Laser-based Segment Classification Using a Mixture of Bag-of-Words

Jens Behley, Volker Steinhage, and Armin B. Cremers

***Abstract*— In this paper, we propose a segment-based object detection approach using laser range data. Our detection approach is built up of three stages: First, a hierarchical segmentation approach generates a hierarchy of coarse-to-fine segments to reduce the impact of over- and under-segmentation in later stages. Next, we employ a learned mixture model to classify all segments. The model combines multiple softmax regression classifiers learned on specific bag-of-word representations using different parameterizations of a descriptor. In the final stage, we filter irrelevant and duplicate detections using a greedy method in consideration of the segment hierarchy. We experimentally evaluate our approach on recently published real-world datasets to detect pedestrians, cars, and cyclists.**

I. Introduction

The detection and recognition of potentially moving objects is crucial for autonomous systems operating in populated environments. Especially self-driving cars need to distinguish reliably between static and dynamic objects, such as pedestrians, vehicles, and bicyclists, to ensure safe operation in crowded and even uncooperative inner-city traffic.

Over the last decade a multitude of image-based approaches for object detection were proposed and achieved promising results on challenging benchmark datasets [1], [2], [3]. As fast three-dimensional laser rangefinders emerged, laser-based object recognition for outdoor applications attracted increasing interest in the robotics community [4], [5], [6]. Laser range scans are an interesting alternative to images, as they are invariant to illumination and directly offer shape information. Furthermore, precise range measurements to objects in the vicinity are essential for collision-free maneuvering [7]. Consequently, Velodyne laser sensors are a de facto standard equipment for self-driving cars [6], [8].

In image-based object recognition, bag-of-word approaches [2] are a well established concept, but in laser-based perception rarely applied. This is remarkable, since they offer by design several properties, which are desirable particularly in laser-based object recognition: (1) bag-of-words are robust to partial occlusions, (2) even if we encounter an under-segmentation, the entries for a certain class should be still visible in a part of the histogram, (3) point descriptors can be computed independently, which makes a concurrent evaluation possible. Thus, bag-of-words extracted from laser data are a fundamental building block of our approach.

Recent work on object detection [1], [3] suggests that it is crucial to incorporate intra-class variations of objects. It has been shown that the performance of an object recognition approach [1] can be significantly improved by learning a mixture of classification models, where specific detectors learn variations of a class. Felzenswalb et al. [1] use a bounding box criterion to initialize different mixture components of a class. In our approach, we will use distance, volume, and the extents of the 3d bounding box as latent variables and additionally learn every mixture component using different parameterizations of a histogram descriptor. This choice is motivated by the distance-dependency of three-dimensional scans, i.e., we can distinguish fine details at small distances, but get only a sparse point cloud at far distances.

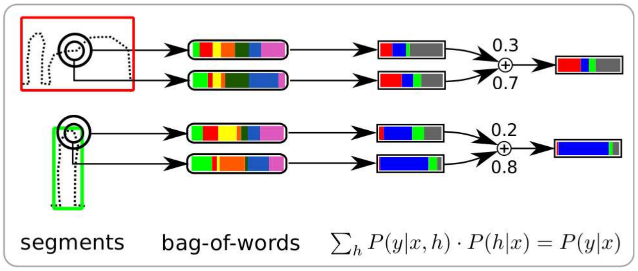

Fig. 1. Overview of our approach. We use a mixture of multiple bag-of-words classifiers learned with different descriptor parameterizations to classify segments generated by a hierarchical segmentation approach.

Overall, our approach is divided in the following three stages. First, we propose a hierarchical segmentation approach resulting in coarse-to-fine hierarchies of segments. Our aim is to reduce the effects of segmentation errors on later stages in the classification pipeline. We explicitly include over- and under-segmentations and let the later stages filter these additional segments.

In the second stage, we employ a mixture of multiple bag-of-word classifiers to classify all extracted segments (see figure 1). We use different parameterizations of a local descriptor for each classifier, which enables the overall approach to adapt to different aspects of the data. The results of the specialized classifiers are averaged using mixture weights jointly learned with the classifiers.

