ECAI 2016
G.A. Kaminka et al. (Eds.)

doi:10.3233/978-1-61499-672-9-304

Robust Real-Time Human Perception with Depth Camera

Guyue Zhang [1], **Luchao Tian** [1], **Ye Liu** [2], **Jun Liu** [1], **Xiang An Liu** [1], **Yang Liu** [1] and **Yan Qiu Chen** [1*]

Abstract. Perception of the presence and position of human is crucial for many kinds of Artificial Intelligence (AI) applications. In this paper, we have developed a novel two-staged method for real-time human detection in depth image. The first stage is to quickly scan through the image to detect possible head-top locations in order to ensure all the candidate locations are included. The second stage is to use a novel head-shoulder descriptor (HSD) which jointly encodes the One-hot Depth Difference information and local geometric characteristics of human upper body to filter the detections so as to keep the genuine human locations and discard false positives. The results show that our approach using only depth data is superior to other methods using color and depth images on four datasets. In addition, our method performs well under weak illumination conditions or even total darkness. Moreover, our system is also able to run in real-time on conventional PC without GPU acceleration.

1 INTRODUCTION

Human detection is an important task due to its wide application in human-computer interaction, intelligent vehicles, autonomous indoor mobile robots, etc. It is also a critical technology in building smart rooms in which intellectual sensors should be aware of users' presence and locations [7, 8, 9]. However, human detection is still a challenging problem especially in occasion of occlusion, posture variations, dynamic and heavily cluttered background or crowd, etc.

Existing methods [6, 24, 28] for conventional video cameras are reported as able to work in well-illuminated environments with relatively simple and stationary background. However, their performance declines quickly if the illumination conditions deteriorate or the background becomes dynamic and complicated.

With the recent rapid development of depth cameras, such as time-of-flight camera and Kinect, human detection becomes more manageable as depth image is relatively insensitive to scene textures, and the depth information acquired by the sensors using actively emitted near infra-red medium is robust against illumination variation of the environment.

There have been methods for detecting human beings with depth cameras [2, 1, 31, 15, 23, 27, 32, 26]. The work reported in [4] adopts a graph-based segmentation algorithm combined with randomized subsampling for depth image segmentation and a set of parameterized heuristics to reduce candidate segments for classification. Wojek et al. [29] combine a full object detector and multiple object part detectors in a mixture of experts based on their expected visibility. Spinello et al. [22] take inspiration from HOG (Histogram of Oriented Gradients) detector which is mainly for color/grayscale image and design HOD (Histogram of Oriented Depths) descriptor for depth data, and then achieve promising human detection result. These methods mentioned above using full-body detectors show their effectiveness in many environments, but encounter challenges in crowded environments where occlusions occur frequently and people are often partially visible.

An upper-body detector is a good choice for robust human detection, since the upper part of human body is less likely to be occluded and less deformable. The approach proposed by Xia et al. [30] combines a 2D head contour model and a 3D head surface model to detect people in indoor environments. Ikemura et al. [12] introduce the notion of Relational Depth Similarity Features (RDSF) based on depth information, which is derived from a similarity of depth histograms and represents the relationship between two local regions. The method presented in [13] uses a continuous normalized-depth template as an upper-body detector for close range and a full-body detector for farther range. Choi et al.'s system [5] integrates multi-hypothesis (including human upper-body shape, human face, human skin, as well as human motion) and shows interesting results for locating people in 3D space. Liu et al. [18] combine a Ring-wedge Mask (RWM) and 2D Joint Histogram of Color and Height (JHCH) information to classify plausible human head candidates. Munaro et al. [19] combines depth-based and color-based techniques in a cascade algorithm to detect people.

In this paper, we propose a novel two-staged approach which fully utilizes the unique characteristics of the human's upper body from depth images only. We first quickly localize all candidate head-tops based on the contour information of human head. Excessive numbers of possible human head candidates are extracted in this stage, which contain true human head regions in the scene (with very low miss rate) as well as false positives. Although the false positive rate is still relative high, we are able to reduce the search space at a very low computational cost. These false positives in the detections from the first stage are further filtered out in the second stage by training a classifier with an effective upper-body head-shoulder descriptor (HSD) which consists of a One-hot Depth Difference Descriptor (ODDD) to describe the occlusion relationship among humans and their surrounding environments, and a Binarized Local Surface Descriptor (BLSD) to characterize the local surface geometrical properties of humans.

