Pedestrian Detection and Tracking Using Three-Dimensional LADAR Data

Luis E. Navarro-Serment, Christoph Mertz, and Martial Hebert

Abstract

The approach investigated in this work employs three-dimensional LADAR measurements to detect and track pedestrians over time. The sensor is employed on a moving vehicle. The algorithm quickly detects the objects which have the potential of being humans using a subset of these points, and then classifies each object using statistical pattern recognition techniques. The algorithm uses geometric and motion features to recognize human signatures. The perceptual capabilities described form the basis for safe and robust navigation in autonomous vehicles, necessary to safeguard pedestrians operating in the vicinity of a moving robotic vehicle.

1 INTRODUCTION

The ability to avoid colliding with other objects is essential in autonomous vehicles, especially in cases where they operate in close proximity to people. The timely detection of a pedestrian makes the vehicle aware of a potential danger in its vicinity, and allows it to modify its course accordingly. There is a large body of work done using laser line scanners as the primary sensor for pedestrian detection and tracking. In our group, we have developed detection and tracking systems using SICK^{TM} laser line scanners; these implementations work well in situations where the ground is relatively flat [5]. However, a 3D LADAR (i.e. one who produces a set of 3D points, or point cloud) captures a more complete representation of the environment and the objects within it. In [6], we presented an algorithm that detects pedestrians from 3D

The Robotics Institute, Carnegie Mellon University, Pittsburgh, PA 15213, e-mail: lenscmu@ri.cmu.edu, cmertz@andrew.cmu.edu, hebert@ri.cmu.edu

data. Its main improvement over the version with 2D data was that it constructs a ground elevation map, and uses it to eliminate ground returns. This allows pedestrian detection even when the surrounding ground is uneven. To classify the humans the algorithm uses motion, size, and noise features. Persons are classified well as long as they are moving. However, there are still too many false positives when classifying stationary humans.

In this paper, we describe a strategy to detect and classify humans using the full 3D point cloud of the object. This will improve the classification of both moving and static pedestrians. However, the improvement will be most significant for static humans. The algorithm quickly detects the objects that have the potential of being humans using a subset of the point cloud, and then classifies each object using statistical pattern recognition techniques. We present experimental results of detection performance using 3D LADAR, which were obtained from field tests performed on a Demo III XUV [7].

2 RELATED WORK

Some researchers have applied classification techniques to the detection and tracking problem. The approach reported in [1] applies AdaBoost to train a strong classifier from simple features of groups of neighboring points. This work focuses on 2D range measurements. Examples using three-dimensional data include [4], where 3D scans are automatically clustered into objects and modeled using a surface density function. A Bhattacharya similarity measure is optimized to register subsequent views of each object enabling good discrimination and tracking, and hence detection of moving objects. In [3], the authors describe a pedestrian detection system which uses stereo vision to produce a 3D point cloud, and then classifies the cloud according to the point shape distribution considering the first two central moments of the 2D projections using a naive Bayes classifier. Motion is also used as a cue for human detection.

In [8] the authors report an algorithm capable of detecting both stationary and moving humans. Their approach uses multi-sensor modalities including 3D LADAR and long wave infrared video (LWIR). Similarly, in [9] the same research group presents a technique for detecting humans that combines the use of 3D LADAR and visible spectrum imagery. In both efforts the authors employ a 2D template to extract features from the shape of an object. Among other differences, as opposed to our work, they extract a shape template from the projection in only one plane, and compute a measure of how uniformly distributed the returns are across the template.

3 ALGORITHM DESCRIPTION

In this section, the algorithm for pedestrian detection and classification is described. In our approach, since operation in real time is a chief concern, we do object detection and tracking in a 2D data subset first, and then use the object's position and size information to partition the set of 3D measurements into smaller groups, for further analysis. We describe these steps in the following sections.

