AN ATTEMPT TO PEDESTRIAN DETECTION IN DEPTH IMAGES

Shengyin Wu[1], Shiqi Yu[1], Wensheng Chen[2]
[1] College of Computer Science and Software Engineering, Shenzhen University
[2] College of Mathematics and Computational Science, Shenzhen University
Shenzhen, P. R. China, 518060
Email: {wushengyin@email., shiqi.yu@, chenws@}szu.edu.cn

***Abstract*—We investigate pedestrian detection in depth images. Unlike pedestrian detection in intensity images, pedestrian detection in depth images can reduce the effect of complex background and illumination variation. We propose a new feature descriptor, Histogram of Depth Difference(HDD), for this task. The proposed HDD feature descriptor can describe the depth variance in a local region as Histogram of Oriented Gradients(HOG) describes local texture cues. To evaluate pedestrian detection in depth images, we also collected a large dataset, which contains not only depth images but also the synchronized intensity images. There are 4673 pedestrian samples in it. Our experimental results show that detecting pedestrians in depth images is feasible. We also fuse the HDD feature from depth images and HOG from intensity images. The fused feature gives an encouraging detection rate of 99.12% at FPPW=10^{-4}.**

***Keywords*—Pedestrian detection, Depth image, HDD**

I. Introduction

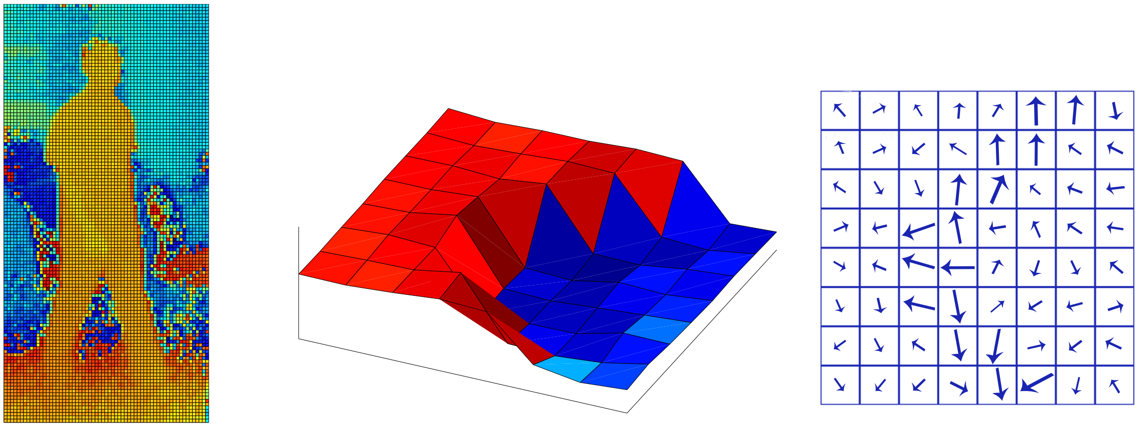
Pedestrian detection has great needs in many fields, such as robotics, surveillance, etc. Pedestrian detection algorithms can improve a system's perception ability. However, due to the variations of illumination, human posture, clothing, complex background and occlusion, it is difficult to detect the human body in images accurately.

In recent decade, pedestrian detection has gained increasing attentions from computer vision researchers due to its rapid development. Many novel methods have been proposed to detect pedestrians in images and videos. Some early contributions employ Haar wavelet [1], [2] and edge orientation histograms [3] as feature descriptor. Leibe *et al.* combined local and global cues via probabilistic top-down segmentation for pedestrian detection [4]. In 2005, Dalal and Triggs proposed a method that takes Histograms of Oriented Gradients (HOG) as the feature descriptor and SVM as the classifier [5]. Their method gives encouraging results. It is a milestone contribution to pedestrian detection, and many newer methods are inspired by HOG. Due to the high computational costs of HOG in real-time detection, Zhu *et al.* adopted AdaBoost and Cascade to accelerate HOG feature descriptor selection and pedestrian detection [6]. To solve the problem of partial occlusion, Wang *et al.* [7] combined HOG and (Local Binary Pattern)LBP as feature descriptor and achieved good performance. Enzweiler *et al.* [9] presented a survey and experiments on pedestrian detection, their experimental results show that when compared with wavelet-based AdaBoost cascade and Neural Network/LRF, the HOG/LinSVM combination obtains an excellent performance at higher resolutions while the processing speed is lower. But the methods mentioned before are just restricted in sliding detect window with only one rigid template when scanning the images. In 2009, Dollár *et al.* benchmark the state of the art pedestrian system and their experimental results and analysis indicated that HOG still competitive [10].