Our contributions are in the following aspects:

- We propose a fast head-top candidate extractor by localizing extreme points along the depth discontinuities and try to ensure all true human head-tops are included, which reduces searching space to obtain high speed.
- We propose a novel upper-body head-shoulder descriptor (HSD), which jointly encodes the information of One-hot depth differ-

[1] School of Computer Science, Shanghai Key Laboratory of Intelligent Information Processing, Fudan University, China. Emails: {guyuezhang13, lctian14, ljun, xaliu13, yliud13, chenyq}@fudan.edu.cn. * Corresponding author.
[2] College of Automation, Nanjing University of Posts and Telecommunications, China. Email: yeliu@njupt.edu.cn.

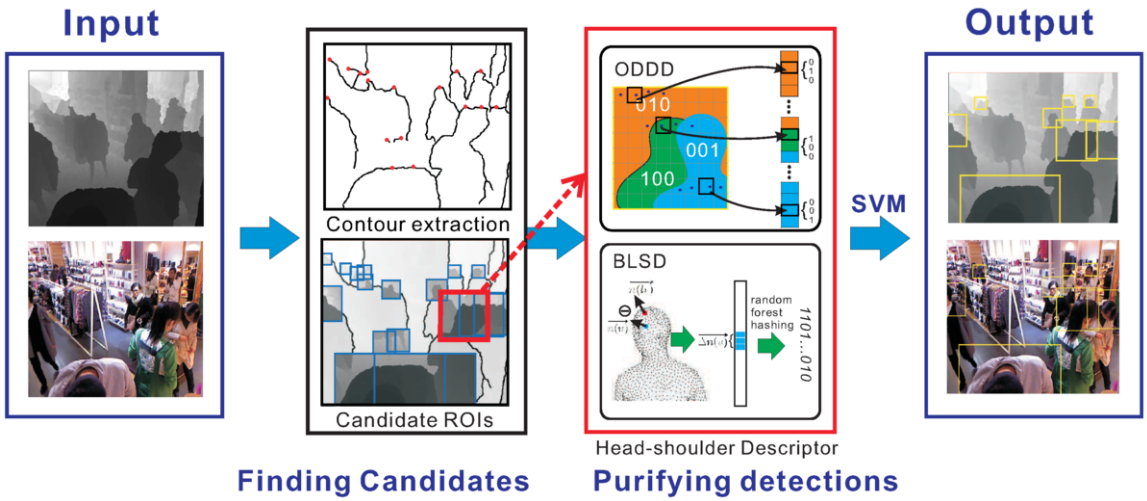


Figure 1: Workflow of the proposed method.

ence and local surface statistics to effectively and efficiently classify detections, aiming at further eliminating false positives while preserving the true detections. Both of the two descriptors are discriminative and compact, and they encode different kinds of information of the upper-body in depth images.

Experimental results have demonstrated the effectiveness of our approach on four available datasets. Although only depth data is employed, the proposed method outperforms existing methods using both depth and RGB data. Using only depth data is also advantageous in environments with dark or volatile illuminations.

2 THE PROPOSED APPROACH

Our method follows a two-stage cascaded structure. The first stage uses extreme points in edge map to detect candidate head-top locations and try to include all the probable locations. The second stage uses a novel head-shoulder descriptor (HSD) to verify the candidates so as to keep the genuine ones and discard false ones. We try to achieve a very low miss rate and expect to efficiently locate the candidates in the first stage, while the subsequent more computationally expensive verification stage only needs to deal with a limited number of candidates rather than on all image pixels. An overview of the proposed detection framework is given in Figure 1.

2.1 Finding Possible Head-top Points

Several existing works also tried to locate the head regions as ROI (Region of Interests) as a preliminary detection stage. For example, [30] and [13] use a depth template to localize the head positions as ROI to reduce the search space. However, the template matching often gets corrupted due to occlusions. [13] and [17] project the 3D point cloud to the ground plane and then use the height information to locate the probable head positions. But these methods require the prior knowledge of the ground plane which is either time consuming or even not possible to estimate.

Our motivation is to quickly and directly find possible head-tops in the depth image without the assist of any point clouds or depth templates, so that the more computationally intensive verification process needs to be applied to only a limited number of candidates rather than all pixels, thus substantially reducing the computation load. We try to ensure all genuine head-top points are included in the responses while allowing some false positives. There are two successive modules in this stage: depth based contour extraction and head-top candidates localization.

