEX ENVIRONMENTS USING SINGLE RGBD CAMERA

Jun Liu, Ye Liu, Ying Cui, and Yan Qiu Chen

School of Computer Science, Fudan University, Shanghai, China
{ljun, yeliu, cuiying, chenyq}@fudan.edu.cn

ABSTRACT

This paper presents a new approach to real-time human detection and tracking in cluttered and dynamic environments by integration of RGB and depth data. We introduce the notion of Point Ensemble Image, which fully encodes both RGB and depth information from a virtual plan-view perspective, and we reveal that human detection and tracking in 3D space can be performed very effectively based on this new representation. Our human detector is able to take advantage of depth data by effectively locate physically plausible candidates as a first step, and then both depth and color information is made full use of in a supervised learning manner at the second stage. 3D trajectories of humans are finally generated by data association in which joint statistics of color and height are computed and compared. Experimental results show that the system is able to work satisfactorily in complex real-world situations.

Index Terms— Human detection, Tracking, RGBD

1. INTRODUCTION

Detection and tracking of human beings in video sequences have wide application in security surveillance [1], human collective behavior study [2], etc. and have attracted a lot of research attention [3][4]. The problem (of detecting and tracking people in complex environments) remains largely open due to many serious challenges, such as occlusion, change of appearance, complex and dynamic background, etc.

Early work [5][6][7][8] used conventional video cameras capturing appearance image sequences 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 but would dramatically deteriorate in more challenging real-world applications. Recently, depth cameras such as Kinect and TOF cameras have become widely available at affordable prices. Studies employing depth cameras [9][10][11] have demonstrated the great value of the depth information for detecting and tracking human beings. In existing works, depth cameras are often placed either vertically overhead [9][12] or at horizontally the same level as humans [10][13]. Occlusion is significantly reduced in the former case, details of the human body however can't be well captured. Advantages and shortcomings in the latter case are just the opposite, human body can be observed more completely but severe occlusions may happen frequently. A oblique view may be a good trade-off between completeness and occlusion. Even with this set-up, 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.

In this paper, we present a real-time system we have developed that is able to detect and track humans in oblique view RGBD videos. The proposed method is three-stage structured. In the first stage, an unsupervised detector is used to retrieve positions that are physically plausible. Then at the second stage, the candidate positions are further refined by a classifier which has been trained off-line with features extracted from multiple cues. Finally, data association is carried out to the detection responses, which generates trajectories. This three-stage structure allows very fast detection and tracking, and yields good performance in a real-world clothing store scene, in which both the human activities and background are quite complex.

2. METHOD

The system presented in this paper consists of three main stages, as shown in Fig. 1. Before these stages, the original RGBD data is transformed to Point Ensemble Image. In this new representation, we use a two-stage method to detect humans. Finally, the trajectories are produced by an effective color-height joint statistics and association method.

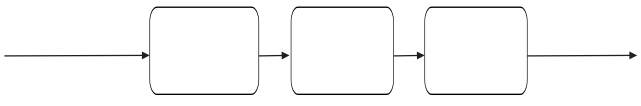


Fig. 1. Overview of the proposed system.

The research work presented in this paper is supported by National Natural Science Foundation of China, Grant No. 61175036.

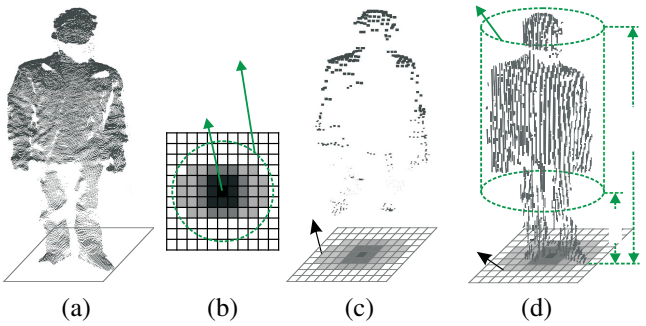


Fig. 2. (a) 3D point cloud of a person. (b) Plan-view of the person. (c) In height map representation, only the upper portion of the body is kept well. (d) In Point Ensemble Image representation, all of the pixels (points) from original RGBD image (3D point cloud) are recorded. *(Note: The green parts of (b) and (d) show the neighborhood area of a point in Point Ensemble Image, more details will be given in Sec. 2.2).*

2.1. Point Ensemble Image Representation

Due to mutual occlusion and adjacency among humans, 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 on the original data, as shown in Fig. 2.

