A Layered Approach to People Detection in 3D Range Data

Luciano Spinello[a,b] **Kai O. Arras**[a] **Rudolph Triebel**[b] **Roland Siegwart**[b]

[a]Social Robotics Lab, University of Freiburg, Germany
[b]Autonomous Systems Lab, ETH Zurich, Switzerland

Abstract

People tracking is a key technology for autonomous systems, intelligent cars and social robots operating in populated environments. What makes the task difficult is that the appearance of humans in range data can change drastically as a function of body pose, distance to the sensor, self-occlusion and occlusion by other objects. In this paper we propose a novel approach to pedestrian detection in 3D range data based on supervised learning techniques to create a bank of classifiers for different height levels of the human body. In particular, our approach applies AdaBoost to train a strong classifier from geometrical and statistical features of groups of neighboring points at the same height. In a second step, the AdaBoost classifiers mutually enforce their evidence across different heights by voting into a continuous space. Pedestrians are finally found efficiently by mean-shift search for local maxima in the voting space. Experimental results carried out with 3D laser range data illustrate the robustness and efficiency of our approach even in cluttered urban environments. The learned people detector reaches a classification rate up to 96% from a single 3D scan.

1 Introduction

Robustly detecting pedestrians is a key problem for mobile robots and intelligent cars. Laser range sensors are particularly interesting for this task as, in contrast to vision, they are highly robust against illumination changes and typically provide a larger field of view.

In this paper we address the problem of detecting pedestrians in 3D range data. The approach presented here uses techniques from people detection in 2D range data for which a large amount of related work exists (Kluge, Köhler, and Prassler 2001; Fod, Howard, and Mataríc 2002; Schulz et al. 2003; Cui et al. 2005; Arras, Martínez Mozos, and Burgard 2007). In early works, people are detected using ad-hoc classifiers, looking for moving local minima in the scan. Learning has been applied for this task by (Arras, Martínez Mozos, and Burgard 2007) where a classifier for 2D point clouds has been learned by boosting a set of geometrical features. As there is a natural performance limit for people detection in a *single* 2D layer of range data, several authors started looking into the use of multiple co-planar 2D laser scanners (Gidel et al. 2008; Carballo, Ohya, and Yuta 2008). Close to our context is the work of (Mozos, Kurazume, and Hasegawa 2010), in which the authors apply boosting on each of three horizontal layers and use a probabilistic rule set to combine the three classifiers assuming a known ground plane.

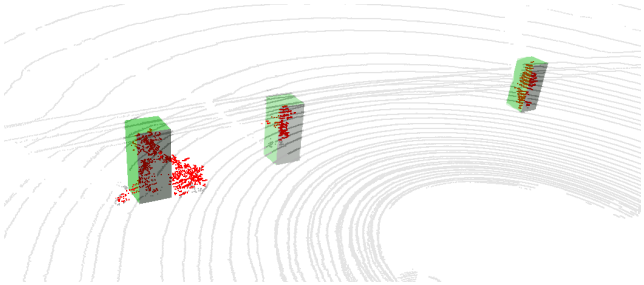

Figure 1: 3D pedestrian detection. A person pushing a buggy, a child and a walking pedestrian are correctly identified in the point cloud.

There is little related work on pedestrian detection in 3D data. (Navarro-Serment, Mertz, and Hebert 2009) collapse the 3D scan into a virtual 2D slice to find salient vertical objects above ground. For these objects, they align a window to the principal data direction, compute a set of features, and classify pedestrians using a set of SVMs. (Bajracharya et al. 2009) detect people in point clouds from stereo vision by processing vertical objects and considering a set of geometrical and statistical features of the cloud based on a fixed pedestrian model. From the comprehensive body of literature on people detection in images, we mention the most related ones, namely the HOG detector by (Dalal and Triggs 2005) and the ISM approach by (Leibe, Seemann, and Schiele 2005). People detection from multimodal data using laser and vision has been presented by (Spinello, Triebel, and Siegwart 2008).

To our knowledge, this work presents the first principled learning approach to people detection in 3D range data. The idea is to subdivide a pedestrian into parts defined by different height levels, and learn a highly specialized classifier for each part. We exploit the fact that most of the commercial 3D laser devices retrieve the environment as a set of individual *scan lines* which are not necessarily co-planar.