3.1 Projection into 2D Plane

To reduce the computational cost of processing the entire point cloud, we initially isolate a 2D virtual slice, which contains only points located at a certain height above ground. As shown in Fig. 1, a 3D scanner produces a point cloud, from which a "slice" is projected onto the 2D plane, resulting in a virtual scan line. This scan line is a vector of range measurements coming from consecutive bearings, which resembles the kind of data reported directly by a line scanner such as the SICK^{TM} laser scanner. This is done by collapsing into the plane all the points residing within the slice, which is defined by its height above the ground.

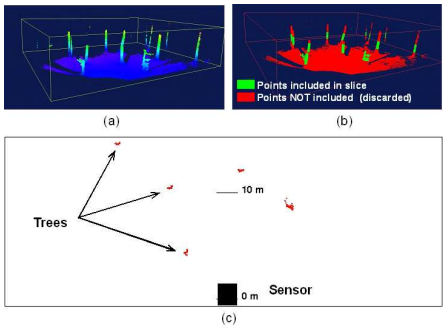

Fig. 1 Projection of virtual scan line. (a) A point cloud is collected from the environment shown. (b) The points located within a certain height above ground are projected into a 2D plane, and processed as if it were a single scan line. The resulting projection is shown in (c), top view

The ground elevation is stored in a scrolling grid that contains accumulated LADAR points and is centered at the vehicle's current position. The points are weighted by age, more recent points have a heigher weight. The mean and standard deviation of the heights of all scan points that are inside each cell are computed, and the elevation is then calculated by subtracting one standard deviation from the average height of all the points in the cell. The key properties of this simple algorithm are that mean and standard deviations can be calculated recursively, and that the elevation is never below the lowest point while still having about 80% of the points above ground.

The system adapts to different environments by varying the shape of the sensing plane i.e., by adjusting the height of the slice from which points are projected onto a two-dimensional plane. Spurious measurements produced

by ground returns are avoided by searching for measurements at a constant height above the ground. Since our research was done in an open outdoor environment, we did not encounter overhanging structures like overpaths or ceilings. These might be topics of future research.

3.2 Motion Features

After detecting and tracking objects using the virtual scan line we can compute a *Motion Score* (MS). The MS is a measure of how confident the algorithm is that the detected object is a human, based on four motion-related variables: the object's size, the distance it has traveled, and the variations in the object's size and velocity. The size test discriminates against large objects like cars and walls. The distance traveled test discriminates against stationary objects like barrels and posts. The variation tests discriminate against vegetation, since their appearance changes a lot due to their porous and flexible nature. The individual results of these tests are scored, and then used to calculate the MS. A detailed description of each test and all parameters involved is presented in [6].

3.3 Geometric Features

To discriminate against static structures, we compute a group of distinguishing geometric features from the set of points belonging to each object being tracked in 2D, and then feed these features to a classifier, which determines whether the object is a human or not. This concept is depicted in Fig. 2.

As shown in Fig. 2(a), the process starts when a point cloud is read from the sensor. We define $Z_j = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ as the set of N points contained in a frame collected at time t_j, whose elements are represented by Cartesian coordinates $\mathbf{x} = (x, y, z)$. The points corresponding to one frame are shown, and are colored according to their height above ground. To avoid the computational cost of processing the entire point cloud, we extract a 2D virtual slice, as described in Section 3.1 (Fig. 2(b)). For each object being tracked, its position, velocity, and size are estimated using the algorithm described in [6]. These values are used to compute the MS. The object's position and size information are used to isolate, from the original point cloud, only those points corresponding to potential humans, as shown in Fig. 2(c). In this way, the three-dimensional information corresponding to each object is recovered in the form of smaller sets of points. At this point, we have a collection of M sets $\{S_1, S_2, \ldots, S_M\}$, where $S_{i \in \{1,2,\ldots,M\}} \subset Z_j$. A feature vector is computed from each of these sets (Figs. 2(d) - (e)), and then fed to a classifier that determines for each object whether it is a human or not, Fig. 2(f). This decision

is made for each object, and is based on the most recent set of points collected from the sensor. The classifier also takes into account the information used to calculate the MS; this is described in a subsequent section.