With depth information, pixels are back-projected into 3D space to construct a 3D point cloud in camera's coordinate. We can establish a new coordinate system with the ground as its XY plane and the vertical upper vector as the Z axis and discretize the ground plane into regular grids. The point cloud, mapped to the new system, is projected to XY plane and forms the plan-view on the floor. With this virtual view, overlapping is dramatically reduced.

In a height map representation [14][15], each cell of the plan-view keeps the highest one (the one with the maximum Z value) of the 3D points which are projected into that cell. Though interesting results have been achieved, much of the detail information is lost [16]. Thus a novel representation, called Point Ensemble Image, is proposed in this paper, which takes advantage of plan-view while preserving the information of all 3D points (see Fig. 2(d)).

In Point Ensemble Image (denoted by E), each cell records an ensemble of the points (a set of the points) which are projected into that cell, so it can be formulated as: $E_{i,j} = \{p \mid p \in P \wedge (p_x, p_y) \in g_{i,j}\}$, where P is the point cloud in the new coordinate system, $p = (p_x, p_y, p_z)$ is a 3D point of P, and $g_{i,j}$ is a cell of the grids.

2.2. Human Detection

Background subtraction [14][17] is widely used in human detection, but in real-world situations, such as supermarkets and clothing stores, it may not work well since background can hardly keep static. Though several depth camera based human detection methods [9][18] without background subtraction have been proposed, in many scenes, their performances will be limited due to mutual occlusions and depth artifacts (optical noise and data loss) , as the shape prior of head and body which is crucial in these methods can be easily corrupted in the case of occlusion and incomplete depth data.

In our work, we do not carry out detection by imposing strong shape prior on targets, instead we do it incrementally. Our two-stage detection method is able to leverage the advantages of depth and RGB data. In the first stage, most of the false positives that are human-physiquely implausible can be efficiently rejected, then the responses are further purified by a learning based classifier in the second stage (see Fig. 3).

Stage 1 - Physically plausible candidate localization.

As a preliminary step, we try to find the points which is higher than all other points in a neighborhood (local height maxima). For a point p in the point cloud, we draw an cylindrical neighborhood with radius $\omega/2$ (ω is the average width of human torso), as shown in Fig. 2. The height values of the points inside the cylinder are compared with that of p. This computation can be very efficient thanks to the Point Ensemble Image. The neighbor points of p can be calculated on the Point Ensemble Image (denoted by E) as:

$$\mathcal{N} = \{x \mid x \in c^{'} \ \wedge \ c^{'} \in E \ \wedge \ |c^{'} - c| < \omega/2\}, \quad (1)$$

where c is the cell of E that p is projected into, $c^{'}$ denotes the nearby cells of c. All the points with the maximum height value in their neighborhoods are regarded as crowns of human heads. To increase reliability and efficiency, only the points whose heights are between h_{max}(we use 0.6m) and h_{min}(we use 2m) are taken into account, which rejects a large amount of unreasonable points, such as those near floor and ceiling.

Even in highly crowded and occluded situations, this stage makes sure that real head-crown locations are contained in the resulted responses. Although they are far from perfect as they contain lots of false positives (see Fig. 3(c)), the space to be searched has been greatly reduced.

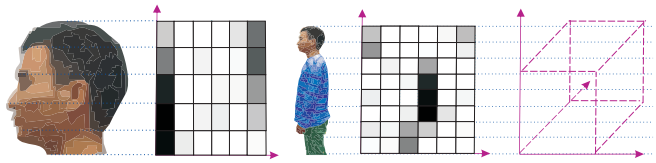
Stage 2 - Learning-based refinement.

