Pedestrian Detection by Fusing 3D Points and Color Images

Ben-Zhong Lin and Chien-Chou Lin

Department of Computer Science and Information Engineering,
National Yunlin University of Science & Technology
123 University Road, Section 3, Douliou, Yunlin 64002, Taiwan, R.O.C.
E-mail: benlin@nvidia.com, linchien@yuntech.edu.tw

Abstract

In this paper, a fusing approach of a 3D sensor and a camera are used to improve the reliability of pedestrian detection. The proposed pedestrian detecting system adopts DBSCAN to cluster 3D points and projects the candidate clusters onto images as region of interest (ROI). Those ROIs are detected by HOG (histograms of oriented gradients) pedestrian detector. Because the DBSCAN groups together 3D points and rejects outlier points correctly, the proposed system has a low false detection rate. The performance is also improved since the proposed system only detects the ROI instead of the whole color image.

Keywords: pedestrian detection; data fusion; LIDAR; histograms of oriented gradients (HOG); multi-sensor system

1. Introduction

Recently, many approaches of pedestrian detection are proposed because the technology is widely used in many applications, e.g., surveillance system, door control system, driving assistant system and home care system. The pedestrian detection systems usually use variant perceptions which may be single sensor or multisensor systems. Generally, the acquired information from those sensors can be categorized as 2D information and 3D information. Some existing detection algorithms use only single type information to recognize pedestrians, e.g., camera. The most advantage of the color image based pedestrian detection is cheap because only a camera is needed. But, the main drawback is too false alarms caused by shadows or occlusion because of lacking depth information. Unlike the color image based pedestrian detection, the 3D information based pedestrian detection systems have more accuracy information which can be used to separate objects more exactly. But, recognition of high dimensional features takes more computation time. Furthermore, a Lidar is more expensive than a camera.

In the past few years, 3D sensors were used in pedestrian detection [1-8]. These systems proposed fusion approaches to handle multisensor information and performed a combinational pedestrian recognition. Most of the pedestrian detection algorithms process the information by three stages: selecting rough ROI, feature extraction, and recognition. This paper proposes an integrated pedestrian detection system which combines range and visual information, gathered by a single-layer LIDAR and a webcam. In order to fusing the 3D points and color image into the same coordinate system, a calibration approach [9] is adopted and reviewed. After calibration, the proposed scheme selects the candidates according to the features of clustered 3D points firstly. Then, those candidates are transformed to images as regions of interest (ROIs). Those ROIs are detected by a pedestrian detector based on histograms of oriented gradients (HOG), associated with a support vector machine (SVM).

The rest of this paper is organized as follow. In the Section II, the related works are reviewed. The proposed recognition algorithm is introduced in Section III. Finally, the simulation results and conclusions are given in Section IV and V, respectively.

Published by Atlantis Press

2. Related Works

The existing multi-sensor approaches of pedestrian detection usually use two or more different sensors to avoid mutual interference. The Lidar is often used for sensing the depth information of the environment. Basically, the Lidar has two types, single-layer Lidar and multi-layer Lidar. In single-layer LIDAR approaches, the 3D scan data on a same horizontal plane is used for detecting and segmenting objects. In [2], the range information and visual information are considered together in the CRF classification. In [3], two monocular color cameras for line and vehicle detection and a LIDAR for tracking are used. A local and a global tracking approaches were proposed for on-road object detection. These approach provided ROIs for visual detection system.

In multi-layer LIDAR approaches, the 3D scan data of the whole environment can obtained to provides more features for recognition. Thus, the multi-layer information is not only to provide the candidates of objects, but also their features are meaningful for object recognition. However, the main drawback is time-consuming because high dimensional features take more time to classify. In [4], Spinello and Siegwart used a multi-layer LIDAR to detect the object's position and the HOG-SVM classifier based on monocular color images to classify the detected objects as pedestrian or nonpedestrian. In [5], a method based on tracking and vision score–based likelihood is proposed. In [6], the approach used a convolutional NN classifier based on monocular gray-scale images to detect pedestrians.