A set of features is computed with the purpose of extracting the most informative signatures of a human in an upright posture from the 3D data. The legs are particularly distinctive of the human figure, so the algorithm computes statistical descriptions from points located around the legs. Similar descriptions are computed from the trunk area, representing the upper body. Additionally, the moment of inertia tensor is used to capture the overall distribution of all points. Finally, to include the general shape of the human figure, we compute the normalized 2D histograms on two planes aligned with the gravity vector.

3.3.1 Feature Extraction

Let $S_k = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ be the set of points belonging to the object k, whose elements are represented by Cartesian coordinates $\mathbf{x} = (x, y, z)$. A set of suitable features is computed from S_k, as depicted in Fig. 2(d), which constitutes a profile of the object.

We begin by performing Principal Component Analysis (PCA) using all the elements of S_k, to identify the statistical patterns in the three-dimensional data (see Fig. 3). This involves the subtraction of the mean $\mathbf{m}$ from each of the three data dimensions. From this new data set with zero mean, we calculate the covariance matrix $\mathbf{\Sigma} \in \Re^{3\times3}$, and the normalized moment of inertia tensor $\mathbf{M} \in \Re^{3\times3}$, treating all points as unit point masses:

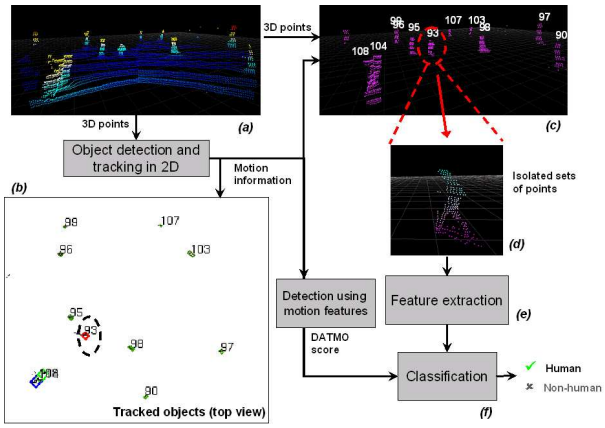

Fig. 2 Improved pedestrian detection. Geometric features present in subsets of the point cloud are used by a classifier to distinguish pedestrians from static objects.

$$\mathbf{\Sigma} = \frac{1}{n-1}\sum_{k=1}^{n}(\mathbf{x}_k - \mathbf{m})(\mathbf{x}_k - \mathbf{m})^T$$

$$\mathbf{M} = \begin{bmatrix} \sum_{k=1}^{n}(y_k^2 + z_k^2) & -\sum_{k=1}^{n} x_k y_k & -\sum_{k=1}^{n} x_k z_k \\ -\sum_{k=1}^{n} x_k y_k & \sum_{k=1}^{n}(x_k^2 + z_k^2) & -\sum_{k=1}^{n} y_k z_k \\ -\sum_{k=1}^{n} x_k z_k & -\sum_{k=1}^{n} y_k z_k & \sum_{k=1}^{n}(x_k^2 + y_k^2) \end{bmatrix}$$

Since both $\mathbf{\Sigma}$ and $\mathbf{M}$ are symmetric, we only use 6 elements from each as features.

Resulting from the PCA are three pairs of eigenvectors and eigenvalues, sorted according to decreasing eigenvalue. Call these eigenvectors $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3$, with their corresponding eigenvalues $\lambda_1 > \lambda_2 > \lambda_3$. We assume that a pedestrian is in an upright position, so the principal component $\mathbf{e}_1$ is expected to be vertically aligned with the person's body[1]. Together with the second largest component $\mathbf{e}_2$, it forms the main plane (Fig.3, center top), and also forms the secondary plane with the smallest component, $\mathbf{e}_3$ (Fig.3, center bottom). We then transform the original data into two representations using each pair of components $\mathbf{e}_1, \mathbf{e}_2$ and $\mathbf{e}_1, \mathbf{e}_3$, from which we proceed to compute additional features (the third possible representation, i.e. using the two smallest components $\mathbf{e}_2, \mathbf{e}_3$, is not used).

