3D Laser Range-based People Detection using Bag-of-Features

Jens Behley, Volker Steinhage, and Armin B. Cremers

***Abstract*— Recognizing pedestrians in sensor data is crucial for autonomous driving. The usage of three-dimensional laser range data for this task recently attracted interest due to advances in sensor technology. In image-based detection and classification, bag-of-feature approaches proved to be a versatile tool for state-of-the-art results. However, they are mostly ignored in laser range-based classification and detection. In this contribution we investigate different important design decisions when implementing bag-of-features for laser-based detection. We discuss in particular the choice of descriptors and parameters for the construction of a suitable vocabulary. The experiments show that a bag-of-feature approach with very simple descriptors can be as effective as other recently proposed state-of-the-art pedestrian detectors with more complex descriptors.**

I. Introduction

A robust and reliable pedestrian detection is essential for the deployment of self-driving cars into everyday traffic. For this purpose, laser rangefinders such as the Velodyne HDL-64E, are the most interesting sensors as they provide precise depth measurements of surrounding objects. The depth information is viable to maneuver collision-free at normal driving speeds through city traffic. A further advantage of laser range data is the simple segmentation compared to the segmentation of images. There are actually a lot of approaches to get coherent and meaningful segments from laser range data, ranging from very simple grid-based approaches to more sophisticated ones [1], [2]. Thus, this paper concentrates on the detection of people and will assumes that segments were generated with a suitable approach.

Bag-of-Features (BoFs) or Bag-of-Words are an attractive approach to extract a meaningful feature representation for such segments. They offer by design several advantageous properties compared to other segment descriptors, which are required especially in laser perception: (1) BoFs are robust to partial occlusions, since the histogram can compensate missing entries — as long as the distribution is the same, it should be still possible to attain a correct classification. (2) Even if we encounter an undersegmentation, the entries for a certain class should be still visible in a part of the histogram. (3) The single feature computations are independent of each other, which makes a concurrent computation possible — current 3d laser rangefinders produce millions of laser points. Nowadays, BoF-based approaches are among the most powerful approaches in computer vision for different tasks, such as object detection [3] and object recognition [4], [5]. Hence, it is remarkable that BoF approaches are mostly ignored in laser-based perception.

J. Behley , V. Steinhage and A. B. Cremers are with the Department of Computer Science III, University of Bonn, 53117 Bonn, Germany.
{behley, steinhage, abc}@iai.uni-bonn.de

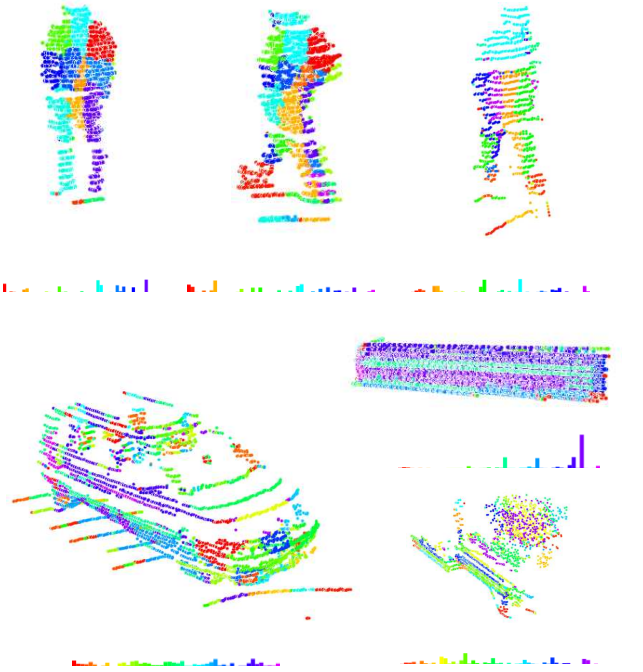


Fig. 1. Bag-of-feature encoding for pedestrians and background objects — a car, vegetation and a wall. Every point is colored according to the corresponding word in the learned vocabulary of size 50. The words in the vocabulary correspond to different simple Distribution Histograms. Below the point clouds are the corresponding histograms shown, where we see differences in the distributions.

A straight-forward approach to extract a BoF using laser range data is to generate depth images and extract image-based features [6]. Nevertheless, converting a laser range scan to a depth image generally entails information loss due to projection to the image plane. In this contribution, we will use the three-dimensional laser range directly data and extract simple and fast to compute three-dimensional histogram descriptors [7].