The building blocks used for classification are *segments*, i.e. groups of consecutive points in each scan line, on which a set of geometrical and statistical features are computed. We do not define a 3D pedestrian shape model beforehand, but instead learn it from labeled data by storing the displacements between the segment centers and the person's center. This allows for a general and robust description for articulated and complex 3D objects. Then, each segment is classified based on the likelihood of belonging to each part. We relate the output of each classifier geometrically by employing a 3.5D voting approach where each segment votes for the center of a person. Areas of high density in the continuous voting space define hypotheses for the occurrence of a person. This allows for robustness against occlusions as not all parts are needed for a detection. Moreover, our approach does not rely on any ground plane extraction heuristics and does not require any motion cues. No tracking is done in this work.

The paper is structured as follows: Sec. 2 explains the preprocessing steps we apply to the 3D data. Sec. 3 describes how we subdivide and learn a 3D person model from data. In Sec. 4 the detection step is presented. Sec. 5 contains the experimental results and Sec. 6 concludes the paper.

2 Preprocessing 3D Range Data

Different systems exist to acquire 3D range data from the environment. Many of them rely on a scanning device that sends out laser rays and measures the distance to the closest object. To acquire a 3D scan, such devices are usually rotated about one of the main axes of the sensor-based coordinate frame. Examples include 2D range finders such as the SICK LMS laser scanner, mounted on a turntable that rotates about its vertical or horizontal axis (Lamon, Kolski, and Siegwart 2006). Other devices, such as the Velodyne HDL-64E, also rotate about the z-axis sending out 64 independent laser beams that are not coplanar. The Alasca XT rangefinder uses a beam deflected by a rotating mirror and 4 receivers. Such sensors return point clouds that consist of individual *scan lines*, i.e. sequences of points that have been measured with the same beam. With some abstraction, we can thus think of such a 3D point cloud as a collection of 2D laser points arranged in *slices* or *layers*. This definition holds also for a wide set of non-laser sensors: range camera data (e.g. Swissranger) or point cloud data from stereo cameras can also be transformed into sets of scan lines by horizontally sampling image pixels.

Formally, we consider a point cloud $\mathcal{X}$ as consisting of layers $\mathcal{L}_i = \{\mathbf{x}_{ij}\}$, where $\mathbf{x}_{ij} = (x_{ij}, y_{ij}, z_{ij})$. In this paper, we demonstrate that by treating a 3D scan as a collection of 2D scans at different levels, known and proven techniques for detecting people in 2D range data can be easily extended to the 3D case, yielding a fast and robust people detector for 3D range data.

2.1 Point Cloud Segmentation per Layer

As a first step of our detection algorithm, we divide each scan line into *segments* using Jump Distance Clustering (JDC). JDC initializes a new segment each time the distance between two consecutive points exceeds a threshold θ_d. As a result, the data is reduced to a smaller number of segments with a higher amount of information than that of the raw data points. We denote each segment as a set $\mathcal{S}_j, j = 1, \ldots, N_i$ of consecutive points where N_i is the number of segments in scan line i. Our algorithm assumes that the 3D scanner rotates about the vertical z-axis, which means that the points in a segment are sorted by ascending azimuth angles. The segments constitute the primal element to extract local information.

Nr	Feature Name	Nr	Feature Name
f_1	Width	f_2	Number of points
f_3	Circularity	f_4	Linearity
f_5	Boundary length	f_6	Boundary regularity
f_7	Mean angular difference	f_8	Mean curvature
f_9	Quadratic spline fitting	f_{10}	Cubic spline fitting
f_{11}	Standard dev. w.r.t. centroid	f_{12}	Mean avg. dev. from median
f_{13}	Kurtosis w.r.t. centroid	f_{14}	Radius
f_{15}	PCA ratio	f_{16}	Bounding box area
f_{17}	Convex hull area		

Table 1: Features used to describe the shape and statistical properties of a segment.