We focus on the points included in the main plane, to analyze the patterns that would correspond to the legs and trunk of a pedestrian, as shown in Fig 3, center top. These zones are the upper half, and the left and right lower halves. After separating the points into these zones, we calculate the covariance matrix (in 2D) over the transformed points laying inside each zone. This results in 9 additional features (3 unique values from each zone).

Finally, we compute the normalized 2D histograms for each of the two principal planes (Fig. 3, right), to capture the shape of the object. We use 14×7 bins for the main plane, and 9×5 for the secondary plane. Each bin is used as a feature, so there are 143 features representing the shape. A total of 164 geometric features are determined for each object.

3.4 Human Detection

A classifier (Fig. 2(f)), composed of two independent Support Vector Machines (SVM) [2], determines for each object whether it is human or not. The first classifier is a SVM that receives the vector of 164 geometric features computed directly from S_k, and scores how closely the set matches a human

[1] Dealing with the violation of this assumption is the focus of current research

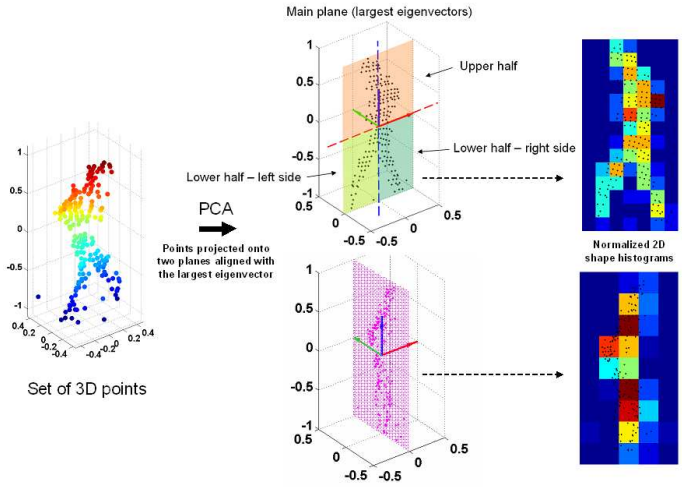

Fig. 3 A set of geometric features is computed from a the set of points belonging to an object.

shape. We call this the *Geometric Score* (GS). The GS is particularly effective for detecting static pedestrians. Similarly, the features used to determine the MS (i.e. object's size, the distance it has traveled, and the corresponding size and motion noises) contain valuable information about the motion of the target. Together with the GS, these features are fed to a second SVM, whose output represents the distance to the decision surface of the SVM. The *Strength of Detection* (SOD), the total measure of how strongly the algorithm rates the object as being a human, is calculated as the logistic function of the distance to the decision surface. This number is reported for each object. If the GS cannot be computed (e.g. insufficient data from a distant target, or violation of the upright position assumption), then the MS is reported as the SOD for that object.

3.4.1 Training

We trained the GS classifier using a combination of simulated and real examples. Because it is impossible to collect enough real data to evaluate perception algorithms in all possible situations, we have created a simulator capable of producing synthetic examples of sensor data. The simulator uses a ray tracing engine to generate a set of ray intersections between sensor and the objects in the scene to simulate. This information is then used to produce synthetic LADAR measurements according to a set of parameters for a particular sensor, as shown in Fig. 4. We trained the GS classifier using over 3500 examples (27.4% humans, 72.6% non-humans). The human set included

62% of simulated examples. The second classifier was trained using only real examples, since the motion and size noises used to determine the MS are of a dynamic nature and consequently harder to simulate efficiently (over 46000 examples: 6% humans, 94% non-humans).

