RELIABLY DETECTING HUMANS WITH RGB-D CAMERA WITH PHYSICAL BLOB DETECTOR FOLLOWED BY LEARNING-BASED FILTERING

Guyue Zhang[⋆] *Jun Liu*[†] *Luchao Tian*[⋆] *Yan Qiu Chen*[⋆ *]

[⋆]School of Computer Science, Shanghai Key Laboratory of Intelligent Information Processing, Fudan University, China
[†]School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore

ABSTRACT

This paper proposes a two-staged approach to real-time human detection in cluttered environments using RGB-D camera. The first stage is a novel physical blob (P-Blob) detection that can quickly find plausible human heads. The second stage uses a combination of human upper-body features to filter out false positives. Experiment results on three publicly available datasets show that the proposed method can reliably detect people in RGB-D video in real time.

Index Terms— Human detection, RGB-D camera, Physical blob detector

1. INTRODUCTION

Detecting people with camera is an important research problem due to its wide application in human-computer interaction, security surveillance etc. It has attracted a great deal of research attention and yet remains challengingly open due to complex background, various illumination conditions, different clothes and poses of people. A large number of human detection methods using conventional visible light video cameras have been proposed [1, 2, 3, 4, 5], which are reported to give good performance when background is relatively simple. However, when the textures of background become complicated, their performance often deteriorates dramatically.

With the advent of commercially available depth cameras, researchers noticed that the depth information is helpful to detect human beings in real world environments. Xia et al.[6] combine a 2-D head contour model and a 3-D head surface model to detect people in indoor environments. Ikemura et al.[7] 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. Spinello et al.[8] design HOD (Histogram of Oriented Depths) descriptor for depth data, and achieve promising human detection result.

*Corresponding author. {guyuezhang13, lctian14, chenyq}@fudan.edu.cn, jliu029@ntu.edu.sg. Thanks to National Natural Science Foundation of China, Grant No. 61175036 for funding.

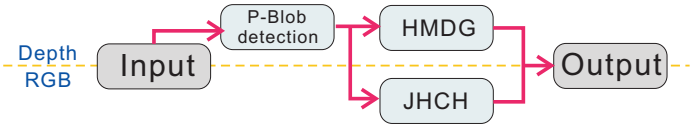


Fig. 1. Workflow of the proposed method.

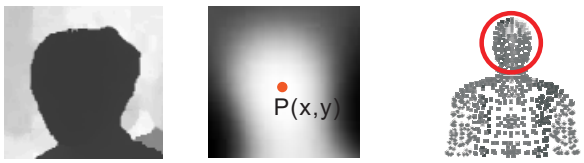
The method presented in [9] uses a continuous normalized-depth template for close range and ground HOG for farther range. Choi et al.'s system[10] integrates multihypothesis and shows interesting results for locating people in 3-D space.

The depth signal is much less sensitive to changes of illumination and textures than RGB signal. Moreover, depth image is easier for segmentation due to the significant depth value discontinues between the foreground and background objects, while segmenting color image is often quite difficult due to complicated textures.

When the environment is crowded with people, human body is frequently occluded and partially visible, so we detect human head to represent whole human since the upper part of human is less deformable or likely to be occluded.

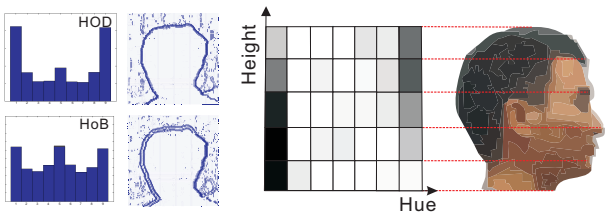
We propose in this paper a novel physical blob (P-Blob) human head detector by combining the image blob detection and filtering in 3-D physical space. Taking advantages of P-Blob, we try to ensure all genuine head regions are included in the responses. For each detected P-Blob, we combine Histogram of Multi-order Depth Gradients (HMDG) and Joint Histogram of Color and Height (JHCH) features as an upper-body descriptor to classify it into human or nonhuman class. An overview of the proposed detection framework is shown in Fig. 1.