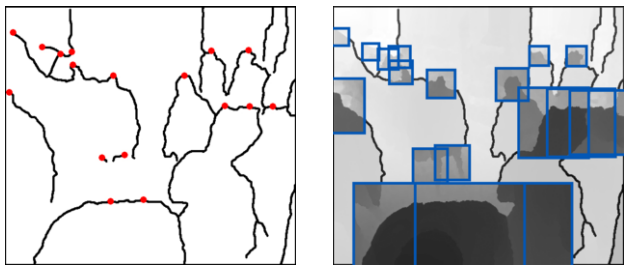
2.1.1 *Depth based Contour Extraction*

Depth data remains continuous within the same object and varies greatly across distinct objects or parts of objects. So depth discontinuities usually indicate true boundaries between two non-touching objects. In real-world scenes, a standing person's head is always sufficiently far away from its surrounding background. This inspires us to first extract the contour of human head based on depth discontinuities and then look for further cues in these contours. This cue makes it possible for us to obtain a set of less noisy contours corresponding to human head boundaries from depth data much easier than from RGB images.

Since depth data generated by depth camera may contain some noise and holes, we use a depth image inpainting technique [20] to reconstruct missing data, then the inpainted depth image is smoothed by a Gaussian filter. Gradient magnitude $M(x, y)$ and gradient orientation $\phi(x, y)$ are calculated with a Canny operator. Canny operator is employed as its outputs are not only isolated edge points but a set of contours with points linked which facilitates our further analysis of the contour. Then we locate every possible edge point by the non-maxima suppression (NMS) and extract contours by double-threshold edge linking scheme[3]. With conventional RGB image, the contours output by Canny can be noisy and fragmented, but as we have discussed above, with depth data we can always obtain clearer and more complete human head contours as shown in Figure 2(a).

2.1.2 *Head-top Candidates Localization*

In most scenarios, the camera is positioned to make the image y-axis inversely aligned with the gravity direction. Now that we have obtained relatively clean and complete contours of human heads, which protrude towards the ceiling of the image, we can find extreme points along the contours in which head-tops are contained as shown in Figure 2 (a). Given a contour consisted of a chain of points $C = \{(x_i, y_i)\}_{i=1}^{n}$, we define the extreme point l_k as the point which



(a) (b)

Figure 2: (a) shows an example of depth based contour extraction and extreme point localizations. The head-tops are in red color with contours in black. In (b), each head-top has its own corresponding ROI illustrated in blue box.

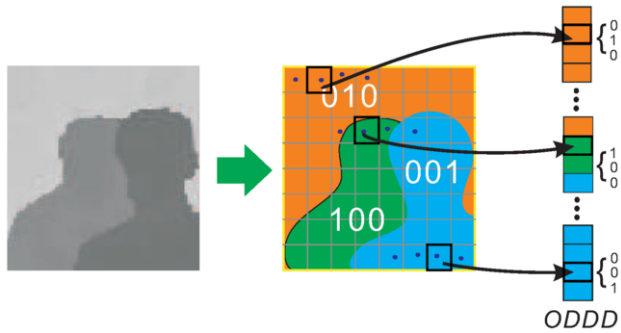


Figure 3: One-hot Depth Difference Descriptor (ODDD). We classify each pixel into three categories: detectee (100), background (010) and overlapper (001). The codes of picked points are concatenated to build a feature vector.

has a local maxima y value:

$$l_k = \{(x_k, y_k) \mid y_k \geq y_j, j \in N(k)\} \tag{1}$$

where $N(k) = \{k-s, ..., k+s\}$ is a neighborhood of point (x_k, y_k). We set $s = 3$ in all the experiments. Searching along all the contours extracted, we can obtain a set of extreme points as $L = \{l_k\}_{k=1}^{m}$, which are considered to be head-top candidates. The head-top candidates can be extracted ultra fast since that 1) we do not have to estimate the ground plane; 2) the extreme points are searched on the extracted contours rather than the whole image. Even in highly crowded environments, it makes sure that people's head-top positions are included in the resultant responses L (see in Section 3.2).

It is a novel idea to extract the candidate points information from edge maps for subsequent processing, which dramatically reduces the searching space and gains high detection speed. It can be easily extended to other vision tasks in depth image, such as object recognition, human activity analysis, and hand gesture analysis.

2.2 Purifying the Detections

We have obtained the probable head-top detection candidates, but the results are over-detected and the false positives should be further filtered out. We train a classifier using a novel head-shoulder descriptor (HSD) combining two features: One-hot Depth Difference Descriptor (ODDD) and Binarized Local Surface Descriptor (BLSD).

2.2.1 Scale Invariant ROI Selection

Conventional object detection in RGB images often involves testing detection windows with different scales to detect objects with different sizes on image, which is time-consuming. With the help of depth information, we can select the ROIs adaptively according to the depth of a 3D point as shown in Figure 2 (b).