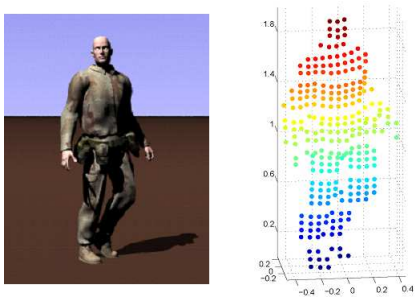

Fig. 4 Simulated target (left), and its corresponding point cloud (right).

We trained both SVMs using a five-fold cross validation procedure. We found that both radial basis function (RBF) and polynomial kernels resulted in similar levels of classification performance. After multiple tests, we determined that a RBF kernel was the best for the calculation of the GS, while a polynomial kernel was preferred for the second classifier.

4 EXPERIMENTAL RESULTS

This section presents the results of several experimental runs. These results were obtained from field tests performed on a Demo III XUV [7]. The data comes from 14 different runs, where the variations include static and moving vehicles, pavement and off-road driving, and pedestrians standing, walking, or jogging. The data was taken at 17 Hz, and the average duration of each run was about 1 minute. There were altogether 48 humans and 1075 non-human objects, where those who came in and out of the field-of-view were counted twice. The ground truth was produced by labelling the data by hand.

In the upper part of Fig. 5 the ROC curve and the precision-recall curves are shown. Each human in one cycle is a positive example and each non-human object in one cycle is a negative example. There are about 6300 positive and 60000 negative examples. These plots illustrate the current performance of our system. The blue traces indicate the MS score, which is our previous detection algorithm. The red traces indicate the geometric score, i.e. the classification using the geometric features computed directly from the object's set of 3D points, but without any motion clues. As seen in the plots, neither algorithm by itself outperforms the other throughout the entire operational range. For low false positive rates the GS is better and at high false positive rates MS is better. As we mentioned earlier, the MS only works for static humans at high false positive rates. The synergistic combination of both results has significantly better performance, as indicated by the black traces.

An alternative representation of ROC and precision-recall is shown in the lower part of Fig. 5, where each object is counted per run. The score of an

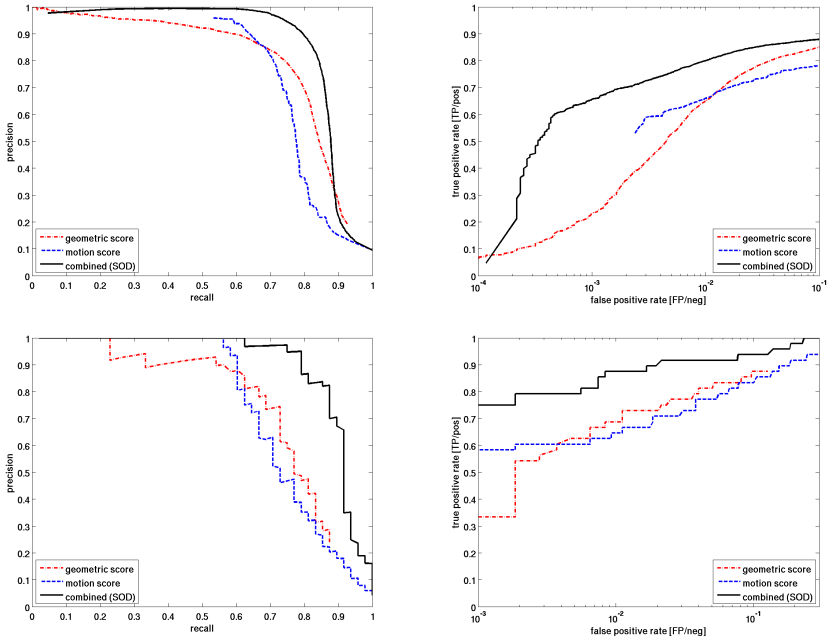

Fig. 5 The plots on the left show the precision-recall and the ones on the right the ROC curves. Shown are the curves produced by using the geometric score (red dashed), the motion score (blue dot-dash) and the combined (black solid). For the upper curves each object in each cycle and in the lower curves each object in the full run is counted as one example.