The contributions of this work include: (1) A novel physical blob (P-Blob) human head detector is proposed to extract plausible head regions in 3-D space and avoid searching over the entire image for subsequent stages. (2) A robust upper-body descriptor jointly encodes shape information from depth data and color information from RGB data is proposed. (3) Experiment results show that the proposed method offers superior performance over existing approaches on three publicly available datasets.



(a) (b) (c)

Fig. 2. (a) A human head region in depth image. (b) Extremal point $P(x, y)$ of DoH response. (c) Diagram of P-Blob in 3-D physical space.



(a) (b)

Fig. 3. Features for filtering. (a) HMDG. (b) JHCH

2. PROPOSED METHOD

The proposed method consists of a physical blob (P-Blob) detection stage and a detection verification stage. The P-Blob detection is achieved through detecting blobs in depth image and then back projecting them to 3-D physical space to select those whose sizes are compatible with that of human head. Then at the second stage, the responses are further purified by a learning-based classifier.

2.1. Physical Blob Detection

The human head is spherically shaped with small size variation. This inspires us to develop method that can detect physical blobs in 3-D space to locate plausible head positions. This is accomplished by detecting image blobs in depth data and back projecting them into 3-D space to obtain their corresponding 3-D sizes and then to filter to keep only those whose sizes are compatible to that of human head.

2.1.1. DoH blob detection in 2-D depth image

From Fig. 2 (a), we can see that the pixels inside the human head region are considerably darker (closer to camera) than the background and the head region approximates a circle. These characteristics make human head appear like a blob. We use Determinant of Hessian (DoH) [11, 12, 13] to detect location of human head as blob like structures. Hessian Matrix for image point $P(x, y)$ and scale s is calculated as:

$$H(P, s) = \begin{bmatrix} L_{xx}(P, s) & L_{xy}(P, s) \\ L_{xy}(P, s) & L_{yy}(P, s) \end{bmatrix} \quad (1)$$

where $L_{xx}(P, s)$ is the convolution result of second order Gaussian derivative $\frac{\partial^2 g(s)}{\partial x^2}$ for the point $P(x, y)$ at scale s (similarly for L_{xy} and L_{yy}). These derivatives are known as Laplacian of Gaussians (LoG). To increase calculating speed, we ues Difference od Gaussians (DoG) to approximate and replace LoG, which can improve the detection efficiency [12].

Since the grey level of human head region is less than the background, we detect blob as the minimum extremal points of DoH responses (Fig. 2 (b)) denoted as:

$$(\hat{P}, \hat{s}) = argminlocal_{(P,s)}(detH(P, s)) \quad (2)$$

where $detH(P, s)$ is the determinant of Hessian Matrix.

The blob points $\hat{P}(\hat{x}, \hat{y})$ and scales $\hat{s}$ are also defined from an operational differential geometric definitions[14] that leads to blob descriptors covariant with translations, rotations and rescalings in the image domain.

We use blob detection in depth image rather than RGB image, because depth image has superior performance in segmentation as it is insensitive to textures. To the best of our knowledge, we are the first to use blob based method to detect human heads in depth image.

2.1.2. Blob filtering in 3-D physical space

The size of a human head in 3-D physical space is a constant while that of its region in depth images varies with its distance from the camera. This prompts us to use its physical size instead of image size to achieve higher detection accuracy.

The relationship between physical quantity and projected quantity is computed as: $h = \frac{d \times h_d}{\lambda}$ where h is the physical quantity of head diameter in world coordinates (in the range of $h_{min} < h < h_{max}$), λ is a constant factor obtained with cameras intrinsic parameters [15] and h_d is the projected quantity of head diameter in image at distance d[16]. By computing the correspondence of human head from depth image, we can remove the unnecessary noises those dissatisfy the size in physical space to get 3-D physical blobs (Fig. 2 (c)).

However, after this stage, there are still noises of DoH extremal points such as hands or other objects in background. So in the next stage, we filter the candidates via learning based classification.