In the final step, we filter duplicated detections. We apply a greedy breadth-first search strategy to ensure consistent final detection hypotheses with maximal confidence.

Related work. Segmentation is a basic pre-processing step applied in many approaches dealing with large-scale 3d point clouds. The main purpose is the reduction of the overall number of points by discarding irrelevant segments and thus a more efficient overall processing.

Most approaches [9], [10], [6] apply two steps: (1) filtering of irrelevant ground points and (2) grouping of the remaining points into coherent and meaningful clusters. A common

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

approach for ground point filtering is a height-based filtering using an elevation map. Petrovskaya and Thrun [11] and also Himmelsbach et al. [10] exploit that ground points should not cause large height differences and inspect locally the smoothness in angular sectors. Klasing et al. [12] directly use the point cloud to determine segments by an efficient distance-based clustering, where each cluster is defined by points with a given maximal distance to each other. In contrast to these approaches, other solutions [13], [14] explicitly exploit the sensing method to attain real-time capable solutions with graph-based methods.

All approaches share a non-trivial selection of suitable parameters, which is usually specific to the task and object classes of interest [15]. We apply multiple stages of the elevation-based segmentation and are therefore more independent of a specific choice of parameters. Our approach possibly generates many more segments than really needed, but we rather filter these irrelevant segments later. Van der Sande et al. [2] also generate an over-complete hierarchy of segments in images, but do not exploit the hierarchy to eliminate duplicate detections.

Segment-based classification of 3d laser range data in urban environments was mainly investigated for classification of dynamic objects. Himmelsbach et al. [9] classify segments represented by a histogram of multiple point-based features and remission intensities using a SVM. Teichman and Thrun [6] use tracking information to smooth the segment-based classification results of an AdaBoost-based approach. The segments are represented by multiple spin images [16] with different resolutions and HoG-features calculated on projections of the point cloud. The feature set is additionally enriched by holistic features, which represent track-based properties, and spin images calculated over accumulated and aligned point clouds from multiple track positions. Himmelsbach et al. [17] use tracking information to correct under- and over-segmentations. In contrast to these approaches, we aim at learning multiple classifiers using a bag-of-words, where each bag uses a different descriptor parameterization.

In computer vision, several authors investigated ways to reintroduce spatial information in the bag-of-words approach [18], [19]. Parizi et al. [19] learn specific bag-of-words models for image regions. In our approach, we use a different probabilistic modeling approach and use specific descriptors per bag-of-word vocabulary.

From the machine learning perspective, our approach is closely related to mixture-of-experts [20], where multiple classifiers are jointly learned for parts of the feature space.

II. APPROACH

Our objective is to determine all segments belonging to the classes pedestrian, car, and bicyclist, using only a single 3d laser range scan. To this end, we view the detection problem as classification task and learn a classifier to output a probability distribution $P(y|\boldsymbol{x})$ for a segment $\boldsymbol{x}$ belonging to either to the target classes or background. In a post-processing step, we finally remove segments belonging to background and also non-maximal detections.

A. Hierarchical Segmentation

Model-free segmentation is usually less involved using laser range data compared to using solely images. This is caused by the availability of depth information, which separates objects from each other and the ground. Hence, in most cases less complex methods are sufficient to attain very good results. Despite this advantage, we still have to cope with under- and over-segmentation – especially in outdoor environments, where distances to objects range from few meters to more than 20 meters. Hence, the point cloud density varies drastically leading to difficulties in finding suitable parameters for distance-based segmentation methods, which result in coherent segments for different ranges.

In this work, we are not aiming at generating a single perfect segmentation, but to generate multiple coarse-to-fine segmentations. Later, we will use the classification results and a greedy filtering approach to remove irrelevant or duplicated segments.

Basic building block of the proposed hierarchical segmentation approach is an efficient elevation map-based segmentation [6], [9]. We start with a partitioning of the scan into a regular grid with grid cell size r_0 and record for every grid cell the smallest and largest height of points. Then, we find connected components of adjacent cells with point height differences larger than a threshold η.