Spinello *et al.* [8] recently compared their pedestrian detector to a BoF approach and in their experiments, BoF performed sub-optimally. We will investigate this behaviour and will show that BoF can be on par with state-of-the-art people detectors. Especially, the choice of the descriptor and a vocabulary of appropriate size will be shown to be essential for a working BoF detection approach. We show that axis-aligned Distribution Histograms — despite missing invariances — are better suited than rotation-invariant Spin Images [9]. Figure 1 depicts the BoF encoding with Distribution Histograms and a small vocabulary.

The paper is organized as follows. Section II summarizes related work with some applications of BoF. In section III we will outline the general BoF framework. After this general overview, we will in section IV experimentally evaluate different parameters for the BoF with person detection. Finally, section V will conclude the paper and outline future work.

II. Related Work

People detection using images attracted the interest of several researches over the past decade. Dollár *et al.* [10] survey and evaluate several state-of-the-art vision-based person detection approaches. The evaluation points out that most of the approaches are not able to detect persons reliably at mid or far ranges. Furthermore, occlusions are not handled such that a high-detection rate can be maintained.

Especially in robotics, laser rangefinders were intensively used to track and also detect people in 2D laser range data [11], [12]. Also multiple layers of 2D laser rangefinders were employed to detect people more reliably by combining evidence of multiple 2d layers [13], [14]. Recently, also fast high-resolution 3D laser rangefinders, such as the Velodyne HDL-64E, were used to detect and track people [15], [16], [17], [8], [18].

In more detail, Spinello *et al.* [8] combine a detection algorithm based on slices of a 3D laser rangefinder with a classification approach to filter out false positives, the so-called the Bottom-up Top-down Detector(BUTD). In the Bottom-up step, a detector [19] is used to get possible person locations in the sensor data. In the Top-down phase, bounding boxes at these locations are validated with a AdaBoost classifier and a learned tesselation scheme. Similarly, Carballo *et al.* [14] use AdaBoost and multiple layers from 2D range scanners to detect people. In their approach, one 2D laser rangefinder is mounted at leg height and one at chest height. Individual detections from these layers are combined into a complete detection hypothesis. Using laser remission-based features the robustness was further improved. Kidono *et al.* [17] use a support vector machine (SVM) with radial basis functions. They also calculate several shape features and a remission histogram. The overall performance of the classifier at mid and far ranges was improved with so-called slice features capturing the widths in different slices. In contrast to these descriminative approaches, Kasestner *et al.* [18] use a generative approach for object detection and tracking.

A segment-based approach was also applied to multi-class classification. Teichman *et al.* [20] present a classification approach based on tracking information. From the tracking information and segmentation a AdaBoost based classsification approach is learned. This approach classifies dynamic objects using segement descriptors and motion descriptors. To filter and enhance the result, a seperate Bayes filter is used to smooth the classification results. Later this approach has been extended to use unlabeled track labels with few labeled tracks for semi-supervised learning [21].

Other applications in laser-based perception also used Bag-of-Features. Endres *et al.* [22] use Bag-of-Features to unsupervised clustering with latent dirichlet allocation [23]. In their approach, a vocabulary is extracted by discretizing a variant of Spin Images [9]. Steder *et al.* [6] use Bag-of-Features with depth images and the so-called NARF descriptor [24] for place recognition.

In contrast to the related approaches, we will use very simple histogram features and combine them into a more descriptive representation. We will show that such a representation also enables state-of-the-art pedestrian detection performance. The choice of the histogram descriptors is driven by our recently published evaluation on histogram descriptors in point-wise classification [7]. In this evaluation we showed that a larger feature size can improve the performance of linear classifiers significantly.

III. Bag-of-Features

The general idea of Bag-of-Features can be intuitively described using the relation to text documents. A document is made of words. Counting these words and inspecting their distribution in the text, we can mostly guess the topic of this document. This idea has been applied to the modeling of text documents, and it can be also used in the context of images. However, we first need to define a vocabulary, as with images and also laser range scans, there is no natural concept like words. As mostly done, we will also use different descriptors to represent the contents of a laser range scan. Hence, in a general bag-of-feature approach we have two stages – first extracting a vocabulary from the laser range scans, and second describing a segment with a histogram of word occurances from this vocabulary.