We then try to determine whether the objects, produced by previous stage, are indeed humans. We learn a robust classifier with two sources of features, which encode shape and color information respectively. The features are as follows:

(1) Histogram of Height Difference (HOHD): The shape of human's upper part is almost distinct from other objects. This can be observed more clearly in plan-view (see Fig.2(b)). The height of the head-crown is larger than nearby points and significantly larger than shoulder areas. This kind of height variation can be described with height differences in a statistical manner. For a local height maximum, its height value is subtracted from those of the points in its cylindrical neighborhood. As we just focus on body's upper part, only a part of neighbor points are selected, which satisfy the following two

(a) (b) (c) (d)

Fig. 3. Detection process. (a)(b) Depth and RGB images. The scene is complex and some people are occluded. (c) Output of Stage 1. (d) Output of Stage 2. Points in the neighborhood of a local height maximum in 3D space are considered to belong to the same object and are given the same color mask.



(a) 2D JHCH of head (b) 2D and 3D JHCHs of body

Fig. 4. (a) 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) 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 gray-scale values (a darker value indicates a higher probability).

criteria: a) the point has the largest height value in the cell it is projected into; b) the height of the point is less than 0.45m smaller than head-crown's. Given the height differences, we construct a histogram, in which a range from 0 to 0.45m are divided into 15 bins. By constructing a histogram, we avoid making hard assumptions on human shape like [9][18], thus this feature is effective on noisy or even fragmentary data.

(2) Joint Histogram of Color and Height (JHCH): Both color distribution [7] and height (or depth) pattern [14][18] have proved to be effective in human detection and tracking. Simple linear combination of the two has also been exploited in [9]. In practice, color and height information can be employed much more effectively if correlation between them is considered. In our work, we propose a color-height joint histogram (we call it 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 is basically collected from hair and face or solo hair (if human is observed from the back) and are located at different height levels. Since the hue component of HSV color space is insensitive 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 neighbor points whose height values are less than 0.2m (approximately the size 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 when dealing with people with different hairstyles and head poses.

We captured several RGBD sequences of different scenes, including office, lounge and clothing store. In all, 3214 frames are selected as sample frames for training. Local height maxima are detected on these sample frames, and a total of 24997 positions are recorded which are then manually labeled as human or nonhuman targets. Features are extracted at these positions for classifier training. Averagely, about 2 persons and 6 other objects are extracted per frame. The training samples involve about 100 different people with various poses, such as standing, walking, sitting, bowing, etc.

The support vector machine (SVM) with a linear kernel is used as our classifier.

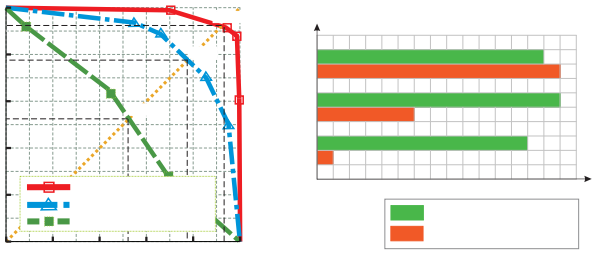
2.3. Human Tracking

Human tracking is facilitated if clean detection responses are given. We use a basic Kalman Filter model and a simple greedy data association technique, in which each track is assigned a response in current frame which is most similar to it and nears the predicted position of Kalman ($\leq$0.6m). The challenge is to define a similarity function that is able to better characterize the differences among different humans.

Generally, different people may have different color-height distributions. To automatically track unknown number of people, the JHCH model described in Sec. 2.2 is utilized here again. Color-height joint histogram is superior to color 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. 2D JHCH 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 varieties of clothes colors. So we adopt the 3D hue-saturation-height histogram to present a person (see Fig.4(b)). We divide the hue range into 9 intervals, saturation range into 5 intervals, and height range (0.6m-2m) 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 the 3D JHCH.