For every segment, we further apply this height-based segmentation, but now with a smaller resolution $r_{i+1} < r_i$ until we reach the desired depth. Thus, we get a smaller grid and consequently can subdivide a segment into smaller sub-segments, if necessary. We get multiple trees containing at every level a finer segmentation of the original point cloud.

B. Learning a mixture of bag-of-words

Given the segment trees, we determine multiple bag-of-word representations using only points from each segment. In particular, we learn multiple vocabularies on subsets of the training data using differently parameterized descriptors.

We are interested in a discriminative classification approach $P(y|\boldsymbol{x})$, and for our purposes we introduce a latent variable h:

$$P(y|\boldsymbol{x}) = \sum_h P(y, h|\boldsymbol{x}) \tag{1}$$

$$= \sum_h P(h|\boldsymbol{x})P(y|h, \boldsymbol{x}) \tag{2}$$

The value of the hidden variable $h \in \{1, \dots, M\}$ depends on the segment $\boldsymbol{x}$ and for each hidden variable we learn a separate multi-class classifier $P(y|h, \boldsymbol{x})$, where $y \in \{1, \dots, K\}$.

We learn for both models $P(h|\boldsymbol{x})$ and $P(y|h, \boldsymbol{x})$ a softmax regression model [21]:

$$P(h = j|\boldsymbol{x}) = \frac{\exp(\boldsymbol{w}_j^T \cdot \boldsymbol{x})}{\sum_k \exp(\boldsymbol{w}_k^T \cdot \boldsymbol{x})} \tag{3}$$

$$P(y = k|h, \boldsymbol{x}) = \frac{\exp(\boldsymbol{w}_{k,h}^T \cdot \boldsymbol{x})}{\sum_l \exp(\boldsymbol{w}_{l,h}^T \cdot \boldsymbol{x})} \tag{4}$$

Here $\boldsymbol{w}_h$ and $\boldsymbol{w}_{y,h}$ represent the weight vectors for every latent variable h and class y, respectively. In the following, we summarize these parameter vectors of all models by $\boldsymbol{\theta}_t = (\boldsymbol{w}_1, \dots, \boldsymbol{w}_M, \boldsymbol{w}_{1,1}, \dots, \boldsymbol{w}_{K,1}, \dots, \boldsymbol{w}_{1,M}, \dots, \boldsymbol{w}_{K,M})$, where t denotes the iteration in the optimization process.

We jointly estimate the parameters of $P(y|\boldsymbol{x}, h)$ and $P(h|\boldsymbol{x})$ using Expectation Maximization [21]. In the E-step, we estimate the distributions $q_h(\boldsymbol{x}_i, y_i)$ for every training instance $(\boldsymbol{x}_i, y_i)$ by $P(h|\boldsymbol{x}_i, y_i)$ using the parameters $\boldsymbol{\theta}_{t-1}$ from the last iteration $t-1$:

$$q_h(x_i, y_i) = P(h|\boldsymbol{x}_i, y_i) \tag{5}$$

$$= \frac{P(y_i|h, \boldsymbol{x}_i)P(h|\boldsymbol{x}_i)}{\sum_k P(y_i|\boldsymbol{x}_i, k)P(k|\boldsymbol{x}_i)} \tag{6}$$

The log-likelihood $L(\boldsymbol{\theta})$ in the M-Step is given by

$$L(\boldsymbol{\theta}_t) = \sum_i \sum_h q_h(\boldsymbol{x}_i, y_i) \log \left[P(h|\boldsymbol{x}_i) P(y_i|h, \boldsymbol{x}_i) \right] \tag{7}$$

Hence, the gradients in respect to the parameters $\boldsymbol{w}_j$ and $\boldsymbol{w}_{k,h}$ are given by:

$$\frac{\partial L}{\partial \boldsymbol{w}_j} = \sum_i \sum_h q_j(\boldsymbol{x}_i, y_i) \left[\mathbf{1}\{h = j\} - P(j|\boldsymbol{x}_i) \right] \boldsymbol{x}_i \tag{8}$$

$$\frac{\partial L}{\partial \boldsymbol{w}_{k,h}} = \sum_i q_h(\boldsymbol{x}_i, y_i) \left[\mathbf{1}\{y_i = k\} - P(k|h, \boldsymbol{x}_i) \right] \boldsymbol{x}_i, \tag{9}$$

where $\mathbf{1}\{\cdot\}$ denotes the indicator function returning 1, if the condition is true, and 0 otherwise.