Although great progresses have been achieved in pedestrian detection, some problems are still challenging, such as illumination changes, complex backgrounds. To solve these problems, one solution is using depth images. Unlike intensity images, a depth image is a 2D array, and the pixel values stored in it are the distances from objects to the depth camera. A depth image is similar to a gray scale (or RGB) image, but the intensity values are replaced with the distance values. The pixel values in a depth image only depend on the distance of objects, and are not sensitive to light intensity and colors. Some researchers have done some pioneer work on pedestrian detection in depth images. Krotosky *et al.* [12] used four cameras to obtain depth images and propose a sterobased pedestrian detection method. Gidel *et al.* employed a multilayer laser scanner to capture depth information and fused them to track pedestrian in urban environment [13]. Rohrbach study on fusing HOG and LEF feature descriptor from depth and intensity to detect pedestrian, but the depth informations are computed from a calibrated stereo camera. [14]

A new method for pedestrian detection in depth images is proposed in the paper. The depth camera we used is a novel Time-Of-Flight (TOF) camera. The TOF camera can capture the whole scene at the same time, and then output intensity images and depth images. The intensity image and depth image are synchronized, which means the positions of the same instance are the same in these two kinds of images. The feature descriptor used here is named as Histogram of Depth Difference (HDD), and it can describe the depth variance in depth images. The SVM is employed as classifier. The main contribution of our work is to experimentally demonstrate that pedestrian detection in depth images is feasible.

The remainder of this paper is organized as follows. In Section II we describe the feature extraction method and the classifier. The dataset involved in the experiments is introduced in Section III. The experiments and analysis are presented in



(a) A depth image. The yellow color means it is closer.

(b) A cell region (8×8) cropped from the depth image(a) in the neck region.

(c) The depth differences of the pixels in the cell.

Fig. 1. A depth image and the depth differences.

Section IV. At last Section V concludes the paper.

II. Proposed Method

A. Histogram of Depth Difference

Histogram of Oriented Gradients is a good local descriptor, many state of the art pedestrian detection methods are inspired by it [7], [6], [15], [16]. Our proposed feature, named as Histogram of Depth Difference (HDD), is also inspired by HOG. As in [5], a detection window can be divided into overlapped blocks, and each block is divided into 4 cells. For a 64×128 pixels window, the cell size is 8×8 pixels, and there are $15 \times 7 = 105$ blocks in a detection window. Surely other values can be chosen for the window size, block size and cell size. For a pixel located at (x, y) in a cell (Fig. 1(b)), its depth difference can be calculated as follows:

$$\begin{aligned} \Delta_x &= \frac{D(x+1, y) - D(x-1, y)}{2} \\ \Delta_y &= \frac{D(x, y+1) - D(x, y-1)}{2} \end{aligned} \tag{1}$$

where $D(x, y)$ is the depth value at (x, y) in a depth image, Δ_x and Δ_y are the depth differences in the X direction and Y direction respectively.

Since the depth difference has two components Δ_x and Δ_y, it can also be represented by a magnitude and an orientation as shown in Fig. 1(c). In the Fig. 1(c), we can find that the depth differences can present the depth variation of depth images. A histogram with orientation bins can be employed to describe a cell region. The difference between HOG and HDD is the orientation bins. The orientation bins for HOG are spaced over $[0°, 180°)$, and this perform best in pedestrian detection [5]. For HDD in our method, we put emphasis on the bins which spaced over $[0°, 360°)$. Because the human body is always closer to the camera than the background, the HDD orientation on the body boundary should point from the body area outward to the background area. It is no need to flip orientation as HOG does. HDD feature can describes the distance variance in depth images, and represents local geometrical structures.

B. The SVM classifier

The Support Vector Machine [17] with a linear kernel trained in our method is the same as which in Dalal and Triggs's [5]. One reason to choose SVM is that it is convenient to compare HDD with HOG under the same classifier. Besides, SVM is also an outstanding classifier for many classification problems.

III. Dataset

To our knowledge, there is still no open depth dataset for pedestrian detection available now, so we collected a dataset, which contains intensity images and synchronized depth images. The depth camera we used is a Time-Of-Flight camera designed by Mesa Imaging AG, and the type of the camera is SwissRanger^{TM} SR4000 [18]. Beside depth images, the camera can also capture synchronized intensity images. The resolutions of the depth images and intensity images are all 176×144. The depth images are distance arrays, and the distances range from 0 to 5 meters. The intensity images are gray scale images, the valid range of intensity data is from 0 to 65535.