For vocabulary extraction, descriptors are usually randomly sampled from a large number of unlabeled segments. Then a vocabulary is learned using a clustering algorithm to find representative descriptors, i.e., the words in the vocabulary. The idea is to represent the contents of a segment by descriptors, which often occur in the world of segments. In the literature, different algorithms for the generation of the vocabulary are presented [4]. An important choice in every BoF approach is the number of words in the vocabulary and usally a larger vocabulary results in better performance. However, a too large vocabulary can also decrease the performance as we get an oversegmentation of the descriptor space. Intuitively, we get multiple words to describe the same concept, which makes it hard to get a general description of an object.

With the learned vocabulary, we build a histogram of word frequencies – the so-called Bag-of-Features or Bag-of-Words – by evaluating the descriptor at regularly spaced locations or using interest points. After determining the descriptors of a segment, we need to convert these descriptors into a histogram of words. This can be done with different quantization methods, e.g., hard and soft assignment. Hard assignments mean we search for the nearest matching word in the vocabulary and increment the count of this word in the histogram. In contrast, soft assignment assigns a descriptor

TABLE I

OVERALL NUMBER OF SEGMENTS IN THE DATASETS.

dataset	pedestrian	bicyclist	car	background
Polyterrasse	3,380	-	-	2,046
Tannenstrasse	3,629	47	131	6,907

TABLE II

NUMBER OF SEGMENTS IN TRAINSET AND TESTSET.

dataset	trainset		testset	
	pedestrian	background	pedestrian	background
Polyterrasse	2,389	1,146	991	900
Tannenstrasse	2,460	6,013	1,169	1,072

to multiple words weighted by the distance to the cluster centers — the count of multiple words is increased.

After this outline of a general Bag-of-Features approach, we now specify the choices for our detection framework. We apply k-means clustering for learning the vocabulary and the cluster centers represent the words in the vocabulary. We will use a dense sampling and compute a descriptor for every point inside the segment. In our approach, we use the euclidean distance of a descriptor to a cluster center and a hard assignment to words. Using the representation by a Bag-of-Features we finally learn a classifier to discriminate pedestrians and background segments. In our approach, we learn Spectrally Hashed Logistic Regression (SHLR) [25] with normalized Bag-of-Feature vectors.

IV. EXPERIMENTAL RESULTS

Datasets. All datasets were recorded with a static Velodyne HDL-64E 3D laser rangefinder mounted on a tripod in a populated urban enviroment [8].

The first dataset "Polyterasse" was recorded in front of the main building of the ETH Zurich, Switzerland. The second data set "Tannenstrasse" has been collected in the city of Zurich at a street crossing. The sensor rotated with a frequency of $5\,\mathrm{Hz}$ and generated around $120,000$ points per revelation. We manually annotated static and dynamic objects of interest with bounding boxes and corresponding labels *pedestrian*, *bicyclist*, *car* and *background*. The *background* class contains segments of building walls, vegetation and poles. The segmentation was performed semi-automatically using a 2D obstacle map applying different resolutions for near and far objects to get tight bounding boxes for pedestrians. Furthermore, we tried to segment nearby people into distinct segments and also annotated all pedestrians with their occlusion – none, partial and full. Thus, it is possible that a person disappears and reappears after being fully occluded and still is assigned to the same track id. Table I summerizes the number of segments used in the evaluation. As Spinello *et al.* [8], [19] point out, the dataset Tannenstrasse is more challenging than Polyterrasse as there are more background objects with person-like extents.

Implementation details. We randomly extract 100.000 descriptors from the training data and cluster the data using k-means [26]. For efficiently searching the nearest descriptor in the vocabulary, we use a kD-tree [27]. We further used an octree to speed up the nearest neighbor search in the 3D point clouds and estimate normals using PCA with a neighborhood of $0.6\,\mathrm{m}$.

Experimental setup. We split all datasets in training data and test data according to the annotation into distinctive tracks of dynamic and static objects. We used $3/4$ of all tracks for training and the rest for testing the detector. We decided to use this setup, as we wanted separate sets, where background segments never appear in the training set. Additionally, we filtered out all segments with a distance larger than $25.0\,\mathrm{m}$, smaller than $1.0\,\mathrm{m}$ height and with less than 100 laser points [8]. Table II summarizes the number of segments used for training and testing the detector.