2.2 Segment shape characterization

In the next step, we compute several *descriptors* for each extracted segment. A descriptor is defined as a function f_k: $\mathcal{S}_j \rightarrow \mathbb{R}$ that takes the M points contained in a segment $\mathcal{S}_j = \{(x_1, y_1, z_1) \ldots (x_M, y_M, z_M)\}$ as an input argument and returns a real value. Most of the features we use ($f_1 \ldots f_8$) have been presented by (Arras, Martínez Mozos, and Burgard 2007) and (Spinello, Triebel, and Siegwart 2008), the following ones ($f_9 \ldots f_{17}$) are added for this particular task:

- *Quadratic spline fitting*: this feature measures the residual sum of squares of a quadratic B-Spline regression s_2 (a piecewise polynomial approximation introduced by (De Boor 1978)) of the points in $\mathcal{S}_j$: $f_9 = \sum_i (s_2(x_i, y_i) - y_i)^2$
- *Cubic spline fitting*: this feature measures the residual sum of squares of a cubic B-Spline regression s_3 of the points in $\mathcal{S}_j$, i.e. $f_{10} = \sum_i (s_3(x_i, y_i) - y_i)^2$
- *Kurtosis with respect to centroid*: the kurtosis is defined as the fourth standardized moment of the cluster $\mathcal{S}_j$, i.e. $f_{12} = \frac{\sum_i (\mathbf{x}_i - \hat{\mathbf{x}}_j)^4}{M \cdot f_{11}^4}$, where f_{11} represents the standard deviation with respect to the centroid, and $\hat{\mathbf{x}}_i$ the center of gravity of $\mathcal{S}_j$.
- *PCA ratio*: this feature is the ratio between the second biggest eigenvalue λ_2 and the biggest eigenvalue λ_1 of the scatter matrix associated with $\mathcal{S}_j$. It measures the aspect ratio of the oriented bounding box, i.e. $f_{13} = \frac{\lambda_2}{\lambda_1 + 1}$
- *Bounding box area*: this feature represents the area of the axis-aligned bounding box of $\mathcal{S}_j$.
- *Convex hull area*: this feature represents the area computed from the convex hull polygon extracted from $\mathcal{S}_i$.

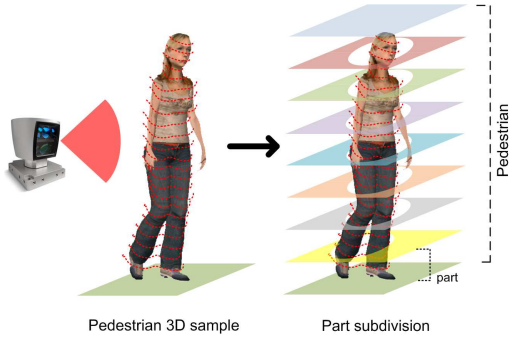

Figure 2: Learning a 3D person model. Objects are vertically divided into K parts. For each part an independent AdaBoost classifier is learned: all the segmented points contained on each scan line are considered as positive samples for the K AdaBoost classifiers.

Table 1 lists all 17 used features. The set of feature values of each segment $\mathcal{S}_i$ then forms a vector $\mathbf{f}_i = (f_1, \ldots, f_{17})$.

3 Learning a 3D Model of People

The appearance of people is highly variable. Humans have different sizes and body shapes, wear clothes, carry bags, backpacks, or umbrellas, pull suitcases, or push buggies. This makes it hard to predefine models for their appearance and motivates a learning approach based on a model that is created from acquired data.

3.1 Definition of Parts

We tackle the problem of the shape complexity of humans by a subdivision into different height layers or *parts* (see Fig. 2). The subdivision is defined beforehand and does not follow an anatomical semantics like legs, trunk, head. Results from computer vision literature (Dalal and Triggs 2005; Zhu et al. 2006; Viola and Jones 2002) and also our own experience show that descriptors computed in geometrically overlapping tessellations are powerful tools for learning an object model. Therefore, for learning a 3D person model we create K different and independent classifiers, each corresponding to a height-divided part of a human.

For training, all the scan lines that fall within a part are considered. The model is learned from subjects with similar heights (within $\pm 15cm$ from the mean). As part classifier we use AdaBoost (Freund and Schapire 1997), a well known machine learning algorithm, that has been proven successful for people detection in 2D range data (Arras, Martínez Mozos, and Burgard 2007).