Using the TOF camera, we captured 4637 pedestrian images and 198 non-pedestrian images in 3 different indoor environments. Pedestrians in these images are all standing or walking. The body orientation to the camera is not limited and can be any direction, the pedestrians are upright , fully visible without any occlusion. The pedestrian positions were manually labeled by rectangles. The 4637 pedestrian (positive) samples are divided into two parts for training and testing respectively. There are 3160 positive training samples, and 1477 positive test samples. Besides, we cropped 14199 negative samples for training, and 56802 negative samples for testing. Some positive and negative samples are shown in Fig. 2 and 3. The dataset is available from http://yushiqi.cn/research/depthdataset.

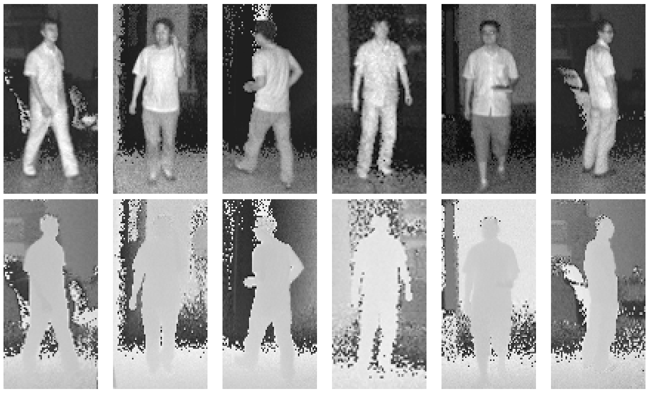


Fig. 2. Some positive samples from our dataset. The intensity images are shown in the first row, while the depth images are in the second.

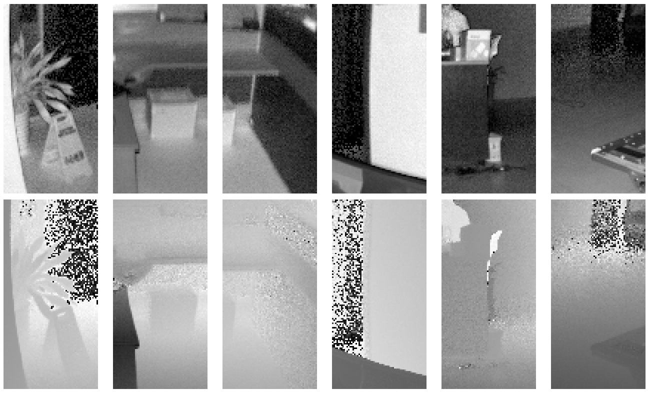


Fig. 3. Some negative samples from our dataset. The intensity images are shown in the first row, while the depth images are in the second.

IV. Experimental Results and Analysis

We evaluated the proposed method on our dataset. The default approach parameters are based on Dalal and Triggs [5]. The samples were normalized to a fixed size 64 × 128 pixels. The cell size is 8 × 8 pixels. The block size is 2 × 2 cells, and the block stride is 8 pixels. For the HDD feature, the orientation is divided into 9 bins which are spaced over $[0°, 360°)$, and each bin is $40°$. There are also 9 bins for HOG, but the bins are spaced over $[0°, 180°)$ in [5]. HDD for a depth sample and HOG for an intensity sample are all be presented as a 3780-D vector.

We use Detection Error Tradeoff (DET) curve [5] to evaluate the classification performance in our experiments. This curve measures the proportion of true detections against the proportion of false positives, which means using Miss Rate versus False Positives Per Window (FPPW). Better performance has lower miss rate and FPPW values.

Two groups of experiments were carried out.

A. *Performance of HDD feature*

The first group of experiments aim at investigate the parameters which affect the performance of HDD feature descriptor, [5] presented that the parameters: cell(8 × 8), block(2 × 2), stride(8) and normalized scheme(L2-Hys) are overall the best choices. But for the bins space, a good orientation coding will lead to a better performance. HDD is different from HOG, as analysis in section II, differences of depths should be more distinct when considering both orientations.

The results of the first group of experiments are shown in Fig. 4. Increasing the orientation bins increases the performance. In our experiment, 18 bins is better than 9 bins. The best curve(miss rate = 3.8% in FPPW = 10^{-4}) is achieved when HDD with 18 orientation bins and bins space($360°$), slightly better than the bins space($180°$). The contribution of these results is two folds. The first is the wide range of clothes and the complex contour between pedestrian and background reduce the discrimination of the signs of contrast. The second one is that the significant signs of contrast the depth information contains leads to a more discriminative representation alongside the contour between pedestrian and background.