We denote an ROI as $r_k = (l_k, w_k, h_k)$ around the head-top point to cover the whole head. Here l_k is a head-top candidate, w_k and h_k are the width and height of selected ROI for l_k, they are adjusted adaptively according to the corresponding depth value d_k. We select the ROI whose left-top corner is $l_k - (0.25h_k, 0.5w_k)$. This is set empirically to make the ROI cover the upper-body and some context. We prefer slightly larger ROIs to diminish the effect of inaccuracy in localizing head-top points. Since human head is roughly spherical, we let ρ denote the head radius in physical quantity while ρ_{d_k} denote the projected radius and $w_k = 3\rho_{d_k}, h_k = 4\rho_{d_k}$. The relationship between physical quantity and projected quantity is $\rho_{d_k} = \lambda\frac{\rho}{d_k}$, where λ is a constant factor obtained with camera's intrinsic parameters [11, 18].

2.2.2 Head-shoulder Descriptor (HSD)

The pattern of human head-shoulder in depth images is distinctive from other objects, so the key problem here is to design a head-shoulder descriptor (HSD) to describe this distinctive property. We design an HSD that encodes two different aspects of information: the relative positions among human upper-body and their surrounding environments and the local surface geometrical characteristics of human upper-body. With these two kinds of information compactly encoded, the HSD is highly discriminative in keeping the true human locations and rejecting false positives.

One-hot Depth Difference Descriptor (ODDD) Depth difference has been explored in human pose estimation in [21]. But it is not suitable for describing the pattern of human head-shoulder due to complex relationships between humans and their surrounding background. Depth difference between human and background varies from less than $1m$ to more than $10m$, making the resulted feature badly scaled and may cause misleading classification results. If we directly utilize depth difference as a feature, it is possible to acquire misleading results or may influence the classification results significantly. So directly employing depth differences to detect human is not a wise choice.

In order to make depth difference more suitable for our classification task, we propose a novel One-hot Depth Difference Descriptor (ODDD). Inspired by [18], we classify each pixel in an ROI into three categories: *detectee* (pixels belonging to the region of human to be detected), *background* (pixels regarded as background) and *overlapper* (pixels that can be considered as objects that occlude the human to be detected) by depth difference values (with threshold σ). We assign a three-bit one-hot code for each category, so for each pixel $u = (x, y)$ the feature $f(u)$ is computed as (Figure 3):

$$f^d(u) = \begin{cases} 100, & |d(u) - d(l_k)| \leq \sigma \\ 010, & d(u) - d(l_k) > \sigma \\ 001, & d(u) - d(l_k) < -\sigma \end{cases} \tag{2}$$

We divide the ROI into $\alpha \times \beta$ cells, and to obtain fast computation speed, one point rather than all points is randomly picked from each cell, the one-hot code of the picked points in all the cells are concatenated to build a feature vector.

The feature extracted in above manner is able to shield the large variation of depth differences among pixels while retaining the occlu-

sion relationships (the category) between humans and their surrounding environments. Also, using one-hot code can facilitate processing for many classification methods. It should be noted that [18] also seek to threshold depth difference value, however, they used a hard template with strong prior knowledge to perform classification. Our work is significantly different as we turn the category information of pixels into features and train a classifier with them, which leverages the advantage of large amount of data to account for various kinds of occlusions and view changes.

Binarized Local Surface Descriptor (BLSD) One-hot Depth Difference Descriptor (ODDD) focuses on describing the complex depth pattern formed by humans and their surrounding environments, but it is ineffective in characterizing the local geometrical property of objects' surfaces, which has been proven to be important in RGBD object recognition tasks [25]. With depth information, the 2D pixels can be reprojected into 3D space as point cloud, which represent the 3D surfaces of the scene. For a pixel location $u = (x, y)$, the 3D surface normal vector can be approximated as $\overrightarrow{n(u)} = \left(\frac{\partial d(u)}{\partial x}, \frac{\partial d(u)}{\partial y}, -1\right)^{\mathrm{T}}$ [25]. Instead of building a histogram of normal vectors which loses the spatial information of points, we propose a binarized local surface descriptor (BLSD) which encodes the local surface smoothness in different spatial locations (Figure 4). The feature is also extracted from the ROI in Sec. 2.2.1.