3.2 Learning the Part Detectors

AdaBoost is a general method for creating an accurate strong classifier by combining a set of weighted weak classifiers, in this case decision stumps. A decision stump h_t defines a single axis-parallel partition of the feature space. The final strong classifier $H(\mathbf{f})$ computed for the feature vector $\mathbf{f}$ is a weighted sum of the T best weak classifiers:

$$H(\mathbf{f}) = \text{sign}\left(\sum_{t=1}^{T} \alpha_t h_t(\mathbf{f})\right), \quad (1)$$

where α_t are the weights learned by AdaBoost.

As people are usually represented only by a few number of data points in a 3D scan, there are many more background segments than segments on people. This makes the training set unbalanced. Now, instead of down-sampling the negative set, which could lead to an under-representation of the feature distribution, we use an adaptive initial weight vector $\mathbf{w}_0 \in \mathbb{R}^N$ where N is the total number of training segments. Usually $\mathbf{w}_0$ is set to a uniform distribution, i.e. $1/N \cdot \mathbf{1}^N$, where $\mathbf{1}^N$ is the vector of dimension N with all entries equal to 1. Instead we use

$$\mathbf{w}^p := \frac{1}{2N_{pos}}\mathbf{1}^{N_{pos}},\ \mathbf{w}^n := \frac{1}{2N_{neg}}\mathbf{1}^{N_{neg}},\ \mathbf{w}_0 = (\mathbf{w}^p, \mathbf{w}^n), \quad (2)$$

where N_{pos}, N_{neg} are the numbers of positive and negative training samples. Thus, the bigger training set – in our case the negative set – obtains a smaller weight.

To avoid early hard decisions in the classification of segments, we apply a sigmoid to the classification result in Eq. (1). This can be interpreted as a measure of likelihood $p(\pi^k \mid \mathbf{f}_i)$ of a segment i, represented by its feature vector $\mathbf{f}_i$, of corresponding to a part π^k of a pedestrian:

$$g_k(\mathbf{f}_i) = \frac{\sum_{t=1}^{T} \alpha_t^k h_t^k(\mathbf{f}_i)}{\sum_{t=1}^{T} \alpha_t^k}, \quad p(\pi^k \mid \mathbf{f}_i) = \left(1 + e^{2-13 g_k(\mathbf{f}_i)}\right)^{-1}, \quad (3)$$

where g_k is the normalized sum of weak classification results of the k-th classifier and π^k is a binary label that is true if segment i corresponds to part k.

In our case we need to classify one part against all others and the background. We therefore face a multi-class classification problem for which we follow a *one-vs-all* strategy: when training a part, all the features of the segments contained in that part are considered positive samples, the features of the background and of the other parts are tagged as negative samples.

3.3 Learning Geometric Relations

So far we described a way to classify parts of a person, now we combine the individual classifications into a full person detector. In computer vision, this problem is addressed using *part constellations* (Fergus, Perona, and Zisserman 2003), Conditional Random Fields (CRF) (Felzenszwalb and Huttenlocher 2005), or implicit shape models (ISM) (Leibe, Seemann, and Schiele 2005). Loosely inspired from the latter, we propose the following voting model.

First, we use the 3D displacement information of segments to define the geometric relations that constitute a 3D person model. Each part and each segment found in a part are considered independently. For a segment $\mathcal{S}_i$ found in part π^k in the training set, we store the 3D displacement vector $\mathbf{v}_i^k$, also called 'vote', i.e. the difference between the center of the person and the center of $\mathcal{S}_i$. Then all votes for part π^k are collected in a set $\mathcal{V}^k$. This information implicitly resembles different body poses of humans. For instance,

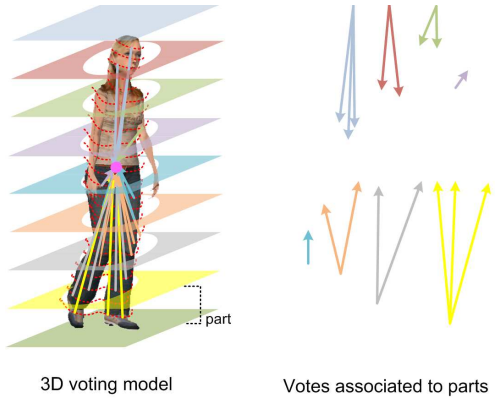

Figure 3: Learning a 3D voting model for training the detector. The displacement, or vote, between a segment and the person center of gravity in 3D associated to each subpart is stored as $\mathcal{V}^k$. Agglomerative clustering is carried out to obtain a more compact representation of $\mathcal{V}^k$.