3. DBSCAN Clustering and the Proposed Algorithm

Since multi-layer LIDAR based pedestrian detection approaches are more computational complex, they are not suitable for real-time applications, such as, driving assistant system. Therefore, in this paper, the single-layer LIDAR is adopted for the time criteria of driving assistant system in the real world. As shown in Fig. 1, the multi-sensor system consists of a single-layer Lidar and a webcam.

In the proposed system, fusion of 3D points and color images is a challenge because an accurate extrinsic calibration of the transformation between the two sensors is required. Several methods for calibrating a single-layer Lidar with respect to a camera have been proposed [9-16]. In this paper, the method proposed in [9] is adopted to calibrate LIDAR-Camera system. The detail is reviewed as follows.

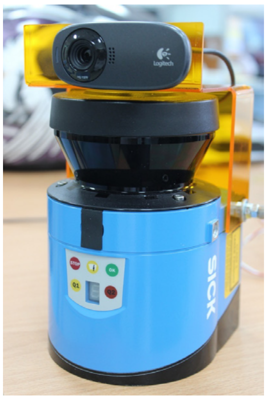


Fig. 1. An integrated multisensor system consists of a 2D laser scanner and a commercial webcam.

Let $P_w = (X_w, Y_w, Z_w)^t$ denote points measured in 3D by the single-layer Lidar. $p_c = (x_{im}, y_{im})^t$ denote pixels locations in the color image and T is the position of the camera optical center. Assuming the Lidar origin point is $(X_w, Y_w, Z_w)=(0, 0, 0)$ in the global coordinate. Furthermore, since the Lidar scans XZ-plane only, it is assumed that $Y_w=0$ for all points acquired from LIDAR. And, since it is difficult to measure T, it is assumed that $T_y = Y_o$. Therefore, the pixel location can be represented as

$$x'_{im} = x_{im} - o_x = \frac{f}{s_x} \frac{r_{11}x_w + r_{13}z_w - T_x}{r_{31}x_w + r_{33}z_w - T_z}, \qquad (1a)$$

$$y'_{im} = y_{im} - o_y = \frac{f}{s_y} \frac{Y_0}{r_{31}x_w + r_{33}z_w - T_z} \qquad (1b)$$

where f is the focal length of the camera, (o_x, o_y) is the position of image center and (s_x, s_y) is the physical size of a single pixel in the image sensor. The r_i denotes the row vector formed by the i-th row of the rotation matrix that determines the camera orientation.

Since there are six unknowns in (1a) and (1b), at least six equations are needed to solve the transformation. In other words, six pairs of corresponding pairs between scan data of LIDAR and 2D image pixels are needed. A zigzag calibration pattern can used to provide such corresponding pairs between two sensors. If there are i corresponding pairs are found, the linear system can be expressed as $AX = 0$ where

$$A=\begin{bmatrix} r_{31}x_{w\_1}x'_{im\_1} & r_{33}z_{w\_1}x'_{im\_1} & -T_z x'_{im\_1} & -f_x r_{11}x_{w\_1} & -f_x r_{13}z_{w\_1} & 1 \\ & & \vdots & & & \\ & & & & & \\ & & \vdots & & & \\ r_{31}x_{wi}x'_{im\_i} & r_{33}z_{w\_i}x'_{im\_i} & -T_z x'_{im\_i} & -f_x r_{11}x_{w\_i} & -f_x r_{13}z_{w\_i} & 1 \end{bmatrix} \quad (2)$$

and

$$X=\begin{bmatrix} r_{31} & r_{33} & T_z & f_x r_{11} & f_x r_{13} & f_x T_x \end{bmatrix}. \quad (3)$$

Then, the proximate values of parameters can be derived by SVD. Since it is assumed that $T_y=Y_o$, only x_m is needed to be solve and y_m is defined as the half of image height.