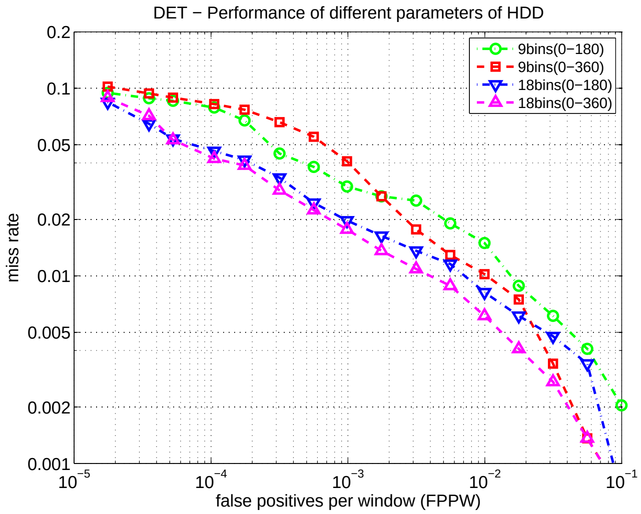


Fig. 4. The performance of different parameters on HDD.

B. *Experiment of different descriptors using FPPW*

The second group of experiments is designed to compare different feature descriptors. The parameters of the feature descriptor we employe are based on the best choice in the first group experiments. The parameters of HDD are based on the best performance in the first group of experiment. On the other hand, those of HOG are set based on Dalal's work [5]. The first experiment employe HDD feature from depth images. The second experiment employe HOG from intensity images, and it is designed to compare with HDD. In the last experiment, HDD from depth images and HOG from intensity images are combined to achieve better performance.

The results of the second group of experiments are plotted as curves in Fig. 5. HDD achieves a miss rate of 3.8% at FPPW=10^{-4}, while HOG achieves 11% in [5]. It is not convincing to declare HDD is superior to HOG since the two experiments were carried out on different datasets. Since we also collected the synchronous intensity images with depth images (Fig. 2 shows that.), we test HOG on our intensity dataset, and achieves a miss rate of 12.3% at FPPW=10^{-4}. Our HOG result is similar with that in [5]. It shows that HDD achieves better performance than HOG. The improvement of HDD should be brought by the robust to illumination change. We also tried to fuse HDD from depth images and HOG from intensity images on our dataset. We concatenate the HOG and HDD feature vector into a new 7560-D (3780 × 2) feature vector as feature descriptor. Our experiment shows that the performance is greatly improved using the fused feature descriptor. The miss rate decreased to 0.88%, and this means the detection rate reaches to 99.12%. It is an encouraging result.

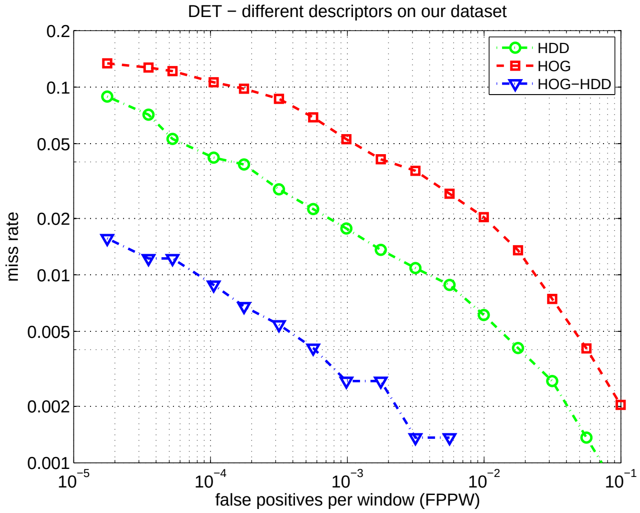


Fig. 5. The performance of 3 different kinds of features on our dataset.

V. Conclusions and Future Work

This paper presents an attempt to detect pedestrians in depth images. A type of new feature descriptor, Histogram of Depth Difference, is proposed to describe local depth variance. HDD shows its great discriminant capability in our experiments. It can decrease the miss rate from HOG's 10.6% to HDD's 3.8%. If combine the depth information with the intensity information, the miss rate can reach to a very low value of 0.88%. Our work on depth images shows that pedestrian detection in depth images is feasible, and it has some unique advantages than that just in intensity images.