In summary, the complete training of the mixture components involves the following steps:

1) Estimate $q_h(\boldsymbol{x}_i, y_i)$ for every labeled segment $(\boldsymbol{x}_i, y_i)$ using eq. (6) and the parameters $\boldsymbol{\theta}_{t-1}$ from last iteration $t-1$.
2) Re-learn vocabularies $\mathcal{V}_k$ over subset $\mathcal{X}_k = \{\boldsymbol{x}_i | q_k(\boldsymbol{x}_i, y_i) \geq q_h(\boldsymbol{x}_i, y_i)\}$.
3) Maximize eq. (7) with respect to $\boldsymbol{\theta}$ after encoding every segment using the newly learned vocabularies $\mathcal{V}_k$.

C. Hierarchical Non-maximum Suppression

Using the learned mixture model, we classify every segment in all hierarchies and for every segment get $P(y|\boldsymbol{x})$. As we now might have contradicting classification in a hierarchy, we have to determine which of the segments are likely to be a correct hypothesis and suppress non-maximal detections.

For this purpose, we use a greedy algorithm starting at the root of every hierarchy and descend the tree in breadth-first order. Background segments are not reported. We mark a segment for the final set of reported segments, when the overlap with non-background parent nodes is smaller than than a threshold γ. In this case, we assume that we found a smaller segment, which is for itself a valid detection, such as a person standing by a car. If the overlap between a node and an ancestral node is larger than γ, we suppress the non-maximal detection, i.e., the hypothesis where $P(y|\boldsymbol{x})$ is smaller. Thus, if an ancestral node classifies a segment differently at a coarser level, we only report the detection with larger confidence.

III. EXPERIMENTS

In this section, we experimentally evaluate segmentation and detection on challenging real-world datasets. First, we compare the proposed hierarchical segmentation with a single layer height-based segmentation. Then, we will use the hierarchical segmentation to extract segments and classify these segments either with a single bag-of-words or the proposed mixture of bag-of-words. Lastly, we compare the single bow and mixture of bow model on pre-segmented data.

Datasets. For evaluation of the complete pipeline, we use the recently published KITTI Vision Benchmark Dataset [8]. We additionally used the Stanford Track Collection (STC) [6] for experiments using the classification model only. All data was recorded using a car equipped with common sensors used in autonomous driving, including a Velodyne laser rangefinder and an inertial navigation system for odometry information. In both datasets, we have to classify cars, cyclists, and pedestrians in everyday traffic situations.

The KITTI dataset contains $7{,}481$ annotated images with additional Velodyne scans and appropriate calibration information. Additionally $7{,}518$ unlabeled test images with laser range scans are provided, where we have to annotate the image with bounding boxes. We have to emphasize that we solely use the laser scans in the following experiments and therefore project scan points using the provided calibration into the image to estimate an image-based bounding box. The detections are evaluated and scored following common image-based detection metrics [22] and must be send to a server-side evaluation script. Thus, we present here results for different parameter values using the training set only and will report results on the testset for a specific parameter setting later. All bounding boxes are annotated with a class label, an occlusion ratio and a truncation value. Depending on these values, bounding box difficulties[1] were defined by Geiger et al. [8], which we will use in the following discussion.

The STC dataset contains roughly $14{,}000$ tracks with segments extracted by a height-based segmentation and 83.3% of all segments are background. Note that we get pre-segmented laser scans and therefore evaluate here only the classification model, either using a single vocabulary or the proposed mixture of multiple vocabularies. We report the classification accuracy of the classifiers.

