FAST HUMAN DETECTION IN RGB-D IMAGES BASED ON COLOR-DEPTH JOINT FEATURE LEARNING

Zhan Hu[1], Haizhou Ai[1], Haibing Ren[2], Yimin Zhang[2]

[1]Tsinghua National Lab for Info. Sci. & Tech., Depart. of Computer Sci.& Tech.,
Tsinghua University, Beijing, 100084, China.
[2]Intel Labs China, Intel Corporation, Beijing, 100086, China
huz14@mails.tsinghua.edu.cn, ahz@mail.tsinghua.edu.cn, {haibing.ren, yimin.zhang}@intel.com

ABSTRACT

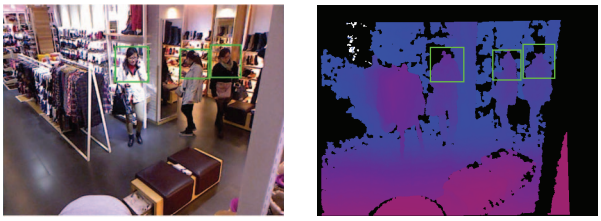
Human detection in RGB-D images is an important yet very challenging task in computer vision. In this paper, we propose a novel human detection approach in RGB-D images, which integrates ROI (region-of-interest) generation, depth-size relationship estimation and a human detector. Our approach has the following advantages: 1) ROI generation and depth-size relationship estimation take full advantage of color and depth information to fast reject about 70% negative samples while maintaining a high recall rate; 2) the cascade-structured human detector can seamlessly concatenate features extracted from both color and depth images; and 3) our method can detect human at a speed of more than 30 fps on 640×480 images on a single laptop CPU without any GPU acceleration. Experiments on challenging public datasets demonstrate the effectiveness of our method.

Index Terms— human detection, RGB-D data, RGB-D camera, real-time system

1. INTRODUCTION

RGB-D camera is a device that can simultaneously capture a color image and a depth image, of which the depth represents the distance between objects and the camera. These cameras have been applied on many platforms[1, 2] notably Microsoft Kinect, Intel Realsense and Asus Xtion. Human detection in RGB-D images is not only an important technology for many potential applications including driverless car, visual surveillance, human-computer interaction and robotic navigation, but also a key prerequisite for human re-identification, human tracking and human retrieval.

In literature, Spinello et al.[3] proposed Histogram of Oriented Depths (HOD) descriptors, whose computation is very similar to HOG[4]. This method utilizes depth information but ignores color and texture information, which is very likely to cause false detections of non-human objects. Besides, it cannot be done in real time without ROI generation. Various approaches, such as AdaBoost[5, 6], SVM[4, 7] and deep neural networks[8], have been applied to human detection in RGB images without depth information. Spinello et al.[3] also developed COMBO-HOD descriptors which extract features from both color and depth images. But this method will fail if the color or depth image is in poor quality, for example, in images captured by a Intel RealSense camera.



(a) Color image (b) Depth image

Fig. 1. Example of detection result for RGB-D camera. The left one (a) is a color image[9][10] and the right one (b) is the corresponding depth image[9][10] captured by an RGB-D camera.

Recent approaches have adopted a strategy that restricts a detector to a pre-generated ROIs. Under the assumption that people are moving on a ground plane, works in [11-14] generate ROIs by removing ground plane and fixed objects (such as walls, buildings and trees) on it. By applying a HOG+SVM[11] detector afterwards, they are able to achieve a real time performance since the search space is deduced. Unfortunately, these methods are restricted to work only for cameras which capture images with a big part of ground plane. Besides, algorithms for ground plane removal, for example the algorithm in Point Cloud Library (PCL), are very sensitive to parameters so it is not robust to scenery changes. What is more, some regions that contain people will be discarded, when a person stands adjacent to fixed objects.