We report the Precision/Recall curves to better asses the performance of the different parameters. We will also report the Equal-Error-Rate(EER) determined on the testsets for comparison with the related work using the same datasets.

For comparison, we calculate a single Spin Image per segment using the extent of the surrounding bounding box. All points of the segment are normalized to $[-1,1]\times[-1,1]\times[-1,1]$ and a descriptor is calculated at the origin of the transformed points. We determine the best setting of the number of bins with a cross validation on the training data. We used 14 bins in the Polyterrasse and 21 bins in the Tannenstrasse dataset.

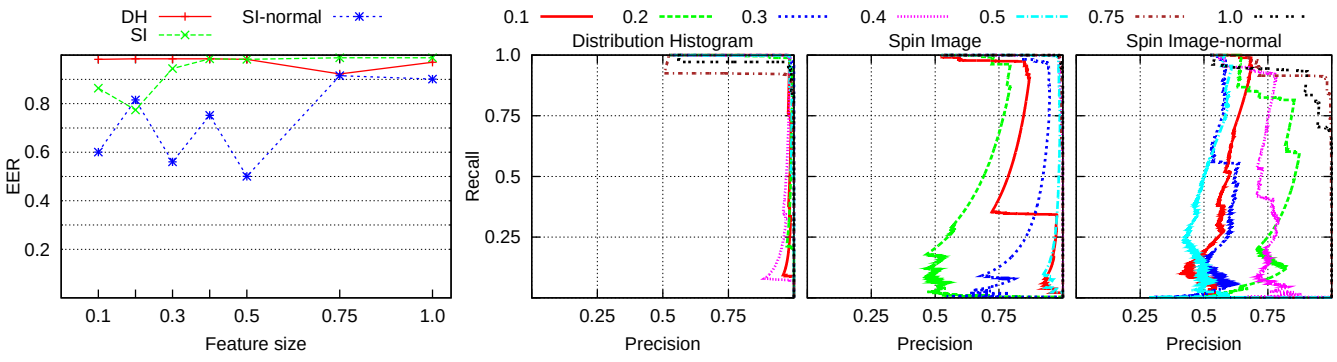
All experiments were performed on an Intel Xeon X5550 with $2.67\,\mathrm{GHz}$ using a single core and $12\,\mathrm{GB}$ memory. We fixed the parameters of the SHLR to 8 bit for the hashing and 4 bit for the similarity search.

A. Descriptors

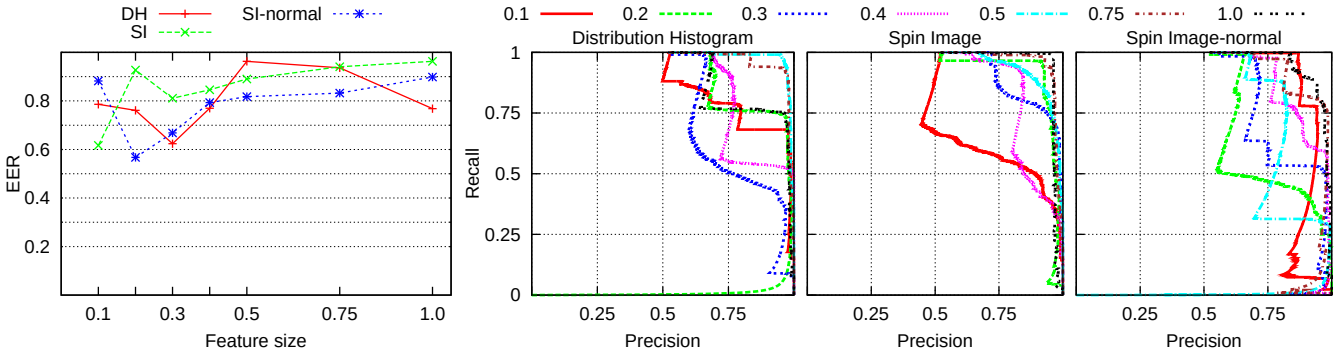
In the presented approach for Bag-of-Features, we compute one descriptor for every laser point. To account for the large number of points and to maintain a reasonable performance, we use very simple and fast histogram descriptors: Spin Images [9] and a variant of the Distribution Histograms [7], where the reference axes are all axis-aligned. Thus, we can circumvent the additional overhead of normal computation. For comparison we also use the standard Spin Images in the local reference frame, i.e., we use a normal-based reference frame.