2.2. Candidate Filtering

For the detection result of first stage, we extract the center point of each detected blob and select a scaled window region as a Region of Interest (ROI) with width W and height H ($W = 1.5 \times h_d, H = 2 \times h_d$) around the center point to cover the whole head in depth image. Then we combine our HMDG and JHCH features as an upper-body descriptor to determine whether the responses produced by previous stage are indeed human heads. The two features describe shape and appearance information of upper body respectively.

(1) Histogram of Multi-order Depth Gradients(HMDG):

Inspired by [17], which achieves good object detection results by fusion of zero-order, first-order, and second-order gradients in RGB image, we present a local 3-D shape feature HMDG combing Histograms of Oriented Depth (HOD) feature[8] and Histograms of Bar-shape (HoB) feature for depth image (see Fig. 3 (a)). HOD takes inspiration from the Histograms of Oriented Gradients (HOG)[1] and follows the similar procedure as HOG for the depth image. To get a feature vector, we compute the first-order depth gradient at each pixel, collect the gradients into cells, count a histogram on each cell then normalize the histograms and concatenate all the histograms.

HoB feature corresponding to the second-order gradients can also contribute to human detection. We employ HoB in depth image to get the second-order depth gradient (r_{xy}, θ_{xy}) at each pixel (x, y) as:

$$\theta = \frac{1}{2} \arctan\left(\frac{2 \cdot I_{xy}}{I_{xx} - I_{yy}}\right) \tag{3}$$

$$r = I_{xx} \cos^2\theta + 2I_{xy} \cos\theta \sin\theta + I_{yy} \sin^2\theta \tag{4}$$

where I is the depth value of ROI in depth image, $\vec{v} = (\cos\theta, \sin\theta)$ is the unit direction, I_{xx}, I_{xy}, I_{yy} are the second-order derivatives of I. As HoB feature has similar cell-based structure with HoD feature, we combine HoB and HoD by following the strategy proposed in [17].

(2) Joint Histogram of Color and Height (JHCH):

We utilize our JHCH feature[18] encoding height level locations of colors on object to characterize appearance information of human head. As show in Fig. 3 (b), the color statistics of human head is basically collected from face and hair or hair alone (when human is observed from the back). With the increase of the height, the probability of skin color may decrease while hair color often increases.

In our upper-body descriptor, we utilize the JHCH with five height intervals and nine color (hue) intervals. The black and white pixels are considered specially and are recorded in extra bins. By characterizing color and height statistically, JHCH works well in coping with people in different head poses and hair styles, and even with small tilt of ROI.

In the classification procedure, we use a linear Support Vector Machine (SVM) to classify the computed upper-body descriptor (by concatenating normalized JHCH and HMDG with equal weights) for an ROI to decide whether the region contains a human head or not.

3. EXPERIMENTS AND DISCUSSIONS

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

Clothing store dataset: This dataset created by us was captured in a clothing store[18]. It contains two video sequences of about 45 minutes long each. The main challenge of this dataset is that the background is complicated and dynamic. People in this dataset take various poses (walking, sitting, and bowing) and they interact with each other frequently (available at `http://www.cv.fudan.edu.cn/humandetection.htm`).

Office dataset: We used the office dataset provided by Choi et al.[10]. It contains 17 video sequences of 2-3 minutes long each. People in this dataset face different directions and take different poses, such as standing and sitting on chairs.

Mobile platform dataset: This dataset is captured by Choi et al.[10] with a Kinect mounted on a mobile platform (a PR2 robot) driving around in a building. It contains 18 video sequences covering different environments such as passageway, office room, and cafeteria. The challenges include various illumination conditions and cluttered backgrounds.

3.1. Implementation Details

The depth image captured by depth sensor often contains information missing and noise. To reduce the impact of artifacts, we apply preprocessing to reconstruct data missing regions using a depth map inpainting technique proposed in [19]. In our experiments, 23 video sequences, captured in laboratory, passageway, lounge and store, are used for training. In all, 4346 frames are selected as sample frames for training. Averagely, about three humans are contained in a frame.