For a pixel location $u = (x, y)$, we first compute the Normal Vector Difference (NVD):

$$f^n(u) = \overrightarrow{\Delta n(u)} = \overrightarrow{n(u)} - \overrightarrow{n(l_k)} \tag{3}$$

where $\overrightarrow{n(u)}$ and $\overrightarrow{n(l_k)}$ are normalized normal vectors at u and l_k. NVD is much more robust to view point change than normal vectors since the normal vector at head-top l_k is substracted. And we also divide the ROI into cells, one point is randomly picked in each cell and its NVD is computed. Concatenating the NVDs of all the sampled points in all the cells will form a feature vector F_k which is highly redundant. Instead of using this feature vector for classification, we convert it to a binary code which is much more compact and can be more effectively processed.

We follow the approach of random forest based hashing to learn the compact binary codes [14]. A set of random forests $\{T_i\}_{i=1}^{M}$ is trained from the training data. Each $T_i = \{t_i^1, t_i^2, ..., t_i^N\}$ is trained using a randomly selected subset of training data and a randomly selected subset of features of F_k, which serves as a binary test for generating one bit of the binary code (i.e. it generates 1 if F_k is classified as positive and 0 otherwise). In this manner, we are able to obtain a compact M bit BLSD which is combined with ODDD as our final binarized features for classification.

2.2.3 *Training and classification*

For training the classfier, we use 7,600 positive and 25,060 negative training samples. These samples are obtained in the following way: firstly detections are generated by the first stage of our method on 23 RGB-D video sequences, then 7,600 true human head-tops and 25,060 false positives are manual selected. Finally, ROIs are determined and from which HSDs are extracted.

In the classification procedure, we use a linear Support Vector Machine (SVM) to classify the Head-shoulder Descriptor (HSD) (by concatenating ODDD and BLSD) for an ROI to decide whether the region contains a human head or not.

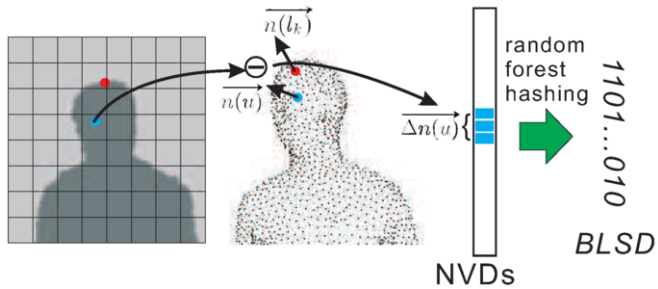


Figure 4: Binarized Local Surface Descriptor (BLSD). Head-top point l_k is in red color and randomly sampled points u are in blue color. Normal vector differences (NVDs) are computed and concatenated, and then are converted into a compact binary code by random forest hashing.

3 EXPERIMENTS AND DISCUSSIONS

We evaluate the detection accuracy and computational efficiency of the proposed method and compare it with other state-of-the-art approaches on four available datasets. These datasets are captured with Kinect at 640×480 resolution.

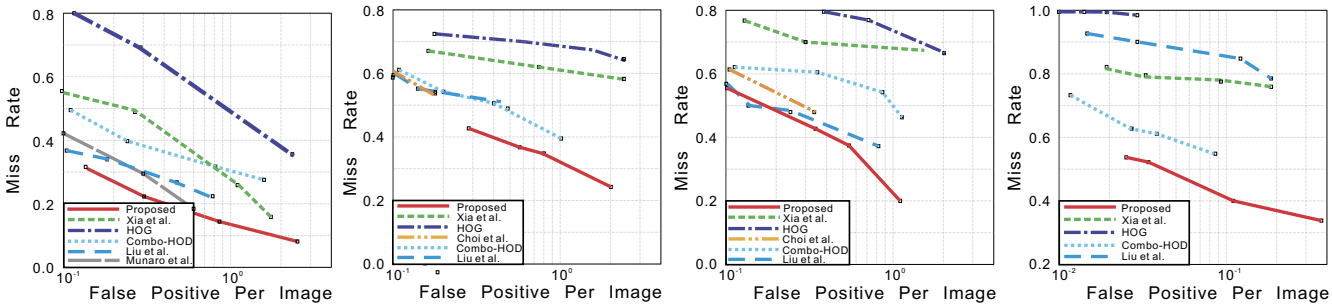
3.1 Datasets and Metrics for Evaluation

The first dataset is the CLOTHING STORE [16, 17]. It contains two video sequences of 45 minutes length each. The scene is cluttered with pillars, hangers, clothes, cabinets and shoe racks. People take on various poses such as walking, sitting and bending, and they interact with each other frequently.