The depth images we captured are in a relative low resolution (176 × 144), and are also noisy. In future we will work on some new depth cameras with high resolution, such Microsoft Kinect for XBOX 360 [19]. Kinect can capture depth images with a resolution of 640 × 480, and the depth values are more accurate than our current SR4000 camera. However, these two kinds of depth cameras depend on their active infrared light source. They can only work indoors, and can not work outdoors. To detect pedestrians outdoors using some other devices such as laser scanners is also our interest.

Acknowledgements: We would like to thank Dr. Kui Jia for his suggestions of our research and Mr. Yongfang Yu for his great work on data collection. This work is partly supported by Shenzhen Fundamental Research Program (Grant No. JC200903170431A) and China Postdoctoral Science Foundation(Grant No. 20090460532) and NSF of China Grant 60873168 and Science & Technology Planning Project of Shenzhen City (JC200903130300A) and Guangdong Province Key Laboratory of Computational Science (201106002) and Shenzhen Fundamental Research Program (Grant No. JC200903170431A).

References

[1] A. Mohan, C. Papageorgiou, and T. Poggio, "Example-based object detection in images by components," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 4, pp. 349–361, Apr. 2001.

[2] P. Viola, M. J. Jones, and D. Snow, "Detecting pedestrians using patterns of motion and appearance," in *Proc. of the Ninth IEEE International Conference on Computer Vision*, Oct. 2003, pp. 734–741 vol.2.

[3] W. T. Freeman, K. Tanaka, J. Ohta, and K. Kyuma, "Computer vision for computer games," in *Proc. of the The Second International Conference on Automatic Face and Gesture Recognition*, Oct. 1996, pp. 100–105.

[4] B. Leibe, E. Seemann, and B. Schiele, "Pedestrian detection in crowded scenes," in *Proc. of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Jun. 2005, vol. 1, pp. 878–885 vol. 1.

[5] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. of the Computer Vision and Pattern Recognition*, Jun. 2005, vol. 1, pp. 886–893 vol. 1.

[6] Q. Zhu, M. C. Yeh, K. T. Cheng, and S. Avidan, "Fast human detection using a cascade of histograms of oriented gradients," in *Proc. of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2006, vol. 2, pp. 1491–1498.

[7] X. Wang, T. X. Han, and S. Yan, "An HOG-LBP human detector with partial occlusion handling," in *Proc. of the IEEE 12th International Conference on Computer Vision*, Oct. 2009, pp. 32–39.

[8] T. Ahonen, A. Hadid, and M. Pietikainen, "Face description with local binary patterns: Application to face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 12, pp. 2037–2041, Dec. 2006.

[9] Enzweiler M. and Gavrila D.M., "Monocular pedestrian detection: Survey and experiments," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, no. 12, pp. 2179 –2195, dec. 2009.

[10] Dollár P., Wojek C., Schiele B., and Perona P., "Pedestrian detection: A benchmark," in *CVPR*, June 2009.

[11] T. Corneliu and S. Nedevschi, "Real-time pedestrian classification exploiting 2D and 3D information," *Intelligent Transport Systems, IET*, vol. 2, no. 3, pp. 201–210, Sep. 2008.

[12] S. J. Krotosky and M. M. Trivedi, "On Color-, Infrared-, and Multimodal-Stereo approaches to pedestrian detection," *IEEE Transactions on Intelligent Transportation Systems*, vol. 8, no. 4, pp. 619–629, Dec. 2007.

[13] S. Gidel, P. Checchin, C. Blanc, T. Chateau, and L. Trassoudaine, "Pedestrian detection method using a multilayer laserscanner: Application in urban environment," in *Proc. of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, Sep. 2008, pp. 173–178.

[14] Rohrbach M., Enzweiler M., and Gavrila D. M., "High-level fusion of depth and intensity for pedestrian classification," in *Proceedings of the 31st DAGM Symposium on Pattern Recognition*, Berlin, Heidelberg, 2009, pp. 101–110, Springer-Verlag.

[15] McAllester D. Felzenszwalb P., Girshick R. and Ramanan D., "Object detection with discriminatively trained part based models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, Sep. 2010.

[16] C. Zeng, H. Ma, and A. Ming, "Fast human detection using MI-SVM and a cascade of HOG-LBP features," in *Proc. of the 17th IEEE International Conference on Image Processing*, Sep. 2010, pp. 3845–3848.

[17] V. N. Vapnik, *The Nature of Statistical Learning Theory*, Springer., 1995.

[18] "MESA Imaging AG.," http://www.mesa-imaging.ch.

[19] "Microsoft Kinect for XBOX 360," http://www.xbox.com/en-US/kinect.