In an earlier evaluation [7], we experimentally evaluated different descriptors using a local and a global reference frame. The local reference frame is determined using the point normal estimated using neighboring points. Another possible choice is the usage of a the up-vector from the robot pose. The second choice proved to be more robust and also improved the classification results significantly.

We choose to use only 3 bins for all histogram descriptors to keep the size of a single descriptor small. In these experiments we learn a large vocabulary of 1600 words and uniformly sample $100,000$ descriptors from the training segments.



(a) Polyterrasse



(b) Tannenstrasse

Fig. 2. Influence of the feature size on the performance of the detector. In (a) and (b) are the precision/recall curves for different feature sizes shown. In the more difficult dataset Tannenstrasse, we need a larger features to capture the differencies between the pedestrians and the background. Especially, with the Spin Images a larger radius is boosting the descriptiveness enormously.

Figure 2 shows the performance for Distribution Histogram and Spin Images with increasing radii. Interestingly, the different spin images in the global and also the local reference frame achieve the highest EER with the largest radius. The distribution histogram achieves the same performance with smaller radii and we can observe a drop in the performance with small and large radii.

The need for features with a radius of at least $0.5\,\mathrm{m}$ can be explained by the resolution of the laser range scanner. A smaller radius does only capture points of a single scanline at large distances.

This also confirms our earlier findings [7] on the feature size in the context of point-wise classification. There we could show a significant increase in the performance with different classifiers with increasing radius of the histogram descriptors. With point-wise classifcation this behavior can be explained by the context, which is implicitly encoded in the histogram. It seems that for the classification of segments of rather different extents — small segments from bushes and pedestrians and large segments from cars and walls — we also need large enough descriptors to capture the difference of the different classes.

With the largest radius we can also use 5 or even more bins to get a finer sampling of the shape at near distances. However, a smaller bin size does also enlarge the risk of discretization errors.

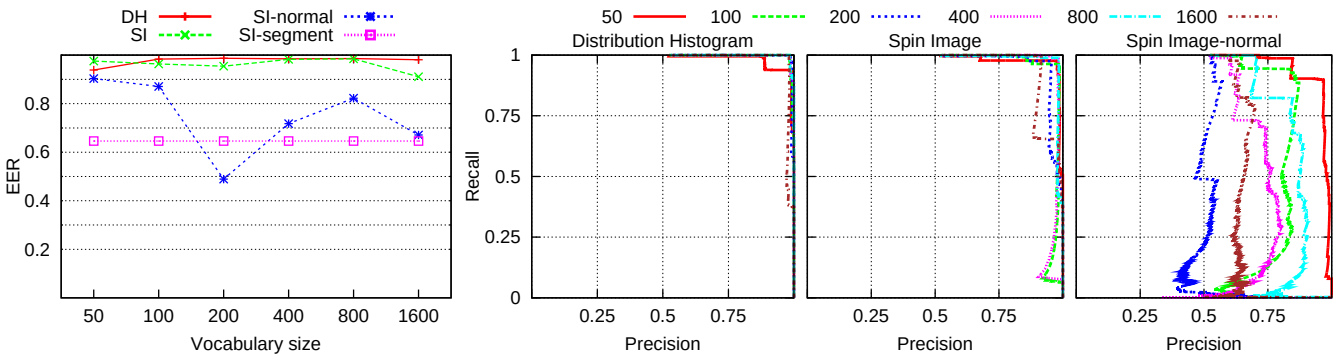
Considering computational performance, we should prefer the smallest possible radius, as we need hundreds of range queries, which are far more expensive at larger radii. In our unoptimized implementation we need approximately $0.01\,\mathrm{s}$ per pedestrian segment with a radius of $0.1\,\mathrm{m}$. With a descriptor radius above of $0.5\,\mathrm{m}$, this increases to $0.1\,\mathrm{s}$. With $10 - 20$ segment from background and pedestrians it takes too long to enable real-time processing.