All reported timings were measured on a system equipped with an Intel Xeon X5550 with $2.67\,\text{GHz}$ and $12\,\text{GB}$ memory using a single thread implementation.

Implementation details. We calculated spin images [16], since these showed good results for point-wise classification in earlier experiments [23] and are relatively fast to compute. All descriptors are calculated using a global reference frame, i.e., we use the z-axis to determine the bin in the histogram. We used for all spin images 5 bins per dimension and performed a bilinear interpolation to calculate the contributions

[1] Bounding boxes are classified into three categories: (1) *easy*: > 40 pixel height, fully visible, $< 15\%$ truncation, (2) *moderate*: > 25 pixel height, at least partial visible, $< 30\%$ truncation, (3) *hard*: > 25 pixels height, at most difficult to see, < 0.5 truncation

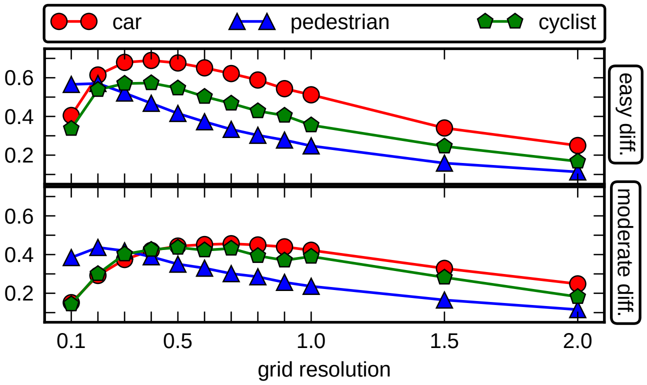

Fig. 2. Overlap with ground truth annotations. Shown is the overlap of the single layer segmentation for 'easy' segments and 'moderate' segments. We get different optimal grid resolutions depending on class and distance.

of every neighboring point. Every descriptor vector is finally normalized using the maximum norm L_∞.

We learn the vocabularies using off-the-shelf k-means clustering [24] and encode the descriptors using a hard quantization, i.e., we search in a kD-tree [25] for the nearest cluster center. Finally, we normalize the resulting bag-of-words vector using the L_1 norm.

A. *Bounding box overlap*

In the first experiment, we investigate the performance of the proposed hierarchical segmentation. For this purpose, we generate segments for all provided training data using either a single layer, two-layer, or three-layer hierarchy. The laser points extracted by these approaches are then projected into the image and an image-based bounding box is determined. For all approaches, we used a minimum height $\eta = 0.3$ and discarded segments with fewer than 50 laser points.

Next, we determine the maximal overlap o_i between annotated bounding boxes A_i and generated bounding boxes B_j using $o_i = \max_j \mathrm{Area}(A_i \cap B_j)/\mathrm{Area}(A_i \cup B_j)$ [22]. The overall overlap score is then averaged over all N scans: $O = N^{-1} \sum_i o_i$.

Figure 2 depicts the class-wise performance of the single layer segmentation with different grid resolutions. As motivated in the beginning, we can see that a generic choice of the resolution parameter is difficult. While for pedestrians a smaller grid is optimal to reduce over- or under-segmentation, the resolution should be larger for cars and cyclists. But also for different distances, we can observe a dependence: nearby objects are better segmented using a smaller resolution, while objects at larger distances are better segmented using a larger resolution. This dependence is hardly surprising, since laser points show a larger sparsity and distance to each other at large distances.

Table I shows the best results of a single, two- and three-layer segmentations, where we selected the best configuration for each segmentation approach using the moderate overall overlap. As can be seen from these results, the proposed multi-layer segmentation approaches clearly outperform the single-layer approach. Especially, the results for pedestrian (increase of up to 0.25 overlap) and cyclist (increase of up to 0.14 overlap) are noteworthy.

resolutions	all	car	pedestrian	cyclist
(0.7)	0.53/0.45	0.62/0.46	0.33/0.30	0.47/0.43
$(1.0, 0.5)$	0.60/0.49	0.68/0.50	0.43/0.37	0.53/0.45
$(1.0, 0.5, 0.2)$	0.69/0.51	0.73/0.52	0.58/0.49	0.61/0.47