The second and third datasets named OFFICE and MOBILE PLATFORM are provided by Choi et al. [5]. The former contains 17 video sequences and was captured in an office room. The environment is cluttered, and people in this dataset face different directions and take various poses, such as standing, walking, and sitting on chairs. The latter was collected with a Kinect mounted on a PR2 robot driving around in a building. It contains 18 video sequences with different scenes. This dataset includes various illumination conditions and cluttered backgrounds.

To the best of our knowledge, there is no publicly available RGB-D dataset captured under dark illumination for person detection. In order to comprehensively assess the performance of our method in such environments, we collected a challenging dataset named DARK. This dataset is captured at night with dark illumination which makes the persons indistinguishable from RGB images, as shown in Figure 7 (d). We will show that our method works well under such weak illumination or totally dark conditions. This new dataset is available at `http://www.cv.fudan.edu.cn/humandetection.htm`.

Also, we evaluate the performance via false-positive-per-image (FPPI) vs. miss-rate in our experiments. FPPI is computed in a standard way, as total number of false positives divided by frame numbers. Four images per second from OFFICE and MOBILE PLATFORM [5], one image every three seconds from CLOTHING STORE and DARK are selected to evaluate the performance [17]. A successful detection is counted if the overlap ratio between the annotated bounding box and the detected bounding box is above 0.5.



(a) CLOTHING STORE (b) OFFICE (c) MOBILE PLATFORM (d) DARK

Figure 5: Comparison results with other approaches on four datasets CLOTHING STORE, OFFICE, MOBILE PLATFORM and DARK.

3.2 Analysis and Evaluation

We evaluate the detection accuracy and computational efficiency of the proposed method, with the overall system performance against other methods for a quantitative evaluation, and then we provide a comprehensive evaluations respectively on the computational times, the role of first stage, the contribution of each feature, and the effect of different distance.

Computational Times The proposed system was measured on a desktop PC with i5-2500 CPU and 8GB RAM, and runs at 40 fps without GPU acceleration, which is faster than the recording speed of Kinect (30 fps).

Overall system performance We compare the proposed system against a conventional HOG detector [6], a depth-based detector proposed by Xia et al. [30], a Combo-HOD detector [22], a color-depth detector proposed by Choi et al. [5], an RWM human locator [18], and cascade classifier proposed by Munaro et al. [19] on four datasets illustrated in Figure 5 (a)-(d). The results shown in Fig. 5 are obtained by using the codes from original authors [18, 19] and by our implementation [6, 22, 30]. The performance of [5] is only evaluated on OFFICE and MOBILE PLATFORM (the performance is reported by the authors), as the source code is not available.

The experiments show that our algorithm outperforms state-of-the-art detectors. In the results, the performance of HOG detector is limited due to the clutter of background and various people's poses in color images. Xia et al.'s method uses a 2D head contour model and a 3D head surface model, which is strongly dependent on human shape, it may fail when in side-view cases. The Combo-HOD detector and Munaro et al.'s method work well in spacious environments, but the performance decreases in our test scenes where people are occluded and posing variedly such as sitting or bowing. The approach proposed by Choi et al. combines multiple cues based on color and depth data. But it may fail because the depth information is not fully exploited. The RWM locator has a strong assumption of the parameter in different categories so that it is easy to fail in divergent scenes of occlusion and tilt.

The proposed method using only depth information provides a more reliable result than the method HOG using RGB data only and the method of Xia et al. using depth only. Moreover, our method even yields higher accuracy than the methods [22, 5, 18, 19] , which utilize both color and depth information. Especially in the DARK dataset, where RGB information is limited and many detector with RGB can not work, our approach is significantly superior to others. Some detecting examples are shown in Figure 7. This demonstrates that our method is quite robust in dealing with real-world challenging tasks including occlusion, variations in postures and clutter, and also dark illumination.

Table 1: Average of miss rate and FPPI in first stage.

Dataset	Miss Rate	FPPI
CLOTHING STORE	0.044	41.12
OFFICE	0.049	50.32
MOBILE PLATFORM	0.027	31.45
DARK	0.057	44.32
Average	0.044	41.80

Contribution of the first stage We evaluate it on four datasets with the results in Table 1 with average miss rate at 4.4% and average FPPI at 41.80. The results show that only a few false positives are contained in the responses. This indicates that the finding possible head-top points stage is very effective for search space reduction. And in this stage, the average run time for one frame is around 10 ms, which is fast enough to ensure the real-time processing. On average, about 25 points from the three hundred thousand pixels in a frame are detected as candidate head-top points (1/12,000).

