# A Two-Step Pedestrian Detection Algorithm Based on RGB-D Data

Qiming Li[1,2(✉)], Liqing Hu[2], Yaping Gao[3], Yimin Chen[3], and Lizhuang Ma[1]

[1] Department of Computer Science and Engineering,
Shanghai Jiaotong University, Shanghai 200240, China
[2] College of Information Engineering, Shanghai Maritime University,
Shanghai 201306, China
qmli@shmtu.edu.cn
[3] Department of Computer Science and Technology, Shanghai University,
Shanghai 200444, China

**Abstract.** A two-step pedestrian detection algorithm based RGB-D data is presented. Firstly, down-sample the depth image and extract the key information with a voxel grid. Second, remove the ground by using the random sample consensus (RANSAC) segmentation algorithm. Third, describe pedestrian characteristics by using Point Feature Histogram (PFH) and estimate the position of pedestrian preliminarily. Finally, calculate the pedestrian characteristics based on color image by Histogram of Oriented Gradient (HOG) descriptor and detect pedestrian using Support Vector Machine (SVM) classifier. The experimental results show that the algorithm can accurately detect the pedestrian not only in the single-pedestrian scene with pose variety but also in the multi-pedestrian scene with partial occlusion between pedestrians.

**Keywords:** Pedestrian detection · RGB-D · Down-sample · PFH · HOG · SVM

## 1 Introduction

Pedestrian detection is a hot research topic in computer vision. It plays an important role in vehicle assistant driving, intelligent video surveillance, human abnormal behavior detection, and so on. Pedestrians are both rigid and flexible, and the appearance is susceptible by wearing, occlusion, scale zoom, perspective and other factors, so the pedestrian detection is still a challenging issue.

Pedestrian detection technology based on RGB image has been mature, but the RGB feature is unitary. With the popularity of depth acquisition device, more and more researchers have begun to study pedestrian detection based on depth image. The classical HOG features used in color image is used to depth image by Spinello [1, 2], the Histogram of Oriented Depth (HOD) descriptor is proposed. Based on the algorithm presented by Spinello, Choi et al. [3] proposed a framework of multiple detector fusion, including skin color detection, face detection, trunk detection and so on. Wu et al. [4] proposed the histogram of depth difference feature. Yu et al. [5] proposed simplified

Z. Shi et al. (Eds.): ICIS 2017, IFIP AICT 510, pp. 288–294, 2017.
DOI: 10.1007/978-3-319-68121-4_31

local ternary patterns features based on depth image. Xia et al. [6] detect human targets by using two-dimensional contour model and 3D facial model. Wang et al. [7] presented a pyramid depth self-similar feature extraction algorithm according to the strong similar depth information of the human local body. Ikemura and Fujiyoshi [8] think that the frequency distribution of the depth values of different objects has a great diversity. They divide the depth map into nonoverlapping blocks and carry on histogram statistics of depth values for each block. The relational depth similarity Features of all blocks in series is used as the feature of the image. In recent years, Cheng et al. [9] proposed a semi supervised learning framework. Wang et al. [10] extract the fusion features of RGB and depth information, and reduce the dimension of the feature by using sparse automatic encoding. Gupta et al. [11] reselect exactly after primary select. Liu et al. [12] use synthetic image to detect pedestrian. Linder and Arras [13] detect pedestrian by using the local multi feature.

In summary, most of the existing pedestrian detection techniques are based on RGB images, so the feature is single, while the pedestrian detection based on depth images is prone to drift. Therefore, a pedestrian detection method based on RGB-D image is proposed in this paper. The algorithm is executed in two steps. Based on the depth image detection, the initial detection results are given first, and then the final detection results are obtained based on the RGB image.

## 2 Two-Step Pedestrian Detection Algorithm

As shown in Fig. 1, firstly, the point cloud expression of the scene is constructed based on depth data, and the down sampling is performed to the point cloud by using the voxel grid method in order to reduce the data size. Then the random sampling consistency segmentation algorithm is adopted to remove the ground plane from point cloud to reduce point cloud data further. Finally the PFH and SVM are used to screen out the candidate pedestrians.