if the training data is acquired during a walking sequence, the segments contained in the lowest body part (associated to feet/ankles) typically have a larger variation of displacements with respect to the center of gravity, i.e. the votes are spread out wider. After all person samples have been processed, a large amount of votes for each part is stored. The final step is then to compress $\mathcal{V}^k$ for each part into a new set $\hat{\mathcal{V}}^k$. This is achieved using agglomerative clustering with average linkage and a distance threshold θ_v (see Figure 3). Then, a weight $\hat{w}_i^k = |\hat{\mathcal{V}}^k|^{-1}$ is assigned to each clustered vote $\hat{\mathbf{v}}_i^k$ of $\hat{\mathcal{V}}^k$, where $|\hat{\mathcal{V}}^k|$ denotes the number of vote clusters for part π^k. The intuition here is that parts with a higher displacement variability imply a lower voting confidence. Finally, to obtain a practical geometrical interpretation of people in 3D, we compute the average bounding box from all person samples.

4 Detecting People in 3D

Our detection method processes a single point cloud as input and retrieves as output a set of detected people. After aquiring a new 3D scan, it is processed by the JDC segmentation step for each scan layer. Then, the feature vector $\mathbf{f}_i$ is computed for each segment found in each part. The likelihood of a segment to belong to a part is obtained by classifying the features with the corresponding AdaBoost model. Thus, we obtain a vector

$$\mathbf{c}_i = \Big(p(\pi^1 \mid \mathbf{f}_i), \ldots, p(\pi^k \mid \mathbf{f}_i)\Big). \tag{4}$$

This defines a multiple weighted part hypothesis, therefore we need to find a way of properly treating this information. Here, we use the voting model learned in the training phase (see Section 3) to generate hypotheses of person centers in 3D. It is important to note that no assumptions about the position of the ground plane are done in this work.

We formulate a 3.5D continuous voting space procedure in which each segment $\mathcal{S}_i$ casts a set of non-negative weighed votes $\mathcal{V}_1, \ldots, \mathcal{V}_K$ in 3D, where K is the number of pedestrian subparts. Each vote set $\mathcal{V}_m$ is weighted by the subpart classification likelihood of equation (4):

$$\rho(k) = \frac{\mathbf{c}_i(k)}{K} \tag{5}$$

where $\mathbf{c}_i(k)$ is the value of the m-element of the vector (4) and K the number of subparts.

All generated votes are collected in a continuous voting space $\mathcal{W}$. High density loci represent hypotheses of pedestrians centers in 3D. Therefore, we estimate the modes of the voting space distribution. This is achieved using Mean Shift estimation (Comaniciu and Meer 2002) with a spherical uniform kernel. Mean shift locates stationary points of a density function given discrete data sampled from that function. As soon as a mode is found, its score is computed:

$$score(\mathbf{x}_k \mid \mathcal{W}) = \left[\sum_j^{\mathcal{N}(\mathbf{x}_k)} \hat{v}_j^{\epsilon(v_j)} \rho(\epsilon(v_j))\right] \frac{\zeta(\mathcal{N}(\mathbf{x}_k))}{K}, \tag{6}$$

where $\mathbf{x}_k$ is a converged mode and $\mathcal{N}(\mathbf{x}_k)$ contains all the indices of the votes contributing to the basin of attraction of $\mathbf{x}_k$, $\epsilon(v_j)$ is a function that returns the part index of the vote v_j, $\hat{v}_j$ the weight value of vote v_j. $\zeta(\cdot)$ is a function that returns the number of parts from which the votes are originated. Thus, $\zeta(\cdot)$ is a modifier that favors people that are explained by more parts than others and it is very useful to decrease strong false positives that receive vote from clutter at the same height. The higher the hypothesis score of equation 6, the higher the likelihood of detecting people in $\mathbf{x}_k$. It is important to notice that this approach implements a form of simultaneous detection and segmentation: votes contributing to a hypothesis identify segments that belong to a person.

5 Experiments

We evaluate our algorithm on two outdoor data sets collected with a Velodyne HDL 64E S2 laser scanner. The first data set, named *Polyterrasse*, has been collected in a large area in the front of the ETH Zurich main building, accessible only to people and bicycles. The second data set, named *Tannenstrasse*, has been collected on a busy street crossing in downtown Zurich with trams, cars, pedestrians, or bicycles.