TABLE I

OVERLAP RESULTS FOR HIERARCHICAL SEGMENTATION (EASY/MODERATE BOUNDING BOXES).

	pedestrian	car	cyclist	background
training	2090/1140	5400/8844	584/401	152158/23827
validation	220/119	571/895	70/43	n/a

TABLE II

SEGMENTS PER CLASS (EASY/MODERATE BOUNDING BOXES)

Despite the significant performance gain, we still have only a maximal average overlap of 0.7 between image-based and laser-based bounding boxes. A major drawback of laser rangefinders is that black objects can not be detected. Therefore a lot of black cars, which can be easily marked in an image, are simply invisible in the laser range data or only represented by non-black parts in the point cloud. Furthermore, glass is often not sensed by the laser sensor either and hence we get very few points on car windows. Thus, segments of cars at larger distance usually do not include the roof part and overlap consequently only partly the annotation in the image, which includes also the windows.

B. *Detection performance*

As introduced earlier, the results presented in this section are generated using a randomly selected validation set. For this purpose, we selected 10% of the training laser scans uniformly at random (see Table II). For training and validation set, we applied the three-layered hierarchical segmentation with $r_0 = 1.0, r_1 = 0.5$, and $r_2 = 0.2$ and ignored segments with less than 50 points and width or length larger than $6\,m$. In the training data, background segments were discarded if the image-based overlap to ground truth annotation was larger than 0.2. We used $\gamma = 0.5$ for the hierarchical non-maximum suppression.

The performance of bag-of-words approaches is primarily influenced by the size of the vocabulary and the choice of the descriptor. Figure 3 shows the influence of the size of the vocabulary and the results for different support radii of the spin images (0.5, 1.0 and $2.0\,m$ radius) with 5 bins in each dimension. The smallest spin image with a support radius of $0.5\,m$ clearly outperforms the larger spin images with larger support radii for the detection of pedestrians. Fine details are more important for the distinction between background and pedestrian. The performance of the other classes is less effected by a specific choice of the radius.

In line with earlier studies on bag-of-words in image-based classification [26], [27], we can also conclude that a larger vocabulary size is beneficial in our application. Especially, in