$$y_m=\frac{H_{im}}{2} \quad (4)$$

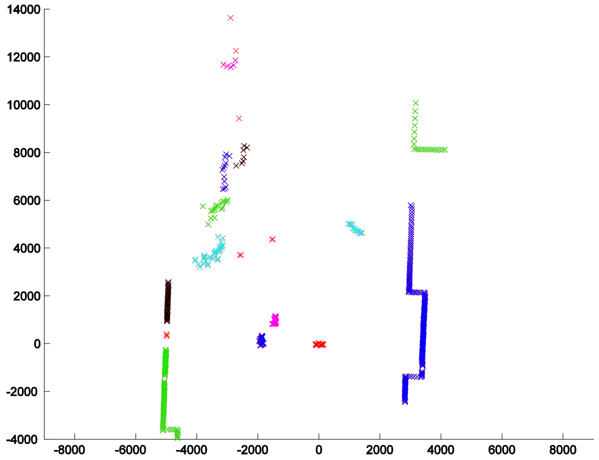


Fig. 2. Points are clustered into several clusters marked in different colors by DBSCAN. The noise points are marked in red.

3.1. *The steps of the proposed system*

The proposed recognition approach has 5 main steps. The steps are described as fallows.

(i) *Acquiring 3D points and color images by LIDAR and webcam.*

(ii) *Clustering the 3D points by DBSCAN. Then, selecting the candidate clusters.*

(iii) *The candidates of 3D points are mapped to the candidate pixels of the image and the rectangles including these pixels are selected as the ROIs.*

(iv) *HOG descriptors are extracted within the ROIs and a trained SVM classifier is used to detecting pedestrians.*

The details of the Step (ii), (iii) and (iv) are described next.

3.2. *DBSCAN Clustering*

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) [17] is adopted to groups together Lidar points with many nearby neighbors and to mark whose nearest neighbors are too far away as outliers points.

Considering a set of points in some space to be clustered, DBSCAN categorizes the points into three types: core points, density-reachable points and outliers. Core points are the point has at least minimum neighbors within a predefined distance. A density-reachable point is a point whose neighbor is a core point (or core points), but less neighbors to be a core point. If points are far away core points and density-reachable points, those points are marked as outlier points.

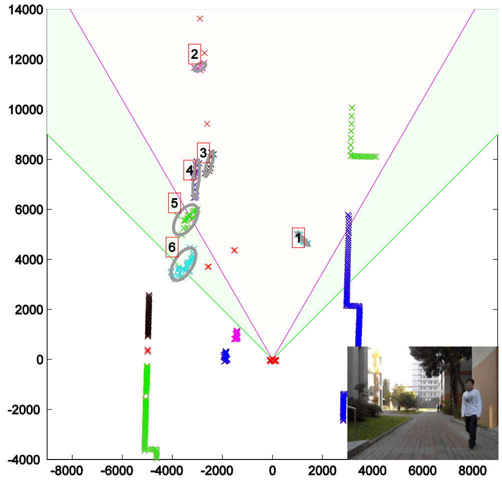


Fig. 3. The results of candidate selection are marked by gray circles. The noise points are colored with red.

Therefore, DBSCAN requires two parameters: the distance restriction of a cluster, ε, and the minimum number of points required to form a cluster, *minPts*. The algorithm starts with an arbitrary unvisited point. If a point contains sufficiently many neighbors within the distance ε, a new cluster is started. Otherwise, the point is labeled as outlier. However, the outlier may be a part of a cluster later if it is a density-reachable point of a cluster. If a new point is added into a cluster, its neighbors within the distance ε are also added into the same cluster. This process is executed iteratively until the whole cluster is completely found. Then, a new unvisited point is selected arbitrarily as a start point of the DBSCAN algorithm. In Fig. 2, the points are clustered into several clusters with different colors by DBSCAN. The noise points are marked in red.

3.3. Candidate Selection and ROI

After the DBSCAN clustering process, points can be separated as noises and clusters. The sizes of clusters, in terms of number of points, are varying. However, the widths of the target objects, pedestrians, are within a reasonable range. In general, since the width of a human body is within 40 cm to 100 cm, the clusters will be selected under the size criteria of cluster. The number of points of a potential cluster can be defined as

$$\Pi = \frac{width\ of\ pedestrian}{d \times \tan\theta} + \Delta t \tag{5}$$

where d is the measured distance of the points and θ is the angle resolution of the Lidar. The Δt is a tolerance. The angle resolution of the Lidar used in this paper is 0.5 degree. Thus, if the number of points of a cluster is within Π_{40} to Π_{100} , the cluster is considered as a candidate. Considering a candidate cluster, P_i, it must satisfy the following rules.