B. *Vocabulary size*

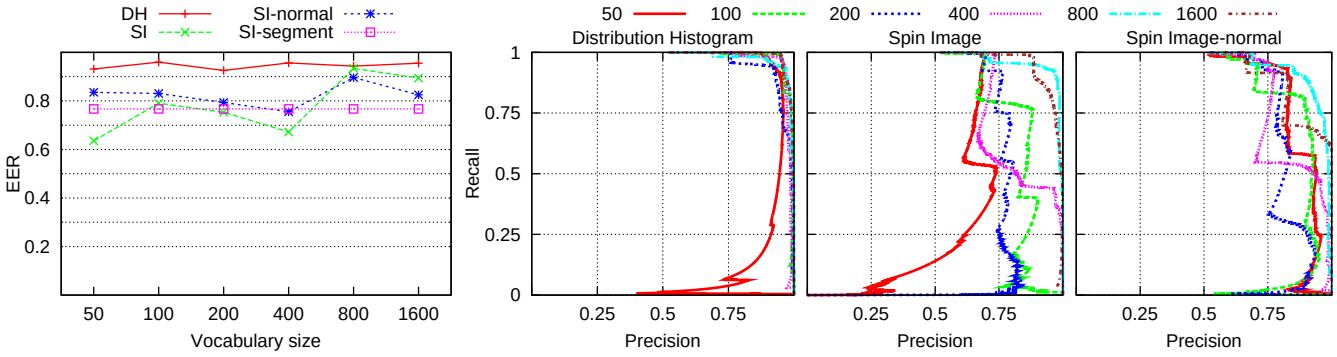
We evaluated the influence of different numbers of words in the vocabulary using a moderate feature radius of $0.5\,\mathrm{m}$. As before we sample $100,000$ descriptors from the training set to generate the vocabulary.

Figure 3 depicts the precision/recall diagrams and also the EER for different codebook sizes. As can be seen from the precision/recall curves, increasing vocabulary sizes leads to improvments for Spin Images and normal-based Spin Images. A clear advantage of the z-axis based descriptors compared to the normal-based spin image is also observable. Surprisingly, the most restricted descriptor — the axis-aligned distribution histogram — attains by far the best performance. With no degree of freedom we might need more words to capture the differencies in the shape. But compared to the time required to compute a single descriptor, we should invest computation time in larger vocabularies than larger features.

With a rotation invariant descriptor, we can get for arbi-



(a) Polyterrasse



(b) Tannenstrasse

Fig. 3. Influence of the vocabulary size on the performance of the detector. In (a) and (b) are the precision/recall curves for different vocabulary sizes shown. In the more difficult dataset Tannenstrasse, we need a large vocabulary to capture the differencies between the pedestrians and the background. In (a) we can see that a vocabulary size of 50 words is almost enough to get a descriptive vocabulary, when we use the Distribution Histogram.

trary rotations of an objects a similar feature representation. But in urban environments this is seldomly needed, as we have rather constrained orientations of the objects. We will encounter pedestrians up-side down or cars not driving on streets. Thus, a restriction of the orientation to the up-vector does not reduce the performance of the descriptors.

By removing almost every degree of freedom, as with the axis-aligned distribution histogram, we get different descriptors for the left and right side of an object. But in the case of the bag-of-features this is not necessarily a disadvantage, if we have enough words to describe these descriptors. Another aspect of the difficulties with the normal-aligned Spin Image is the noisiness in the estimated normal. In our experiments, we calculated the normal using PCA. This estimate is affected by the size of the neighborhood, which we choose to be $0.6\,\mathrm{m}$ to get normal estimates at large range. But this large radius makes it hard to get consitent normals for the pedestrians — moving one arm affects the normal estimates on the body with such large radius.