3.2. Analysis and Evaluation

We determine false positives per image (FPPI) vs. miss rate curves, and evaluate the performance by selecting one image every three second from clothing store dataset and four images per second from office and mobile camera datasets.

In the evaluation frames, the upper bodies of people are hand-annotated with bounding boxes, and 3-D locations are inferred from the depth images and bounding boxes[20]. Two evaluation criteria from [10] are used to determine a positive detection. The first is based on the overlapping degree between the detected regions and the ground truth bounding boxes. The second is based on 3-D distance.

First, we compare our proposed system against a conventional HOG detector[1], a depth-based detector proposed by Xia et al.[6], a novel Combo-HOD detector [8], and a color-depth detector proposed by Choi et al.[10] on the three datasets illustrated in Fig. 4. As the source code of [10] is not available, it is only evaluated on office and mobile platform datasets provided by the authors. The results show that our proposed method outperforms other four methods. HOG detector[1] is limited due to the complex textures of background. Xia et al.'s method [6] uses a strong assumption on human shape by a 2-D head contour model and a 3-D

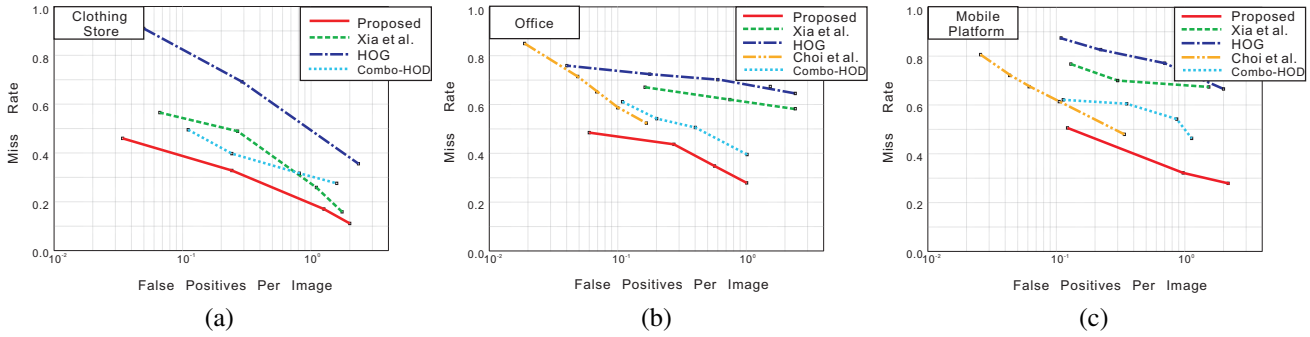


Fig. 4. FPPI vs. miss rate curves. Left to right: results on clothing store dataset, office dataset and mobile platform dataset.

head surface model, which may fail for side-view cases. The Combo-HOD detector[8] consists of an HOG detector on RGB image and an HOD detector on depth image. This full-body detector may work well in spacious scenes, but in our test scenes where people are crowded and occluded, the performance significantly decreases. Choi et al.'s method[10] combines multiple cues based on color and depth data. The color information is situation-dependent and only a depth template is utilized, so it also does not work well in our test scenes.

Then we analyze the performance of our P-Blob detection stage on clothing store dataset. Averagely, only 10 false positives are recorded in a frame, which demonstrates that the searching space for subsequent stages is greatly reduced. Meanwhile, the miss rate of this stage is low (about 0.06).

Next, we quantitatively evaluate single feature (HMDG or JHCH) by turning off one feature at a time to compare the individual contribution of the features used in our proposed descriptor on our clothing store dataset. Fig. 5 (a) shows that turning off HMDG or JHCH may dramatically decrease the performance. Combining two features together outperforms using any single feature.

Finally, we compare the effect of the different distance range from camera. The depth data for far distance may be more fragmentary and noisier than that of close distance[21]. Fig. 5 (b) shows the contrastive results of the far range ($> 3.5m$) and the close range ($\leq 3.5m$). We can draw the conclusion that the proposed method yields higher detection accuracy for closer humans.