$$\begin{cases} P_i \in C, if\ \ \Pi_{40} \le \dfrac{width\ of\ P_i}{d_i \times \tan\theta} \le \Pi_{100} \\ P_i \notin C, if\ \ \dfrac{width\ of\ P_i}{d_i \times \tan\theta} > \Pi_{100} \vee \dfrac{width\ of\ P_i}{d_i \times \tan\theta} < \Pi_{100} \end{cases} \tag{6}$$

These candidates are mapped into 2D image by the transform matrix derived in (1a) and (1b) next. The result of candidate selection of Fig. 2 is shown in Fig. 3. The FOV of the color image is between two pink lines and the extended invisible areas which can be detected by laser scanner are marked as green area.

This step improves the performance of the pedestrian detection significantly because it extracts the ROIs of 2D image. While the conventional HOG-base pedestrian detection algorithm is searching the whole image, the proposed algorithm is detecting those ROIs. The details of ROI selection and HOG-base pedestrian detection algorithm are introduced next.

Because the candidate clusters are transformed to a horizontal line segment of 2D image pixels by (1a) and (1b), only the width of a ROI is available. Generally, in a 2D image, a pedestrian area is a rectangle with a fixed aspect ratio. In this paper, the adopted aspect ratio, 0.5, is obtained by analyzing the INRIA Person dataset. All points of a candidate are projected onto points, *(x, y)*, of the color image by (1a) and (1b). If the maximal and minimal x-coordinates of these pixels are $x_v^{\min}$ and $x_v^{\max}$ respectively, the ROI can be defined as

$$ROI(x,y) = \left\{ (x,y) \in S_i,\ \begin{matrix} x_v^{\min} - \Delta d \le x \le x_v^{\max} + \Delta d, \\ |y - y_v| \le 2\left(x_v^{\max} - x_v^{\min}\right) \end{matrix} \right\} \tag{7}$$

3.4. HOG Based Pedestrian Detection by SVM classifier

In this step, the obtained ROIs are checked by HOG based pedestrian detection algorithm. Histogram of Oriented Gradients (HOG) [18-20] is a feature descriptor which counts occurrences of gradient orientation in localized portions of an image. HOG is computed on a dense grid of uniformly spaced cells and uses overlapping local contrast normalization for improved accuracy. In [18], the object is modeled as local shape and appearance using well-normalized dense histograms of gradient orientation (HOG). Local gradients are binned according to their orientation, weighted by their magnitude, within a spatial grid of cells with overlapping block-wise contrast normalization. Within each overlapping block, a feature vector is extracted by sampling the histograms from the contributing spatial cells. The feature vectors for all blocks are concatenated to yield a final feature vector which is subject to classification using a linear support vector machine (linSVM) [21].

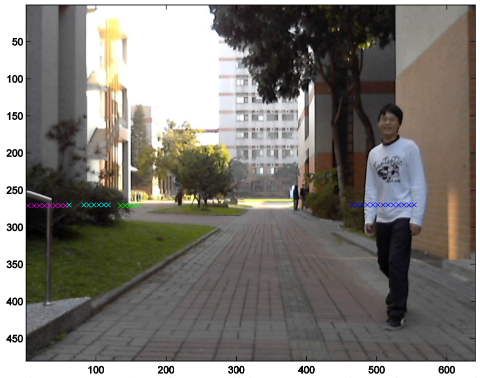


Fig. 4. Points, (X,Y,Z), of a candidate are projected onto points, (X, Y), of the 2D image.

In this paper, all the training samples are scaled to and utilized at a fine resolution of 64 × 128 pixels. The training stage utilizes gradients with [−1, 0, 1] mask, 9 orientation bins, spatial binning (2 × 2 blocks of 8 × 8 pixel cells) as well as overlapping block contrast normalization (L2-norm). The training dataset is INRIA Person dataset [21], which were used in [18]. In order to prevent mass false positives, the proportion of positive and negative samples is 1:2 when training the desire model.