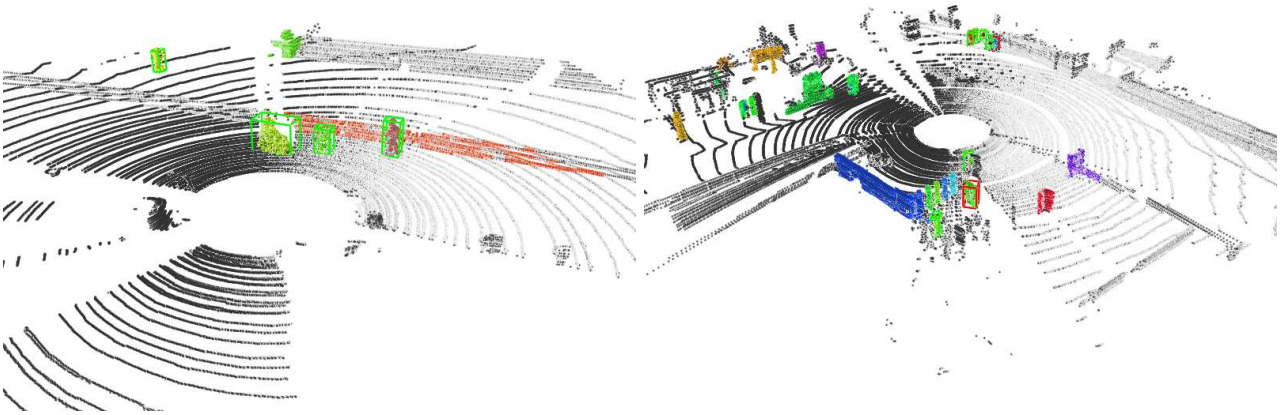
In conclusion we have shown that a BoF approach can yield state-of-the-art results. The best results are achieved with the Distribution Histogram in the Polyterrasse dataset an EER of 98.6% with a codebook size of 200 words and Distribution Histograms with radius 0.5, but this is mainly caused by the rather simple background objects. In the more challenging dataset Tannenstrasse we get an EER of upto 95.9% using Distribution Histograms with $0.5\,\mathrm{m}$ radius and a vocabulary size of 100 words, which is on par with the BUTD [8]. Figure 4 shows results of our approach with the evaluated datasets. Missing detection or false positives are mainly caused by severe partial occlusions.

V. Conclusion and Future Work

In this contribution, we have shown that a bag-of-feature approach with very simple descriptors can be used to get a very descriptive represtenation of a segment. This representation enables a state-of-the-art performance with the detection of pedestrians in three-dimensional laser range data. In a first set of experiments we have shown that some descriptors require a very large radius and also a large vocabulary. The results also suggest that with the Distribution Histogram, we can get similar results with smaller descriptors.

The results serve as an important building block for other interesting applications. The combination with a tracking approach [16] is the next obvious step. Besides this, we also plan to investigate the performace in multi-class segment-based classification [20]. The points mentioned in section I, i.e., robustness to partial occlusions and undersegmentation, should also be experimentally verified.

The computation of the descriptors currently is the bottleneck in the overall framework. Thus, the investigation of sub-sampling schemes and faster methods to evaluate the descriptors is necessary to achive real-time performance. As the descriptor computations are independent, we assume they



(a) (b)

Fig. 4. Detection results with the Distribution Histogram on the (a) Polyterrasse and (b) Tannenstrasse datasets. Extracted segments are depicted by colored points. True positive detections are shown with a green bounding box and false positive/negative detections are shown by a red bounding box.

can be done using a GPU. An avenue is also the investigation of other desriptors and vocabulary learning algorithms [4], pooling schemes [28] and the usage of spatial pyramids [29] to further improve the expressiveness of the representation.

VI. Acknowledgement

We want to thank Luciano Spinello and collegues for providing the datasets used in the evaluation.

References

[1] K. Klasing, D. Wollherr, and M. Buss, "A Clustering Method for Efficient Segmentation of 3d Laser Data," in *ICRA*, 2008, pp. 4043–4048.

[2] F. Moosmann, O. Pink, and C. Stiller, "Segmentation of 3D Lidar Data in non-flat Urban Environments using a Local Convexity Criterion," in *IV*, 2009, pp. 215–220.

[3] K. E. A. van der Sande, J. R. R. Uijlings, T. Gevers, and A. W. M. Smeulders, "Segmentation as Selective Search for Object Recognition," in *ICCV*, 2011, pp. 1879–1886.

[4] A. Coates, H. Lee, and A. Y. Ng, "An Analysis of Single-Layer Networks in Unsupervised Feature Learning," in *AISTATS*, vol. 15, 2011, pp. 215–223.

[5] A. Coates, B. Carpenter, C. Case, S. Satheesh, B. Suresh, T. Wang, D. J. Wu, and A. Y. Ng, "Text Detection and Character Recognition in Scene Images with Unsupervised Feature Learning," in *ICDAR*, 2011, pp. 440–445.

[6] B. Steder, M. Ruhnke, S. Grzonka, and W. Burgard, "Place Recognition in 3D Scans Using a Combination of Bag of Words and Point Feature based Relative Pose Estimation," in *IROS*, 2011, pp. 1249–1255.

[7] J. Behley, V. Steinhage, and A. B. Cremers, "Performance of Histogram Descriptors for the Classification of 3D Laser Range Data in Urban Environments," in *ICRA*, 2012, to appear.