Further detection needs to combine with RGB image. At present, the HOG feature is still the most effective and widely used in the pedestrian detection method based on RGB image, so the HOG feature and the SVM classifier are used in the further accurate detection to get the final pedestrian detection results.

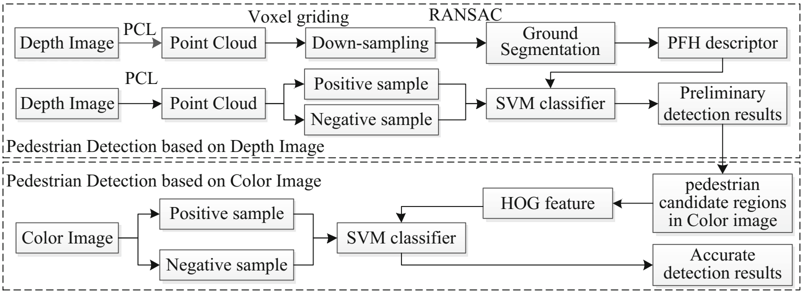

**Fig. 1.** Algorithm flow

### 2.1 Down-Sampling to Depth Point Cloud

The point cloud expression of the scene is constructed based on depth image. However, the data size of point cloud is relatively large. In order to improve the detection speed, the point cloud data must be down-sampled on the premise of keeping the shape of the cloud. In each frame, the spatial data is divided into a set of voxels. The coordinates of all points within each voxel are approximated to the coordinates of the voxel's center. A 3D voxel grid is created based on the point cloud data by using the point cloud library (PCL), and the center of gravity is approximately expressed by the coordinates of all points in voxel. Voxel can reduce the order of magnitude of data processing, reduce the calculation of pedestrian characteristic, and achieve the real-time performance. In addition, this operation is also conducive to constant point cloud density, makes the cloud data volume consistent in the unit space, and makes the point cloud density dependent on the distance from the sensor no longer. In this paper, the selected size of voxel is 0.08 m, and the original point cloud and the filtered result are shown in Fig. 2a and b respectively.

### 2.2 Ground Segmentation

We make a hypothesis that people walk on the ground in this paper. The ground plane is estimated by several points marked in advance on the ground plane. According to the estimated ground plane, we can remove the corresponding voxel point cloud, and further reduce the irrelevant data. The RANSAC segmentation algorithm is adopted to design a model for segmentation judging, and the input data is iteratively extracted by the judgment criterion. First, set thresholds for all points of the point cloud dataset, and randomly select points from the grid point cloud data and calculate the parameters of the given judgment model. If the distance between the point and the determining model is smaller than the threshold, then the point is internal, otherwise it is external out of points. The external points will be removed. As shown in Fig. 2c, the voxels of the ground plane are removed and the volume of the target image is greatly reduced.

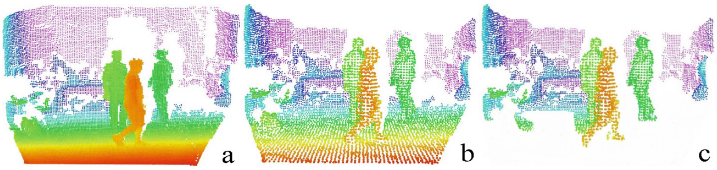

**Fig. 2.** Point cloud grid: (a) the initial point cloud; (b) point cloud after grid and down-sampling; (c) point cloud after ground segmentation

### 2.3 Preliminary Detection Based on Point Cloud

By parameterized querying the difference between point and point, PFH forms a multidimensional histogram to describe the spatial geometric attributes of the K points in neighborhood. The high-dimensional hyperspace provides measurable information space for features. The six dimensional pose of a surface is invariant and robust to different sampling densities or neighborhood noises. PFH representation is based on the relationship between the point and its K neighborhood and their estimated normal, considers the interaction between the normal directions, captures the best changes of the point cloud surface, and describes the geometric features of data samples. For example, for the query point P, the radius of its affected region is *r*, and all its K neighborhood elements are completely connected to each other in a connected region. By computing the relation between the two connected points in the connected region, the point feature histogram is the final PFH descriptor.