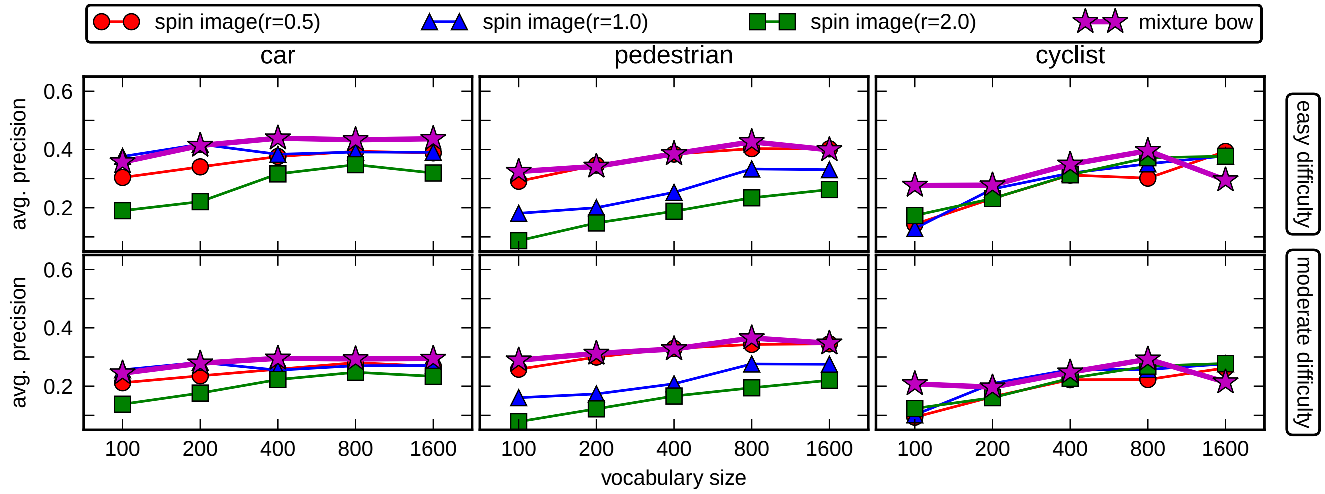

Fig. 3. Influence of the codebook size on single-layer bow with different parameterizations of spin images and the proposed mixture of bag-of-words. The upper row depicts the results for ’easy’ bounding boxes, and the lower row shows the results for ’moderate’ bounding boxes.

approach	car	pedestrian	cyclist
LSVM-MDPM-sv [28]	0.68/0.56	0.47/0.39	0.38/0.29
LSVM-MDPM-us [1]	0.66/0.55	0.45/0.38	0.35/0.27
Mixture bow	0.36/0.23	0.44/0.31	0.28/0.21

TABLE III

RESULTS ON THE KITTI TESTSET (EASY/MODERATE).

approach	car	pedestrian	cyclist	overall
AdaBoost [6]	95.8%	98.3%	98.4%	93.1%
Mixture bow	95.0%	98.3%	98.4%	92.3%
single bow (1.0 m)	91.6%	97.8%	97.7%	87.8%
single bow (2.0 m)	91.7%	97.5%	96.8%	86.7%
single bow (0.5 m)	89.4%	97.8%	96.3%	83.8%

TABLE IV

CLASSIFICATION ACCURACY FOR THE STC DATASET.

case of cyclists and pedestrians we see a significant increase in performance with more words.

The mixture of bag-of-words combines all three descriptor radii and in the first iteration of the EM algorithm we split the training data depending on the distance of the bounding box into three subsets. However, the hidden variable model $P(h|\boldsymbol{x})$ is learned using distance, volume, and the extent of the 3d bounding box. The mixture of bag-of-words improves the results especially with smaller vocabularies.

Table III finally shows the average precision of our approach compared to image-based approaches on the testset. We choose a vocabulary size of 800 as this showed the best performance in the experiments on the validation set. The other image-based approaches use a latent variable model of Felzenswalb et al. [1] and an extension of this approach by Geiger et al. [28]. We have to emphasize again that we solely use laser range information and compare all approaches with image-based overlap metrics. Thus, the extracted segments and consequently the bounding boxes are affected by the insufficiencies of the laser rangefinder discussed earlier.

We often see false positive detections of cars in areas with vegetation and pedestrians are often confused with pole-like structures or small bushes. Another reason for wrong detections are mismatches between the annotated image-based bounding box and the bounding boxes generated from the laser data. Particularly, the car detections are strongly affected by too low overlap values, as we need at least 0.7 overlap between ground truth annotation and detections instead of 0.5 overlap for the other classes.

The complete classification of a single frontal laser range scan currently needs $2.53\,\mathrm{s}$ on average, where almost all time ($2.46\,\mathrm{s}$) is needed to calculate the bag-of-words. The hierarchical segmentation using three layers needs $15.7\,\mathrm{ms}$ on average.

C. Classification performance

Table IV show the results on the STC dataset in comparison to an AdaBoost-based approach presented by Teichman et al. [6]. In these experiments, we used the provided segmentations and applied our classification model with a single bag-of-words and the mixture of multiple bag-of-words. In contrast to the previous experiments, we used $1{,}600$ words for each bag-of-words vocabulary, but the other parameters remained unchanged. The presented results clearly show the advantage of the mixture of multiple vocabularies over a single vocabulary and comparable performance to the state-of-the-art.

IV. CONCLUSION AND FUTURE WORK

In this paper we introduced an approach for segment-based classification using a mixture of different bag-of-words vocabularies. For segmentation, we proposed a new hierarchical combination of coarse-to-fine segmentations, which allowed us to extract more reliably suitable segments. We have shown that a mixture of bag-of-word classifiers outperforms a single vocabulary bag-of-words approach on challenging real-world datasets. Finally, we presented a greedy non-maximum suppression considering the hierarchy of segments.