An important point, that the size of a detection window is related to its depth value, is ignored from previous work. In the image pyramid and sliding window detection approach, it will waste a lot of time at detecting through all the sub-windows. Therefore, we present a way of estimating size

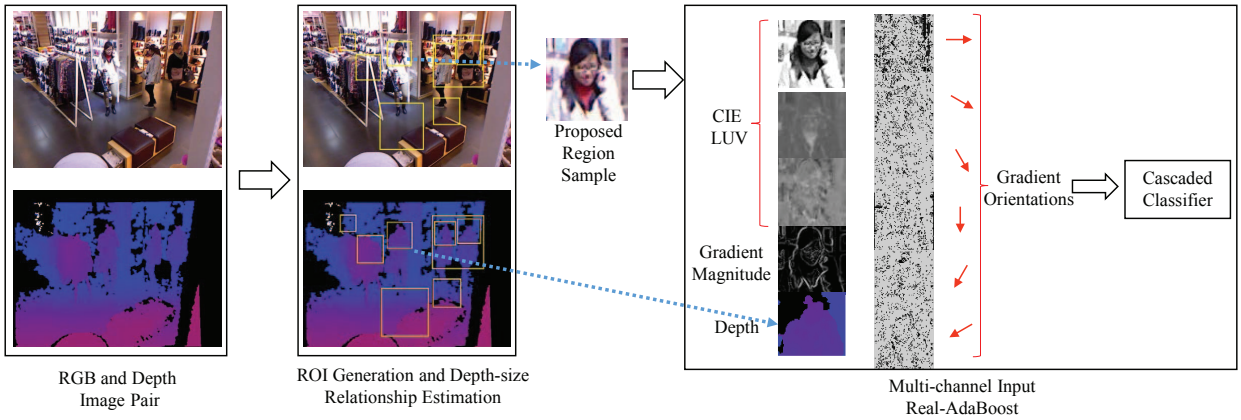


Fig. 2. **Overview of the proposed method.** For each color and depth image pair[9, 10], we use our ROI generation method to generate ROIs. Then, the depth-size relationship estimation method is applied to filter the detection windows. After that, the remaining detection windows will be verified by the human detector which is trained with multi-channel input with Real-AdaBoost.

of detection window according to its depth value. Our main contributions in this paper include: 1)we propose a novel approach to fully exploit both color and depth information for our human detector, which can still works well when the quality of color image or depth image is bad; 2)we propose a robust ROI generation approach to reject about 30% negative samples without ground plane removal and our depth-size estimation method can reject about 50% negative samples with a high recall rate; 3)the proposed depth-size relationship estimation approach is capable of accelerating the detection process; and 4) we conduct comparative experiments to analyze the contributions of different features extracted from depth image.

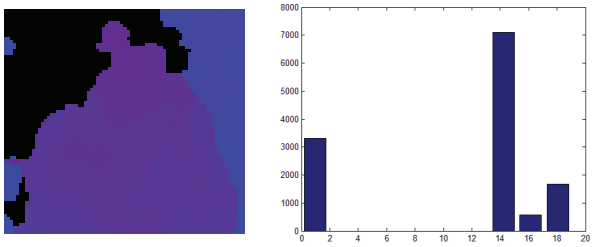


Fig. 3. We construct a histogram of a depth image region to calculate representative depth value. The left image[9, 10] is a depth image region and the right one is its corresponding histogram.

2. THE PROPOSED METHOD

2.1. Overview of the Proposed Method

Fig. 2 shows the overview of our human detection approach. For each color and depth image pair, we use our ROI generation method to generate ROIs. Then, the depth-size relationship estimation method is applied to filter detection windows. After that, the remaining detection windows will be verified by the human detector which is trained with multi-channel input with Real-AdaBoost.

2.2. ROI Generation and Depth-size Relationship Estimation

We follow the main idea of binarized normed gradients (BING)[12] to generate ROIs. BING takes advantage of a linear SVM, using very simple 64 bit binary feature extracted from color image to scan the image in different scales. And it can achieve 300 fps on a single laptop CPU. But we do not apply non-maximal suppression on these bounding boxes after scanning. Instead, we apply our depth-size estimation method to eliminate some bounding boxes. After that, we merge all these remaining bounding boxes into large regions as our ROIs. Besides, only positive human samples, instead of general objects, are taken into account to train the BING model. This ROI generation method can generate people area proposals efficiently and robustly according to our experiments in Section 3.

In order to estimate the relationship between depth value and its corresponding window size, we first need to define a representative depth value for the window. We make a histogram of depth image region as Fig. 3. The area containing human can be separated from non-people area by adaptive mean thresholding method in the histogram, and the major one of these two parts is the area containing people. The average depth value of human part is a representative depth value.