Contribution of the two features We compare each feature used in the proposed descriptor separately. In this experiment, we test each feature at a time to compare the detection results on CLOTHING STORE. As illustrated in Figure 6 (a), using only one feature (ODDD or BLSD) may dramatically decrease the performance than combining ODDD and BLSD together. In addition, we compare the influence between depth difference only and ODDD in Figure 6 (a). It indicates that the detection performance of ODD-D can be improved in average precision much higher than directly using depth differences.

Impact of the distance We evaluate the effect of the distance from camera on detection performance. As the Microsoft Kinect sensor has a practical ranging limit of $0.8m - 3.5m$ distance [10], the long distance out of the practical range may be more fragmentary and noisier than the short distance. Figure 6 (b) depicting the analysis of the effects of long range ($> 3.5m$) and short range ($\leq 3.5m$) shows that the proposed method can provide higher detection accuracy for nearer humans.

Figure 7: Examples of detection results on (a) CLOTHING STORE, (b) OFFICE, (c) MOBILE PLATFORM, and (d) DARK.

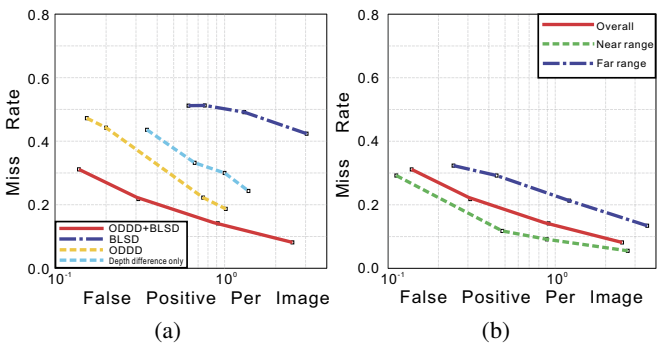


Figure 6: (a) Contribution analysis of each detection feature. The red curve represents the results of the method with both ODDD and BLSD features. Purple and yellow curves show the results of a specific feature either BLSD or ODDD. Blue curve represents depth difference feature. (b) presents the detection performance for different distance ranges ($>$3.5m or $<$3.5m).

4 CONCLUSION

We have presented in this paper a novel two-staged method for detecting humans in depth images. The possible human head-top points are extracted in the edge map by the first stage. These candidates are then fed to the second verification stage to output the final detection results. Experiment results show that the proposed method (without RGB information) can reliably detect people in complex, dynamic and even dark environments in real time with high accuracy, and even outperforms state-of-the-art approaches that use RGB-D data.

ACKNOWLEDGEMENTS

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

REFERENCES

[1] Z. Cai, J. Han, L. Liu, and L. Shao, 'Rgb-d datasets using microsoft kinect or similar sensors: a survey', *Multimedia Tools and Applications*, 1–43, (2016).

[2] M. Camplani, A. Paiement, M. Mirmehdi, D. Damen, S. Hannuna, T. Burghardt, and L. Tao, 'Multiple human tracking in rgb-d data: A survey', *arXiv*, (2016).

[3] J. Canny, 'A computational approach to edge detection', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (6), 679–698, (1986).

[4] B. Choi, C. Meriçli, J. Biswas, and M. Veloso, 'Fast human detection for indoor mobile robots using depth images', in *IEEE International Conference on Robotics and Automation*, pp. 1108–1113. IEEE, (2013).

[5] W. Choi and S. Pantofaru, C.and Savarese, 'A general framework for tracking multiple people from a moving camera', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **35**(7), 1577–1591, (2013).

[6] N. Dalal and B. Triggs, 'Histograms of oriented gradients for human detection', in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pp. 886–893. IEEE, (2005).

[7] D. Focken and R. Stiefelhagen, 'Towards vision-based 3-d people tracking in a smart room', in *International Conference on Multimodal Interfaces*, pp. 400–405. IEEE, (2002).

[8] S. A. Guomundsson, R. Larsen, H. Aanæs, M. Pardas, and J. R. Casas, 'Tof imaging in smart room environments towards improved people tracking', in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1–6. IEEE, (2008).

[9] J. Han, E. J Pauwels, P. M De Zeeuw, and P. HN De With, '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**(2), 255–263, (2012).

[10] J. Han, L. Shao, D. Xu, and J. Shotton, 'Enhanced computer vision with microsoft kinect sensor: A review', *IEEE Transactions on Cybernetics*, **43**(5), 1318–1334, (2013).

[11] C Herrera, J. Kannala, J. Heikkilä, et al., 'Joint depth and color camera calibration with distortion correction', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **34**(10), 2058–2064, (2012).