object is the mean of the score over all cycles, with a minimum of 10 cycles. As mentioned above, there are 48 humans (= positive examples) and 1075 other objects (negative examples). A noteworthy operating point is where there are basically no false positives (rate is 10^{-3}) and still the true positive rate is 0.75.

We identified several cases where performance is decreased. Typically, partial views of a human (e.g. only the upper body and part of the legs are seen by the sensor) result in false negative detections. Also, false detections occur when a target is only partially inside the sensor's field of view. Similarly, pedestrians in non-upright positions usually result in false negative detections. This is expected, because of the particular attention paid to legs and torso when extracting geometric features. An exception to this is the case of kneeling humans, which we have been able to detect consistently, though they are usually borderline classified as humans. Solving these problems is the focus of current research.

5 CONCLUSION

We described a pedestrian detection and tracking system using only three-dimensional data. The approach uses geometric and motion features to recognize human signatures, and clearly improves the detection performance achieved in our previous work. The set of features used to determine the human and motion scores was designed to detect humans in upright positions. To increase the robustness of detection of humans in other postures, in future research we will investigate ways of extracting signatures from the point cloud that are highly invariant to deformations of the human body.

6 ACKNOWLEDGMENT

We thank General Dynamics Robotic Systems for their support. This work was conducted through collaborative participation in the Robotics Consortium sponsored by the U. S. Army Research Laboratory under the Collaborative Technology Alliance Program, Coop. Agreement DAAD19-01-209912.

References

1. K. O. Arras, O. M. Mozos, and W. Burgard. Using Boosted Features for the Detection of People in 2D Range Data. In *Proc. of the 2007 IEEE Int. Conf. on Robotics and Automation*, pp. 3402-3407, Roma, Italy, 10-14 April, 2007.
2. C. J. C. Burges. A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery*, 2, 121-167, 1998. Kluwer Academic Pub. Boston.
3. A. Howard, L. H. Matthies, A. Huertas, M. Bajracharya, and A. Rankin. Detecting Pedestrians with Stereo Vision: Safe Operation of Autonomous Ground Vehicles in Dynamic Environments. *Proc. of the 13th. International Symposium of Robotics Research*, November 26-29, 2007.
4. D. Morris, B. Colonna, and P. Haley. Ladar-based Mover Detection from Moving Vehicles. In *Proc. of the 25th Army Science Conference*, November 2006.
5. L.E. Navarro-Serment, C. Mertz, and M. Hebert. Predictive Mover Detection and Tracking in Cluttered Environments. *Proc. of the 25th. Army Science Conference*, November 27-30, 2006.
6. L. E. Navarro-Serment, C. Mertz, N. Vandapel, and M. Hebert. LADAR-based Pedestrian Detection and Tracking. *1st. IEEE Workshop on Human Detection from Mobile Platforms,* Pasadena, California, May 20th. 2008.
7. C.M. Shoemaker and J. A. Bornstein. The Demo III UGV Program: a Testbed for Autonomous Navigation Research. *Proc. of the IEEE Int. Symposium on Intelligent Control*, Gaithersburg, MD, September 1998, pp. 644-651.
8. S. Thornton, M. Hoffelder, and D. Morris. Multi-sensor Detection and Tracking of Humans for Safe Operations with Unmanned Ground Vehicles. *1st. IEEE Workshop on Human Detection from Mobile Platforms,* Pasadena, California, May 20th. 2008.
9. S. Thornton and R. Patil. Robust Detection of Humans Using Multi-sensor Features. *Proc. of the 26th. Army Science Conference*, December 1-4, 2008.