Overall, the proposed method can reliably detect people in challenging scenarios including complex background and variations in postures. Fig. 6 shows some detecting examples in three datasets. The proposed system runs at about 20 fps on a desktop PC with a i5-2500 CPU, 8GB RAM without GPU acceleration.

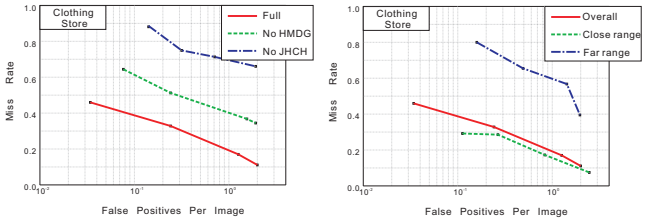


Fig. 5. (a) Evaluation of two features. (b) Impact of the distance.

Fig. 6. Examples of detection results on clothing store dataset (top row), office dataset (middle row), and mobile platform dataset (bottom row).

4. CONCLUSIONS

We have presented in this paper a robust human detection method that can deal with complex and dynamic environments with RGB-D camera. The proposed method accomplishes the detection task via two stages and achieves both high speed and accuracy.

5. REFERENCES

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

[2] X. Wang, T. X. Han, and S. Yan, "An hog-lbp human detector with partial occlusion handling," in *Computer Vision, IEEE 12th International Conference on*, 2009, pp. 32–39.

[3] S. Tang and S. Goto, "Histogram of template for human detection," in *Acoustics Speech and Signal Processing (ICASSP), IEEE International Conference on*, 2010, pp. 2186–2189.

[4] G. K. Vinay, S. M. Haque, R. V. Babu, and K. R. Ramakrishnan, "Human detection using sparse representation," in *Acoustics, Speech and Signal Processing (ICASSP), IEEE International Conference on*, 2012, pp. 1513–1516.

[5] A. Satpathy, X. Jiang, and H. L. Eng, "Human detection using discriminative and robust local binary pattern," in *Acoustics, Speech and Signal Processing (ICASSP), IEEE International Conference on*, 2013, pp. 2376–2380.

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

[7] S. Ikemura and H. Fujiyoshi, "Real-time human detection using relational depth similarity features," in *Computer Vision–ACCV 2010*, pp. 25–38. Springer, 2011.

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

[9] 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 *Robotics and Automation (ICRA), IEEE International Conference on*, 2014, pp. 5636–5643.

[10] W. Choi, C. Pantofaru, and S. Savarese, "A general framework for tracking multiple people from a moving camera," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 35, no. 7, pp. 1577–1591, 2013.

[11] Z. M. Qian, X. E. Cheng, and Y. Q. Chen, "Automatically detect and track multiple fish swimming in shallow water with frequent occlusion," *PLoS ONE*, vol. 9, no. 9, 2014.

[12] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[13] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (surf)," *Computer vision and image understanding*, vol. 110, no. 3, pp. 346–359, 2008.

[14] T. Lindeberg, "Feature detection with automatic scale selection," *International journal of computer vision*, vol. 30, no. 2, pp. 79–116, 1998.

[15] C. Herrera, J. Kannala, and J. Heikkilä, "Joint depth and color camera calibration with distortion correction," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 34, no. 10, pp. 2058–2064, 2012.

[16] 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*, vol. 31, pp. 177–185, 2015.

[17] Y. Jiang and J. Ma, "Combination features and models for human detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 240–248.

[18] 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), 20th IEEE International Conference on*, 2013, pp. 3088–3092.

[19] F. Qi, J. Han, P. Wang, G. Shi, and F. Li, "Structure guided fusion for depth map inpainting," *Pattern Recognition Letters*, vol. 34, no. 1, pp. 70–76, 2013.

[20] 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*, vol. 53, pp. 16–23, 2015.

[21] J. Han, L. Shao, D. Xu, and J. Shotton, "Enhanced computer vision with microsoft kinect sensor: A review," *Cybernetics, IEEE Transactions on*, vol. 43, no. 5, pp. 1318–1334, 2013.