[12] S. Ikemura and H. Fujiyoshi, 'Real-time human detection using relational depth similarity features', in *Asian Conference on Computer Vision*, 25–38, Springer, (2011).

[13] O. H. Jafari, D. Mitzel, and B. Leibe, 'Real-time rgb-d based people detection and tracking for mobile robots and head-worn cameras', in *IEEE International Conference on Robotics and Automation*, pp. 5636–5643. IEEE, (2014).

[14] X. Li, C. Shen, A. Dick, and A. Hengel, 'Learning compact binary codes for visual tracking', in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2419–2426, (2013).

[15] T. Linder and K. O Arras, 'People detection, tracking and visualization using ros on a mobile service robot', in *Robot Operating System (ROS)*, 187–213, Springer, (2016).

[16] J. Liu, Y. Liu, Y. Cui, and Y. Q. Chen, 'Real-time human detection and tracking in complex environments using single rgbd camera', in *Image*

Processing (ICIP), 2013 20th IEEE International Conference on, pp. 3088–3092. IEEE, (2013).

[17] J. Liu, Y. Liu, G. Zhang, P. Zhu, and Y. Q. Chen, 'Detecting and tracking people in real time with rgb-d camera', *Pattern Recognition Letters*, **53**, 16–23, (2015).

[18] J. Liu, G. Zhang, Y. Liu, L. Tian, and Y. Q. Chen, 'An ultra-fast human detection method for color-depth camera', *Journal of Visual Communication and Image Representation*, **31**, 177–185, (2015).

[19] M. Munaro, C. Lewis, D. Chambers, P. Hvass, and E. Menegatti, 'Rgb-d human detection and tracking for industrial environments', in *Intelligent Autonomous Systems*, 1655–1668, Springer, (2016).

[20] F. Qi, J. Han, P. Wang, G. Shi, and F. Li, 'Structure guided fusion for depth map inpainting', *Pattern Recognition Letters*, **34**(1), 70–76, (2013).

[21] J. Shotton, T. Sharp, A. Kipman, A. Fitzgibbon, M. Finocchio, A. Blake, M. Cook, and R. Moore, 'Real-time human pose recognition in parts from single depth images', *Communications of the ACM*, **56**(1), 116–124, (2013).

[22] L. Spinello and K. O. Arras, 'People detection in rgb-d data', in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 3838–3843. IEEE, (2011).

[23] S-Z Su, Z-H Liu, S-P Xu, S-Z Li, and R. Ji, 'Sparse auto-encoder based feature learning for human body detection in depth image', *Signal Processing*, **112**, 43–52, (2015).

[24] B. Tan, J. Zhang, and L. Wang, 'Semi-supervised elastic net for pedestrian counting', *Pattern Recognition*, **44**(10), 2297–2304, (2011).

[25] S. Tang, X. Wang, X. Lv, T. X Han, J. Keller, Z. He, M. Skubic, and S. Lao, 'Histogram of oriented normal vectors for object recognition with a depth sensor', in *Asian Conference on Computer Vision*, 525–538, Springer, (2013).

[26] L. Tian, G. Zhang, M. Li, J. Liu, and Y. Q. Chen, 'Reliably detecting humans in crowded and dynamic environments using rgb-d camera', in *Proceedings of the IEEE International Conference on Multimedia and Expo*, (2016).

[27] X-T Truong, V. N. Yoong, and T-D Ngo, 'Rgb-d and laser data fusion-based human detection and tracking for socially aware robot navigation framework', in *2015 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, pp. 608–613. IEEE, (2015).

[28] X. Wang, T. X. Han, and S. Yan, 'An hog-lbp human detector with partial occlusion handling', in *IEEE International Conference on Computer Vision*, pp. 32–39. IEEE, (2009).

[29] C. Wojek, S. Walk, S. Roth, and B. Schiele, 'Monocular 3d scene understanding with explicit occlusion reasoning', in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1993–2000. IEEE, (2011).

[30] L. Xia, C-C Chen, and JK Aggarwal, 'Human detection using depth information by kinect', in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pp. 15–22. IEEE, (2011).

[31] H. Xue, Y. Liu, D. Cai, and X. He, 'Tracking people in rgbd videos using deep learning and motion clues', *Neurocomputing*, **204**, 70–76, (2016).

[32] G. Zhang, J. Liu, L. Tian, and Y. Q. Chen, 'Reliably detecting humans with rgb-d camera with physical blob detector followed by learning-based filtering', in *Acoustics, Speech and Signal Processing (ICASSP), IEEE International Conference on*, pp. 2004–2008, (2016).