We divide the detectable depth range into N parts. For each part, we construct the histogram of the sizes of the windows. The mode $\overline{s}$ of the histogram is selected by mean shift[13] as the mean of the human sizes and the variance (σ) of the mode is also estimated. Variation of these size is modeled as a single global Gaussian distribution $G(\overline{s}, \sigma)$.

Test window will be verfied by a Gaussian distribution function according to its representative depth value. if

$$|s - \overline{s}| > 3\sigma. \tag{1}$$

then this window will be filtered and will not be passed to the human detector. In Equation (1), s is the size of current detection window.

2.3. Human Detector

RGB images are rich in color and texture information but are sensitive to illumination changes. On the contrary, depth images are robust to illumination changes, but become fragile when the sensor's signal is weak. Therefore, depth-color joint feature is able to provide more reliable information than features extracted only from either depth image or color image.

Our human detector, based on color-depth joint feature, is a Real-AdaBoost detector whose inputs are multi-channel features. Unlike the commonly used boosted detector that select features computed on one single channel (gray image as an example), our detector attempts to learn features extracted on multiple channels. To train the Real-AdaBoost human detector based on color-depth joint feature, we introduce LUV channels, HOG channels, depth channel, HOD (depth gradient magnitude and depth gradient orientation histogram) channels into our detector. Porikli[14] presents an efficient approach to compute histogram through histogram integral images. Dollar et al.[5] compute gradient orientation histograms using histogram integral image according to [14]. Dollar et al.[5] also introduce a way to compute gradient magnitude. We introduce our way to compute depth gradient magnitude and depth gradient orientations. Depth gradient magnitude is can be obtained:

$$G(x, y) = \sqrt{G_h^2(x, y) + G_v^2(x, y)}, \tag{2}$$

where $G(x, y)$, $G_h(x, y)$ and $G_v(x, y)$ are the gradient magnitude, horizontal gradient and vertical gradient at point (x, y) respectively. Depth gradient orientation histogram is the weighted gradient orientation whose weight is gradient magnitude. The m-th depth gradient histogram (total M depth gradient orientations) at point (x, y) $H_m(x, y)$ can be obtained by:

$$H_m(x, y) = \begin{cases} G(x, y) & arctan(\frac{G_v(x,y)}{G_h(x,y)}) \in [\theta_1, \theta_2) \\ 0 & otherwise \end{cases} \tag{3}$$

where $\theta_1 = \frac{(m-1)mod(M)}{M}\pi$ and $\theta_2 = \frac{m}{M}\pi$.

For each Haar feature h_c, all samples will be divided into N bins according to their Haar feature value, where c denotes the input channel. the sample distribution of bin i is as follows:

$$W_{+1}^i = \sum_{y_s=+1} w_s \qquad W_{-1}^i = \sum_{y_s=-1} w_s, \tag{4}$$

where y_s is the label of s-th sample and w_s is its weight. The Loss function F of h_c is:

$$F(h_c) = 2\sum_i \sqrt{W_{+1}^i W_{-1}^i}. \tag{5}$$

We choose the Haar feature h_c^{opt} as the current weak classifier

$$h_c^{opt} = \underset{h_c \in \chi}{\operatorname{argmin}} F(h_c), \tag{6}$$

where χ is the Haar feature pool. And the sequence of h_c^{opt} is our color-depth joint feature.

Work in [5] proved that, for people detection in RGB images, features in LUV+HOG (gradient magnitude and gradient orientation histograms) can achieve top performance. We perform experiments to show which channels should be extracted from depth image and then incorporated with LUV and HOG. During the training stage of our real-AdaBoost human detector, the input channel and its corresponding Haar feature are automatically selected according to the training loss function. In this way, our method can seamlessly join features extracted from color image and depth image.

3. EXPERIMENTS

We perform experiments on Human Detection And Tracking Dataset (HDAT)[9, 10] and RGB-D People Dataset (RGBDP)[3, 15].

Human Detection And Tracking Dataset. This dataset contains $3,001$ RGB-D frames acquired in a clothing store with Kinect mounted at 2.2m height with oblique angle of about 30 degrees with respect to the floor on one Friday evening. there are some clothes in this store that looks very similar to human bodies. And the background of this scene is very complex.