We collected 900 full-view point clouds for the first set and 500 for the second set. The sensor rotates with a frequency of $5Hz$ at a maximum range limited to $20m$. This produces around $120,000$ points per 3D scan. In each frame, people are manually annotated by a bounding box if they are represented by at least 200 points and exceed $1.20m$ in height. A second type of annotations are made for people represented by at least 100 points and $1m$ height.

5.1 Training

We train with 203 persons, standing still and walking. We define a subdivision of 9 parts at different heights given in Table 2. The vote clustering threshold is set to $\theta_v = 25cm$, the JDC threshold is set to $\theta_d = 40cm$. Training has been done on 2592 background segments and 7075 people segments from the *Polyterrasse* data set.

Part number	Part height	# of votes	# pos. training segments
1	$[0m, 0.2m]$	7	500
2	$[0.2m, 0.4m]$	9	814
3	$[0.4m, 0.6m]$	4	751
4	$[0.6m, 0.8m]$	3	811
5	$[0.8m, 1.0m]$	6	866
6	$[1.0m, 1.2m]$	3	881
7	$[1.2m, 1.4m]$	3	903
8	$[1.4m, 1.6m]$	3	917
9	$[1.6m, 2.5m]$	2	632

Table 2: Person parts subdivision and vote set.

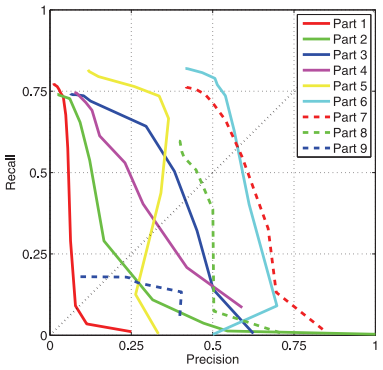

Figure 4: Precision-recall curve of the individual part detectors under the assumption of a known ground plane. Upper body parts (nr. 6-8) show better performance partly due to a better separation from the background, a higher point density, and a smoother shape.

The right-most column in Table 2 gives the resulting number of training samples for each part. The third column in Table 2 contains the number of votes for each part after the vote clustering step. Note that lower parts, related to legs, return more votes due to the varying displacements of the segments during walking. Only 20 decision stumps have been learned to avoid overfitting.

5.2 Quantitative Performance Evaluation

We evaluated each part detector on 440 frames not in the training set. To analyze the performance of the individual classifiers, we assumed to know the ground plane and selected the correct detector for each scan line. Figure 4 shows the precision-recall curve for each part classifier. It is interesting to see that the individual detection performances are rather poor. The detectors show a high recall behavior, high detection rates cause big quantities of false positives.

Figure 5 shows the overall precision-recall graph of the proposed method applied to both data sets. Detections are counted as true positives if the bounding box overlaps with a manually labeled person by more than 60% to account for metric inaccuracies in the annotation and the detection. Adopting the no-reward-no-penalization policy from (Enzweiler and Gavrila 2009), when a detection matches an annotation of the second type, no true positives or no false positives are counted.

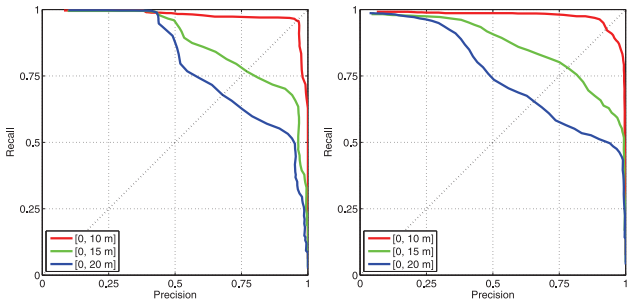

Figure 5: Evaluation for 3D people detection. Each figure depicts precision-recall graphs at different ranges: from 0 to $10m$, $15m$, and $20m$. **Left**: Detection performance for the *Polyterrasse* data set, the Equal Error Rates (EER) are 96%, 71%, 65%. **Right**: Precision-recall graph for the *Tannenstrasse* data set, EER values are 95%, 76%, 63%.