After removing the ground data from the point cloud, the false positive rate strategy is adopted for preliminary pedestrian detection, and the accurate detection is performed in the next step. The advantage of this approach is that it can allow as many candidate targets as possible pass this initial screening to minimize the omission ratio. First, scan the whole image with a fixed size detection window, then describe the characteristics of pedestrian cloud by using PFH, and select the candidate pedestrians using SVM classifier at last.

Kinect sensor is used to capture data in 4 different indoor scenes, including classroom, restaurant, dormitory and library. In order to make the depth information more accurate, only the data the shooting distance of which is between 3–8 m is selected. After manually tailoring the picture, 1218 pairs of positive samples and 3000 pairs of negative samples are obtained. Among them, the negative samples are randomly extracted from pictures that contain no human body. The sample size is $64 \times 128$. Finally, extract the PFH features of positive and negative samples from the depth images and input them into SVM for training.

### 2.4 Accurate Detection Based on RGB Image

The accurate pedestrian detection algorithm using HOG features is as following. A fixed size detection window is used, and the window is divided into a grid the unit of which is cell. The gradient direction of the pixels in each cell is computed into a one-dimensional histogram. The intuitive formulation is that the local appearance and shape can be well described by the distribution of local gradients without having to know the exact locations of these gradients in cell. A set of cell is aggregated into block for local contrast normalization. The histogram of all the blocks is concatenated to form the descriptor vector of the detection window. This descriptor vector is used to train the linear SVM classifier. When detecting pedestrian, the detection window slides in different scale spaces of the image and the HOG descriptors of each position and scale are calculated. Then a trained SVM classifier is used to classify, and finally the location of the pedestrian is obtained.

Compared with the depth SVM classifier trained in Sect. 2.3, the training process of the RGB SVM classifier is basically the same, but it uses HOG features. Considering the real-time application requirement, the modified HOG [14] is adopted to reduce dimensionality. Similarly, the training set constructed in Sect. 2.3 is also used for training the RGB SVM classifier.

## 3 Results

Based on the above idea, the related algorithms are implemented. The development environment is the Visual C++ platform, the software that drives the Kinect is OpenNI interface (Prime Sense Company), and the PCL is used for processing the 3D point clouds, deep and color images. The hardware platform is core duo CPU@2.2 GHz, 4 GB memory, GTX 660 graphics card and Kinect for Windows.

Experiments were carried out on 382 pairs of images containing 1201 human bodies, and the detection rate achieves an average performance of 97.75%. The experiment is divided into two types: single scene and multi person scene. The experimental results show that the pedestrian detection algorithm based on RGB-D has good detection results even in the circumstance when the pose of pedestrian is changeable in single scene. The multi person scene is divided into two cases: with occlusion and without occlusion. The algorithm has good performance in both cases. Some experimental results are shown in Fig. 3.

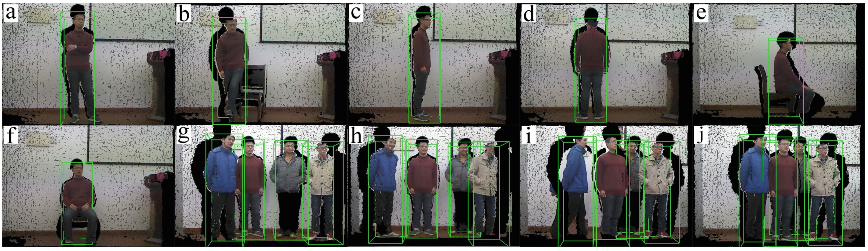

**Fig. 3.** Pedestrian detection results: (a) single scene (stand, front); (b) single scene (stand, front); (c) single scene (stand, side); (d) single scene (stand, back); (e) single scene (sit down, side); (f) single scene (sit down, front); (g) multi person scene (no occlusion); (h) multi person scene (no occlusion); (i) multi person scene (occlusion); (j) multi person scene (occlusion).