RGB-D People Dataset. $3,399$ RGB-D frames are captured in a university hall from three Kinect sensors. This dataset contains standing and upright walking people in different perspectives with different levels of occlusions. And some people cannot be seen clearly due to the un-ideal illumination condition.

Experimental Results. Our ROI generation can reject averagely 31.6% regions with 99.7% recall on HDAT Dataset and 36.4% regions with 99.4% recall on RGBDP Dataset. Depth-size relationship estimation can reject another 42.7% detection sub-windows with 100% recall on

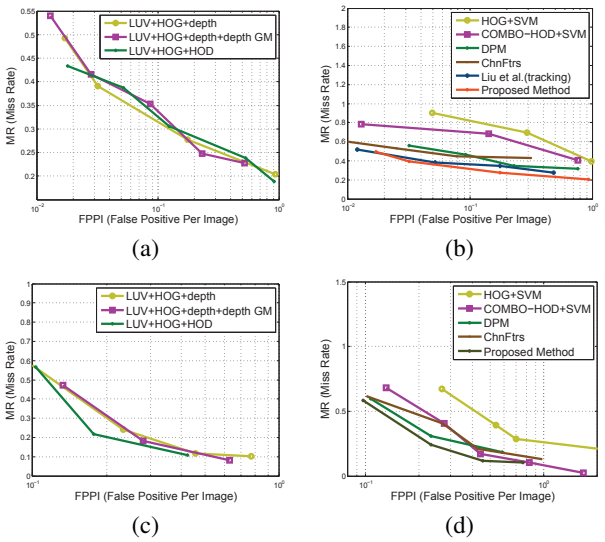
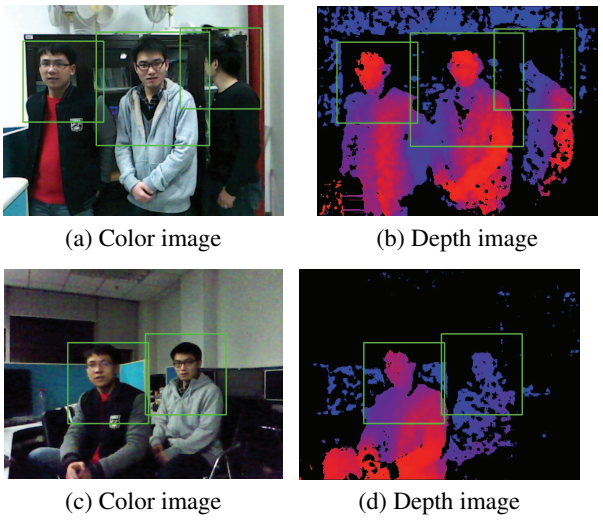


Fig. 4. **Experimental results** (a)(c) present performance differences between features extracted from depth image. (c)(d) show the comparison of ROC curves between our method and other methods. Results of (a)(b) are performed on H-DAT Dataset[9, 10] and (c)(d) is performed on RGBDP Dataset[3, 15].

HDAT Dataset and 53.6% sub-windows with 99.8% recall on RGBDP Dataset.

In Fig.4(a)(c), we present the performance comparison of LUV+HOG+depth, LUV+HOG+depth+depth gradient magnitude and LUV+HOG+depth+HOD channels. According to our experimental results, detector with LUV+HOG+depth+HOD input channels can achieve a slightly better performance than the other two. But we choose LUV+HOG+depth as our final detector input channels as a compromise between performance and detection speed.

Fig.4(b) shows the experimental results of our human detector based on color-depth joint feature on HDAT Dataset compared with HOG+SVM, COMBO-HOD+SVM (COMBO-HOD derives from HOD[3]), DPM[7], ChnFtrs[5] and Liu et al[10]. Fig.4(d) presents the results on RGBDP Dataset compared with HOG+SVM, DPM, COMBO-HOD+SVM and ChnFtrs. We can see that our approach outperforms all these techniques above. When FPPI is 0.1, the detection rate is improved by about 7.5% on HDAT Dataset and 5% on RGBDP Dataset, comparing with method in Liu et al.[10] and DPM[7] respectively. Our approach is also demonstrated on our own dataset captured by a Intel RealSense camera. Fig.5 shows some examples of the detection results. (a)(c) are color images and (b)(d) are their corresponding depth image. From these figures we can see that our approach still performs well even if color image or depth image is in bad quality.