4. Simulation Results

In the section, simulation results are presented for pedestrian detection performed on a notebook with an Intel Core i5 450M 2.4GHz CPU running the Windows

XP SP3 operating system and programming on MS Visual C++ 2008 with OpenCV 2.2 library. Training dataset of HOG-SVM is the INRIA Person dataset [22]. Test data includes 3 types of pedestrians: front-view pedestrian, side-view pedestrian, back-view pedestrian. The first test scenario is that people walks randomly within 3 M to 15 M distance in 797 continuous frames which consist of color image and 3D point set.

Partial results are shown in Figure 5. The results of the proposed algorithm are in left-hand side and the results of Dalal's algorithm are in right-hand side. In Fig. 5(b), Dalal's algorithm has false alarm which is a trees. The proposed algorithm can identify the pedestrian with partial occlusion as shown in Fig. 5(c), but Dalal's algorithm [18] considers the two people as one as shown in Fig. 5(d).

In Tab. 1, a comparison of two methods is given. Since some pedestrians with less than three LIDAR points are ignored by the proposed algorithm, the hit rate of the proposed algorithm is less than the Dalal's algorithm. However, the false alarm rate is reduced significantly with the proposed algorithm. Furthermore, the computation time of the proposed algorithm is also reduced because the proposed algorithm only detects the ROIs not the whole image.

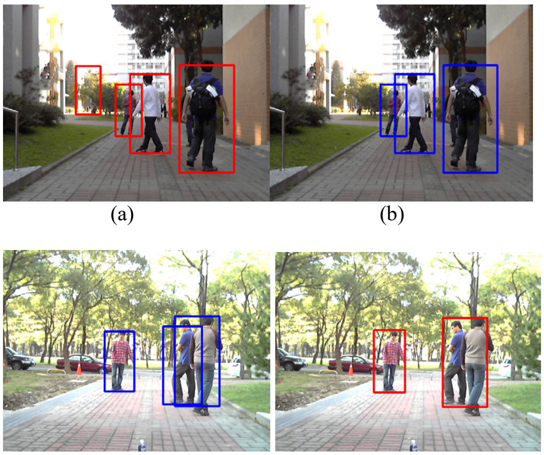
Table 1. Comparisons of accuracy and performance

	*Hit Rate	**False Alarm Rate	Time/797 frames (sec)
The proposed method	71.7% (1351/1885)	18.3% (304/1655)	578
Dalal [17]	75.6% (1425/1885)	40.9% (989/2414)	985

*Hit Rate: True Positive Alarm/Ground Truth
**False Alarm Rate: False Alarm/Total Alarm

5. Conclusions

In this paper, a fusing approach of a 3D sensor and a camera are used to improve the reliability of pedestrian detection. The false alarm rate is reduced significantly. However, since some pedestrians with less than three Lidar points are ignored in our approach, the detection rate is decreased a little bit. Furthermore, the proposed algorithm is also more efficient because it only detects the ROIs instead of the whole image. Every frame can be detected within 0.75 second. In the experimental results, the true positive rate (TPR) of the proposed system is up to 71.7% and the false positive rate is 13%.



(a) (b)

(c) (d)

Fig. 5. The results of the proposed algorithm are in left-hand side and the results of Dalal's algorithm [18] are in right-hand side.

References

1. Muhammad Arshad Awan, Zheng Guangbin, Cheong-Ghil Kim, Shin-Dug Kim, Human Activity Recognition in WSN: A Comparative Study. *International Journal of Networked and Distributed Computing*, Vol. 2, No. 4, pp. 221-230, 2014.
2. Douillard, B., Fox, D., Ramos, F., A spatiotemporal probabilistic model for multi-sensor object recognition. *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 2402–2408, 2007 (IROS 2007), San Diego, CA.
3. Cheng, H., Zheng, N., Zhang, X., Qin, J., & van de Wetering, H., Interactive road situation analysis for driver assistance and safety warning systems: Framework and algorithms. *IEEE Transactions on Intelligent Transportation Systems*, Vol.8, No.1, 157–167, 2007.
4. Spinello, L., & Siegwart, R., Human detection using multimodal and multidimensional features. *Proceedings of the IEEE International Conference on Robotics and Automation*, pp. 3264– 3269, 2008 (ICRA 2008), Pasadena, CA.
5. Pangop, L. N., Chapuis, R., Bonnet, S., Cornou, S., & Chausse, F., A Bayesian multisensory fusion approach integrating correlated data applied to a real-time pedestrian detection system. *Proceedings of the IEEE Workshop on Perception, Planning and Navigation for Intelligent Vehicles*, Nice, France, 2008.
6. Szarvas, M., Sakai, U., & Ogata, J., Realtime pedestrian detection using lidar and convolutional neural networks. *Proceedings of the 2006 IEEE Intelligent Vehicles Symposium*, pp. 213–218, Tokyo, 2006.
7. Cristiano Premebida, Oswaldo Ludwig, and Urbano Nunes, LIDAR and Vision-Based Pedestrian Detection System, *Journal of Field Robotics* Vol. 26, No. 9, pp.696–711, 2009.