The performance increase over the individual part detectors is significant. The false positive rate is greatly decreased while the true positive rate is increased. This means that the part classifiers are diverse and complementary: if some parts do not return positive classifications, others do. This property is likely to explain the result shown in Fig. 1 where the detector correctly finds the child although no child was in the training set.

The figure also shows how the detection performance decreases with distance from the sensor. For the *Polyterrasse* data set, the Equal Error Rate (EER) for ranges between 0 to $10m$ is 96%, to $15m$ it is 71% and to $20m$ it is 65%. For the *Tannenstrasse* data set, the respective numbers are 95%, 76%, and 63%. The decay is mainly caused by point sparsity that leads to oversegmentation and less distinctive descriptors. This loss of detail renders the distinction of people from vertical structures of similar size such as traffic signs or pillars more difficult. The overall performance is comparable in both data sets although the *Polyterrasse* environment is well structured while the crossing of the *Tannenstrasse* is a rather busy place with clutter and all kinds of dynamic objects. Note that the person model was learned only with data from the *Polyterrasse* environment.

In our evaluation set, people are described by 1062 points on average when they are in a range of 0 to $10m$, by 557 points in a $15m$ range and by 290 points in a $20m$ range. Therefore, a 73% decrease in the number of points, from $10m$ to $20m$ range, causes only a 23% performance loss. Fig. 6 shows the detection results of two example frames.

A C++ implementation of the proposed method, not optimized for speed, obtains $\sim 1Hz$ detection frequency in an average 3D scan, limited to $10m$ maximum range, of around $75,000$ points.

6 Conclusions

In this paper we presented a principled learning approach to people detection in 3D range data. Our approach subdivides a person into parts at different height levels. For each part a specialized AdaBoost classifier is created from a set of geometrical and statistical features computed on segments. The

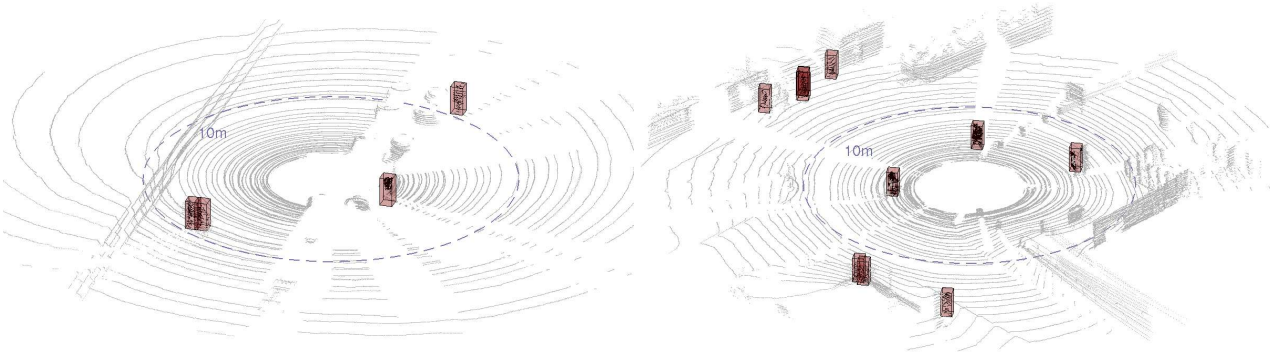

Figure 6: Two example frames showing detection results as red boxes. **Left**: A frame of the *Polyterrasse* data set. Closely walking people and a partly visible individual standing close to the sensor are correctly found. **Right**: A frame of the *Tannenstrasse* data set showing a cluttered urban street crossing with people, pillars, street signs, and a tram just entering the scene. People are correctly detected while crossing the street and walking at a large distance to the sensor. There are two false positives in the lower left part of the picture caused by a glass window with vertical steel poles.

classifiers mutually enforce their evidence across different heights by voting into a continuous space. This approach allows for the detection of people from partial views and does not require knowledge of the ground plane. In experiments with two different data sets in cluttered urban environments, a classification rate up to 96% has been achieved which is a high value given that we detect people from a single 3D scan. In future work, we plan to combine this method with tracking to integrate detection results over time.

Acknowledgements

This work has been supported by the German Research Foundation (DFG) under contract number SFB/TR-8 and EU Project EUROPA-FP7-231888.