(a) Color image (b) Depth image

(c) Color image (d) Depth image

Fig. 5. Examples of detection results for images captured by a Intel RealSense camera. (a)(c) are color images and (b)(d) are their corresponding depth image.

Table 1. Human detection speed on a laptop computer

	Time cost	
	Human Detection And Tracking	RGB-D People
Serial-1	67ms	58ms
Parallel-1	24ms	20ms
Serial-2	113ms	108ms
Parallel-2	42ms	36ms

The speed of our approach is shown in Table.1. Comparing with results in Serial-2 and Parallel-2, Serial-1 and Parallel-1 can achieve approximately a 2-fold acceleration of the detection process with ROI extraction and depth-size relationship estimation.

4. CONCLUSION

In this paper, we introduce a novel human detection approach with ROI generation, depth-size relationship estimation and a human detector based on color-depth joint features. Our ROI generation can efficiently reject non-people area with a very high recall rate, and is more robust than methods based on ground plane or fixed objects removal. The depth-size relationship estimation filters the detection windows with little accuracy loss. In the meantime, our human detector integrated with both color and depth information is demonstrated to achieve state-of-the-art performance.

5. ACKNOWLEDGEMENTS

This work was supported by the Natural Science Foundation of China (Project Number 61521002).

6. REFERENCES

[1] Krystof Litomisky, "Consumer rgb-d cameras and their applications," *Rapport technique, University of California*, p. 20, 2012.

[2] Leandro Cruz, Djalma Lucio, and Luiz Velho, "Kinect and rgbd images: Challenges and applications," in *Graphics, Patterns and Images Tutorials (SIBGRAPI-T), 2012 25th SIBGRAPI Conference on*. IEEE, 2012, pp. 36–49.

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

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

[5] Piotr Dollár, Zhuowen Tu, Pietro Perona, and Serge Belongie, "Integral channel features," in *BMVC*, 2009, vol. 2, p. 5.

[6] Genquan Duan, Chang Huang, Haizhou Ai, and Shihong Lao, "Boosting associated pairing comparison features for pedestrian detection," in *Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on*. IEEE, 2009, pp. 1097–1104.

[7] Pedro Felzenszwalb, David McAllester, and Deva Ramanan, "A discriminatively trained, multiscale, deformable part model," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE, 2008, pp. 1–8.

[8] Wanli Ouyang and Xiaogang Wang, "Joint deep learning for pedestrian detection," in *Computer Vision (ICCV), 2013 IEEE International Conference on*. IEEE, 2013, pp. 2056–2063.

[9] Jun Liu, Ye Liu, Ying Cui, and Yan Qiu Chen, "Real-time human detection and tracking in complex environments using single rgbd camera," in *Image Processing (ICIP), 2013 20th IEEE International Conference on*. IEEE, 2013, pp. 3088–3092.

[10] Jun Liu, Ye Liu, Guyue Zhang, Peiru Zhu, and Yan Qiu Chen, "Detecting and tracking people in real time with rgb-d camera," *Pattern Recognition Letters*, vol. 53, pp. 16–23, 2015.

[11] Chih-Chung Chang and Chih-Jen Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011, Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

[12] Ming-Ming Cheng, Ziming Zhang, Wen-Yan Lin, and Philip Torr, "Bing: Binarized normed gradients for objectness estimation at 300fps," in *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*. IEEE, 2014, pp. 3286–3293.

[13] Dorin Comaniciu and Peter Meer, "Mean shift: A robust approach toward feature space analysis," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 24, no. 5, pp. 603–619, 2002.

[14] Fatih Porikli, "Integral histogram: A fast way to extract histograms in cartesian spaces," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*. IEEE, 2005, vol. 1, pp. 829–836.

[15] Matthias Luber, Luciano Spinello, and Kai O Arras, "People tracking in rgb-d data with on-line boosted target models," in *Intelligent Robots and Systems (IROS), 2011 IEEE/RSJ International Conference on*. IEEE, 2011, pp. 3844–3849.