## 4 Conclusion and Future Work

A two-step pedestrian detection algorithm based RGB-D data is presented in this paper. Firstly, the algorithm captured the depth image and color image by using OpenNI SDK to drive Kinect camera. Then down-sampled the depth image and extracted the key information by voxel gridding method. After that, eliminated the ground by using the RANSAC segmentation algorithm and further screened the point cloud data independent with the target pedestrian. Next, described pedestrian characteristics by using PFH histograms and preliminary estimated the position of pedestrian. Finally, the pedestrian characteristics based on color image were further calculated by HOG descriptor and SVM classifier was used for pedestrian detection. The experimental results show that the algorithm can accurately detect the pedestrian not only in the single scene with pedestrian posture changing (stand, sit down, front, side and back) but also in the multi person scene with occlusion and without occlusion.

Although the algorithm has a high performance in a certain range of indoor applications, there are still some problems to be solved in the wide range scene applications. Limited by Kinect capability, the scope that can obtain the depth information is limited. When the distance between the target and the camera is more than 4 m, the amount of target information will sharply reduce. So, how to detect pedestrian based on RGB-D in a large range application scene is a further research direction.

**Acknowledgement.** This work is supported by Natural Science Foundation of China (Grant No. 61472245) and Shanghai Municipal Natural Science Foundation (Grant No. 14ZR1419700 and 13ZR1455600).

## References

1. Spinello, L., Arras, K.O.: People detection in RGB-D data. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS) **38**(2), 3838–3843 (2011)
2. Luber, M., Spinello, L., Arras, K.O.: People Tracking in RGB-D Data with On-line Boosted Target Models. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS) **30**(1), 3844–3849 (2011)
3. Choi, W., Pantofaru, C., Savarese, S.: Detecting and tracking people using an RGB-D camera via multiple detector fusion. In: IEEE International Conference on Computer Vision Workshops (ICCV), pp. 1076–1083 (2011)
4. Wu, S., Yu, S., Chen, W.: An attempt to pedestrian detection in depth images. In: Third Chinese Conference on Intelligent Visual Surveillance (IVS), pp. 97–100 (2011)
5. Yu, S., Wu, S., Wang, L.: Sltp: a fast descriptor for people detection in depth images. In: IEEE Ninth International Conference on Advanced Video and Signal-based Surveillance (AVSS), pp. 43–47 (2012)
6. Xia, L., Chen, C.C., Aggarwal, J.K.: Human detection using depth information by kinect. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPR), pp. 15–22 (2011)
7. Wang, N., Gong, X., Liu, J.: A new depth descriptor for pedestrian detection in RGB-D images. In: 21st International Conference on Pattern Recognition (ICPR), pp. 3688–3691 (2012)
8. Ikemura, S., Fujiyoshi, H.: Real-time human detection using relational depth similarity features. In: Kimmel, R., Klette, R., Sugimoto, A. (eds.) ACCV 2010. LNCS, vol. 6495, pp. 25–38. Springer, Heidelberg (2011). doi:10.1007/978-3-642-19282-1_3
9. Cheng, Y., Zhao, X., Huang, K., Tan, T.: Semi-supervised Learning for RGB-D Object Recognition. In: 22nd International Conference on IEEE Pattern Recognition (ICPR), pp. 2377–2382 (2014)
10. Wang, A., Lu, J., Wang, G., Cai, J., Cham, T.J.: Multi-modal unsupervised feature learning for RGB-D scene labeling. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8693, pp. 453–467. Springer, Cham (2014). doi:10.1007/978-3-319-10602-1_30
11. Gupta, S., Arbeláez, P., Girshick, R., et al.: Aligning 3D models to RGB-D images of cluttered scenes. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4731–4740 (2015)
12. Liu, J., Liu, Y., Zhang, G., et al.: Detecting and tracking people in real time with RGB-D camera. Pattern Recogn. Lett. **53**, 16–23 (2015)

13. Linder, T., Arras, K.O.: Real-time full-body human attribute classification in RGB-D using a tessellation boosting approach. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 1335–1341 (2015)
14. Maji, S., Berg, A., Malik, J.: Classification using intersection kernel support vector machines is efficient. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1–8 (2008)