As different body parts have different stabilities in image (the upper part is more reliable as it is less likely to be occluded), different weights are assigned to the points at different height levels by employing a weighting function $w(x_i) = e^{\alpha * h(x_i)}$, where $h(x_i)$ is the height value (in meters) of point x_i, and α (we use 0.2) is a control factor. So the r-bins 3D JHCH $H_t = \{H_t^k\}_{k=1,2,\ldots,r}$ of target t can be calculated as:

$$H_t^k = f * \sum\nolimits_{x_i \in \mathcal{N}} \big(w(x_i) * \delta[b(x_i) - k]\big), \qquad (2)$$



(a) Detection performance (b) Tracking performance

Fig. 5. Comparison of performance.

where δ is Kronecker delta function, $b(x_i)$ means the bin that x_i should be assigned to, $\mathcal{N}$ is the neighbor points set, and $f = 1/\sum_{x_i \in \mathcal{N}} w(x_i)$ is the normalization factor. Then we can define the similarity function between targets a and b as:

$$\zeta = \gamma * \rho[H_a, H_b] + (1 - \gamma) * \varphi[L_a, L_b], \tag{3}$$

where L_i is the spatial-location of object i $(i = a, b)$, γ is the weighting factor, ρ and φ are the histograms' Bhattacharyya similarity and targets' spatial-location similarity respectively, which can be formulated as follows:

$$\rho[H_a, H_b] = exp\big(- \mu * \big(1 - \Sigma_{k=1}^{r} \sqrt{H_a^k H_b^k}\big)^2\big), \tag{4}$$

$$\varphi[L_a, L_b] = exp\big(- \upsilon * D(L_a, L_b)\big), \tag{5}$$

where D is the Euclidean distance between a and b, and both μ and υ are control factors (in our work,μ is 30 and υ is 0.01).

In the system, the similarities of detection responses and tracks are calculated. The responses which have no matching tracks will be regarded as new objects and tracks which have no matching responses for more than 10s will be terminated.

3. EXPERIMENTS

The experimental evaluation of the proposed system is performed using real-world RGBD videos captured with Kinect in a clothing store on a Friday evening. The Kinect was mounted at 2.2m height with a oblique angle of about 30 degrees with respect to the floor. The system runs on a desktop equipped with a quad-core Intel i5-2500 CPU, 8 GBs RAM. On this system, the average frame rate is 29Hz.

To evaluate the performance of the proposed detector, a total of 400 frames are uniformly sampled from a 45-minute sequence, which contains unscripted behaviors of customers and salesclerks in the store. The detection method performs well in this challenging dataset. And we have compared our algorithm with a conventional appearance-based detector HOG [5] and a recent depth-based method Model Fitting [18]. We plot the precision-recall curves of these different techniques and compare their BreakEven Points (BEP, precision equals recall), see Fig. 5(a). The precision is defined as $\frac{TP}{TP+FP}$ and recall is defined as $\frac{TP}{TP+FN}$. TP, FP, and FN are true positive, false positive and false negative respectively.

(a) (b) (c) (d)

Fig. 6. (a) Human tracking in complex scenes. (b) After short-time occlusion, the tracks are right re-attached. (c) People are crowded. (d) Boxes are passed between two persons.

The experimental results show that the proposed detector outperforms HOG and Model Fitting based approaches over the entire precision-recall range and obtains a BEP of 93%. Detecting persons only in RGB data, the HOG detector gets a relatively low BEP due to the high complexity of the background and people's actions. The Model Fitting method utilizes a 2D head contour and a 3D head surface model to match heads in depth image. As in our testing scene, partial occlusion of head occurs frequently and depth data captured by Kinect suffers from severe data missing (in Fig. 3(a), the depth data of the furthest man's head is incomplete), consequently, the performance of Model Fitting is limited. The proposed detector utilizes a two-stage human detector by combing RGB and depth data, which can also perform well when partial occlusion and depth data loss occur.

The tracking performance of our method is also quantitatively evaluated. In our evaluation, we only focus on two significant errors: track lost (fail to re-associate a track after occlusion) and ID switch (swapping identities of two tracked persons), as false positives and false negatives have been covered in the detection results. We compare 3D JHCH with independent color or height histogram in data association. The comparison results of tracking error counts are shown in Fig. 5(b). We can see by constructing a joint histogram of color and height, the number of tracking errors is reduced.