8. K. Kidono, T. Miyasaka, A. Watanabe, T. Naito, and J. Miura, Pedestrian recognition using high-definition LIDAR, *Proceedings of the IEEE Symposium on Intelligent Vehicles* (IV2011), pp. 405-410, 2011.
9. Andrew R. Willis, Malcolm J. Zapata, James M. Conrad., A linear method for calibrating LIDAR-and-camera systems, *Proceedings of the IEEE International Symposium on Modeling, Analysis & Simulation of Computer and Telecommunication Systems*, pp. 1-3, Sept. 2009.
10. C. H. Chen and A. C. Kak, Modelling and calibration of a structured light scanner for 3D robot vision, *Proceedings of the IEEE International Conference on Robotics and Automation*, vol. 2, pp. 807-815, 1987.
11. V. Niola, C. Rossi, S. Savino, and S. Strano, A method for the calibration of a 3-D laser scanner, *Robotics and Computer-Integrated Manufacturing*, vol. 27, no. 2, pp. 479-484, 2011.
12. K. Kwak, D. Huber, J. Chae and T. Kanade, Boundary detection based on supervised learning, *Proceedings of the IEEE International Conference on Robotics and Automation*, pp. 3939-3945, 2010.
13. Q. Zhang and R. Pless, Extrinsic calibration of a camera and laser range finder (improves camera calibration), *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, Vol. 3, pp. 2301-2306, 2004.
14. G. Li, Y. Liu, L. Dong, X. Cai, and D. Zhou, An algorithm for extrinsic parameters calibration of a camera and a laser range finder using line features, *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 3854-3859, 2007.
15. S. Wasielewski and O. Strauss, Calibration of a multi-sensor system laser rangefinder/camera, *Proceedings of the IEEE Symposium Intelligent Vehicles*, pp. 472-477, 1995.
16. Kwak, K., Huber, D. F., Badino, H., & Kanade, T., Extrinsic calibration of a single line scanning lidar and a camera. *Proceedings of the 2011 IEEE/RSJ International Conference on Intelligent Robots and Systems* (IROS), pp. 3283-3289, 2011.
17. Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, A density-based algorithm for discovering clusters in large spatial databases with noise. *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, AAAI Press. pp. 226–231, 1996.
18. N. Dalal and B. Triggs, Histograms of oriented gradients for human detection, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (CVPR 2005), Vol. 1, pp. 886-893, 2005.
19. Suard, F., Rakotomamonjy, A., Bensrhair, A., & Broggi, A. (2006, June). Pedestrian detection using infrared images and histograms of oriented gradients. *Proceedings of the 2006 IEEE Symposium on Intelligent Vehicles*, pp. 206-212, 2006.
20. Ryosuke Yamanishi, Ryoya Fujimoto, Yuji Iwahori and Robert J.Woodham, Hybrid Approach of Ontology and Image Clustering for Automatic Generation of Hierarchic Image Database, *International Journal of Networked and Distributed Computing*, Vol. 3, No. 4, 234-242, 2015.
21. C. Cortes and V. Vapnik, Support-vector networks, *Machine Learning*, vol. 20, pp. 273-297, 1995.
22. Dalal, N., & Triggs, B. INRIA person dataset. Online: http://pascal. inrialpes. fr/data/human, 2005.