References

Arras, K. O.; Martínez Mozos, O.; and Burgard, W. 2007. Using boosted features for the detection of people in 2d range data. In *Int. Conf. on Rob. & Autom. (ICRA)*.

Bajracharya, M.; Moghaddam, B.; Howard, A.; Brennan, S.; and Matthies, L. 2009. Results from a real-time stereo-based pedestrian detection system on a moving vehicle. In *Workshop on People Detection and Tracking, IEEE ICRA*.

Carballo, A.; Ohya, A.; and Yuta, S. 2008. Fusion of double layered multiple laser range finders for people detection from a mobile robot. In *IEEE Int. Conf. on Multisensor Fusion and Integration for Intelligent Systems (MFI)*.

Comaniciu, D., and Meer, P. 2002. Mean shift: A robust approach toward feature space analysis. *IEEE Trans. on Pattern Analysis & Machine Intell.* 24:603–619.

Cui, J.; Zha, H.; Zhao, H.; and Shibasaki, R. 2005. Tracking multiple people using laser and vision. In *Int. Conf. on Intel. Rob. and Sys. (IROS)*.

Dalal, N., and Triggs, B. 2005. Histograms of oriented gradients for human detection. In *IEEE Conf. on Comp. Vis. and Pat. Recog. (CVPR)*.

De Boor, C. 1978. *A Practical Guide to Splines*. Springer-Verlag.

Enzweiler, M., and Gavrila, D. 2009. Monocular pedestrian detection: Survey and experiments. *IEEE Trans. on Pattern Analysis & Machine Intell.* 31(12):2179–2195.

Felzenszwalb, P., and Huttenlocher, D. 2005. Pictorial structures for object recognition. *Int. Journ. of Comp. Vis.* 2:66–73.

Fergus, R.; Perona, P.; and Zisserman, A. 2003. Object class recognition by unsupervised scale-invariant learning. *IEEE Conf. on Comp. Vis. and Pat. Recog. (CVPR)*.

Fod, A.; Howard, A.; and Mataríc, M. 2002. Laser-based people tracking. In *Int. Conf. on Rob. & Autom. (ICRA)*.

Freund, Y., and Schapire, R. E. 1997. A decision-theoretic generalization of on-line learning and an application to boosting. *Jour. of Comp. and System Sciences* 55(1).

Gidel, S.; Checchin, P.; Blanc, C.; Chateau, T.; and Trassoudaine, L. 2008. Pedestrian detection method using a multilayer laserscanner: Application in urban environment. In *Int. Conf. on Intel. Rob. and Sys. (IROS)*.

Kluge, B.; Köhler, C.; and Prassler, E. 2001. Fast and robust tracking of multiple moving objects with a laser range finder. In *Int. Conf. on Rob. & Autom. (ICRA)*.

Lamon, P.; Kolski, S.; and Siegwart, R. 2006. The smartter - a vehicle for fully autonomous navigation and mapping in outdoor environments. In *CLAWAR*.

Leibe, B.; Seemann, E.; and Schiele, B. 2005. Pedestrian detection in crowded scenes. In *IEEE Conf. on Comp. Vis. and Pat. Recog. (CVPR)*.

Mozos, O. M.; Kurazume, R.; and Hasegawa, T. 2010. Multi-part people detection using 2d range data. *International Journal of Social Robotics* 2(1).

Navarro-Serment, L.; Mertz, C.; and Hebert, M. 2009. Pedestrian detection and tracking using three-dimensional ladar data. In *Int. Conf. on Field and Service Robotics*.

Schulz, D.; Burgard, W.; Fox, D.; and Cremers, A. 2003. People tracking with a mobile robot using sample-based joint probabilistic data association filters. *Int. Journ. of Rob. Research* 22(2):99–116.

Spinello, L.; Triebel, R.; and Siegwart, R. 2008. Multimodal people detection and tracking in crowded scenes. In *AAAI Conf. on Artif. Intell. (AAAI)*.

Viola, P., and Jones, M. 2002. Robust real-time object detection. *Int. Journ. of Comp. Vis.*

Zhu, Q.; Yeh, M. C.; Cheng, K. T.; and Avidan, S. 2006. Fast human detection using a cascade of histograms of oriented gradients. In *IEEE Conf. on Comp. Vis. and Pat. Recog. (CVPR)*.