Fig. 6 shows some of the experimental results in the clothing store environment. It can be seen that in this highly complex and dynamic situation, our system performs well in human detection and tracking.

4. CONCLUSION

We present in this paper a human detection and tracking system that can work effectively and efficiently in real-world environments. The three stages of the system accomplish the tasks incrementally, and a novel representation (Point Ensemble Image) greatly facilitates the computation in these stages. The experimental results show that the system can run fast and robustly in real-world challenging situations.

5. REFERENCES

[1] X. Liu, P. H. Tu, J. Rittscher, A. Perera, and N. Krahnstoever, "Detecting and counting people in surveillance applications," in *IEEE Conference on Advanced Video and Signal Based Surveillance (AVSS)*, 2005.

[2] M. Butenuth, F. Burkert, A. Kneidl, A. Borrmann, et al., "Integrating pedestrian simulation, tracking and event detection for crowd analysis," in *IEEE ICCV Workshop on Modeling, Simulation and Visual Analysis of Large Crowds*, Nov. 2011.

[3] F. Fleuret, J. Berclaz, R. Lengagne, and P. Fua, "Multicamera people tracking with a probabilistic occupancy map," *IEEE TPAMI*, vol. 30, no. 2, pp. 267–282, Feb. 2008.

[4] D. Mitzel and B. Leibe, "Close-range human detection for head-mounted cameras," in *British Machine Vision Conference (BMVC)*, Sept. 2012.

[5] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *CVPR*, 2005.

[6] M. Han, A. Sethi, W. Hua, and Yihong Gong, "A detection-based multiple object tracking method," in *International Conference on Image Processing (ICIP)*, Oct. 2004.

[7] P. Perez, C. Hue, J. Vermaak, and M. Gangnet, "Color-based probabilistic tracking," in *European Conference on Computer Vision (ECCV)*, 2002.

[8] M. Andriluka, S. Roth, and B. Schiele, "People-tracking-by-detection and people-detection-by-tracking," in *CVPR*, 2008.

[9] B.-K. Dan, Y.-S. Kim, Suryanto, J.-Y. Jung, and S.-J. Ko, "Robust people counting system based on sensor fusion," *IEEE Trans. on Consumer Electron.*, vol. 58, no. 3, pp. 1013–1021, 2012.

[10] J. Salas and C. Tomasi, "People detection using color and depth images," in *Mexican Conference on Pattern Recognition (MCPR)*, 2011.

[11] O. Arif, W. Daley, P. Vela, J. Teizer, and J. Stewart, "Visual tracking and segmentation using time-of-flight sensor," in *IEEE International Conference on Image Processing (ICIP)*, Sept. 2010.

[12] A. Bevilacqua, L. Di Stefano, and P. Azzari, "People tracking using a time-of-flight depth sensor," in *IEEE International Conference on Video and Signal Based Surveillance*, Nov. 2006.

[13] L. Spinello and K.O. Arras, "People detection in rgb-d data," in *IEEE / RSJ International Conference on Intelligent Robots and Systems (IROS)*, Sept. 2011.

[14] M. Harville and D. Li, "Fast, integrated person tracking and activity recognition with plan-view templates from a single stereo camera," in *CVPR*, 2004.

[15] S. Bahadori, L. Iocchi, G.R. Leone, and D. Nardi, "Real-time people localization and tracking through fixed stereo vision," *Appl. Intell.*, vol. 26, no. 2, pp. 83–97, 2007.

[16] S. Yous, H. Laga, and K. Chihara, "People detection and tracking with world-z map from a single stereo camera," in *The Eighth International Workshop on Visual Surveillance (VS2008)*, 2008.

[17] S.A. Guethmundsson, M. Pardas, J.R. Casas, et al., "Improved 3d reconstruction in smart-room environments using tof imaging," *Computer Vision and Image Understanding*, vol. 114, pp. 1376–1384, 2010.

[18] L. Xia, Chia-Chih Chen, and J.K. Aggarwal, "Human detection using depth information by kinect," in *Workshop on Human Activity Understanding from 3D Data in conjunction with CVPR (HAU3D)*, June 2011.