[8] L. Spinello, M. Luber, and K. O. Arras, "Tracking People in 3D Using a Bottom-Up Top-Down Detector," in *ICRA*, 2011, pp. 1304–1310.

[9] A. Johnson and M. Hebert, "Using spin images for effcient object recognition in cluttered 3D scenes," *TPAMI*, vol. 21, no. 5, pp. 433–449, 1999.

[10] P. Dollár, C. Wojek, B. Schiele, and P. Perona, "Pedestrian Detection: An Evaluation of the State of the Art," *TPAMI*, vol. 34, no. 4, pp. 743–761, 2012.

[11] D. Schulz, W. Burgard, D. Fox, and A. B. Cremers, "People Tracking with a Mobile Robot Using Sample-based Joint Probabilisitic Data Association Filters," *IJRR*, vol. 22, no. 2, pp. 99–116, 2003.

[12] K. O. Arras, S. Grzonka, M. Luber, and W. Burgard, "Efficient People Tracking in Laser Range Data using a Multi-Hypothesis Leg-Tracker with Adaptive Occlusion Probabilities," in *ICRA*, 2008, pp. 1710–1715.

[13] O. M. Mozos, R. Kurazume, and T. Hasegawa, "Multi-Part People Detection Using 2D Range Data," *IJSR*, vol. 2, no. 1, pp. 31–40, 2010.

[14] A. Carballo, A. Ohya, and S. Yuta, "People Detection using Range and Intensity Data from Multi-Layered Laser Range Finders," in *IROS*, 2010, pp. 5849–8854.

[15] L. E. Navarro-Serment, C. Mertz, N. Vandapel, and M. Hebert, "Pedestrian Detection and Tracking Using Three-dimensional LADAR Data," *IJRR*, vol. 29, no. 12, pp. 1516–1528, 2010.

[16] F. Schöler, J. Behley, V. Steinhage, D. Schulz, and A. B. Cremers, "Person Tracking in Three-Dimensional Laser Range Data with Explicit Occlusion Adaption," in *ICRA*, 2011, pp. 1297–1303.

[17] K. Kidono, T. Miyasaka, A. Watanabe, T. Naito, and J. Miura, "Pedestrian Recognition Using High-definition LIDAR," in *IV*, 2011.

[18] R. Kasestner, J. Maye, Y. Pilat, and R. Siegwart, "Generative Object Detection and Tracking in 3D Range Data," in *ICRA*, 2012, to appear.

[19] L. Spinello, K. O. Arras, R. Triebel, and R. Siegwart, "A Layered Approach to People Detection in 3D Range Data," in *AAAI*, 2010.

[20] A. Teichman, J. Levinson, and S. Thrun, "Towards 3D Object Recognition via Classification of Arbitrary Object Tracks," in *ICRA*, 2011, pp. 4034–4041.

[21] A. Teichman and S. Thrun, "Tracking-based semi-supervised learning," in *RSS*, 2011.

[22] F. Endres, C. Plagemann, C. Stachniss, and W. Burgard, "Unsupervised Discovery of Object Classes from Range Data using Latent Dirichlet Allocation," in *RSS*, 2009.

[23] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *JMLR*, vol. 3, no. Jan, pp. 993–1022, 2003.

[24] B. Steder, R. B. Rusu, K. Konolige, and W. Burgard, "Point Feature Extraction on 3D Range Scans Taking into Account Object Boundaries," in *ICRA*, 2011, pp. 2601–2608.

[25] J. Behley, K. Kersting, D. Schulz, V. Steinhage, and A. B. Cremers, "Learning to Hash Logistic Regression for Fast 3D Scan Point Classification," in *IROS*, 2010, pp. 5960–5965.

[26] D. Arthur and S. Vassilvitskii, "k-means++: The Advantages of Careful Seeding," in *SODA*, 2007, pp. 1027–1035.

[27] S. Arya, D. M. Mount, N. S. Netanyahu, R. Silverman, and A. Y. Wu, "An optimal algorithm for approximate nearest neighbor searching in fixed dimensions," *JACM*, vol. 45, no. 6, pp. 891–923, 1998.

[28] Y.-L. Boureau, F. Bach, Y. LeCun, and J. Ponce, "Learning Mid-Level Features For Recognition," in *CVPR*, 2010, pp. 2559–2566.

[29] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond Bag of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories," in *CVPR*, 2006, pp. 2169–2178.