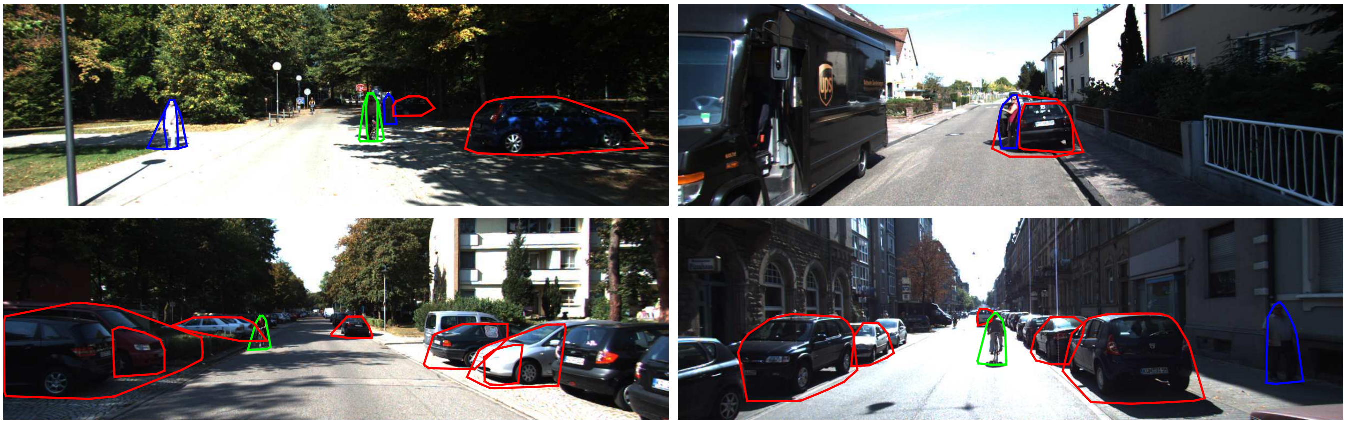

Fig. 4. KITTI detection results of the mixture of bag-of-words for cars (red), pedestrian (blue), and cyclists (green). We show the convex hulls of the projected laser range points.

Based on these promising results, we will investigate more efficient ways of computing the bag-of-word representation – either using sub-sampling or a concurrent computation. Furthermore, we plan to investigate other methods for vocabulary learning and encoding [26], [27], which could further improve the classification results. Using other type of side information, e.g. images or map data, could be exploited to detect black objects or might be used to filter false positives. Integration of tracking information to smooth classification results [6] or fix segmentation errors [17] is also a promising avenue for future research.

V. ACKNOWLEDGMENTS

Thanks to Andreas Geiger for quick help with the KITTI dataset and providing calibration matrices for the test dataset. Special thanks to Florian Schöler, Daniel Seidel, and Jenny Balfer for fruitful discussions.

REFERENCES

[1] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, “Object Detection with Discriminatively Trained Part-Based Models,” *TPAMI*, vol. 32, no. 9, pp. 1627–1645, 2010.

[2] 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.

[3] Q. Chen, Z. Song, Y. Hua, Z. Huang, and S. Yan, “Hierarchical Matching with Side Information for Image Classification,” in *CVPR*, 2012, pp. 3426–3433.

[4] 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.

[5] X. Xiong, D. Munoz, J. A. Bagnell, and M. Hebert, “3-D Scene Analysis via Sequenced Predictions over Points and Regions,” in *ICRA*, 2011, pp. 2609–2616.

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

[7] F. Hoeller, T. Röhling, and D. Schulz, “Offroad Navigation using Adaptable Motion Patterns,” in *ICINCO*, 2010, pp. 186–191.

[8] A. Geiger, P. Lenz, and R. Urtasun, “Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite,” in *CVPR*, 2012, pp. 3354–3361.

[9] M. Himmelsbach, T. Luettel, and H.-J. Wuensche, “Real-time Object Classification in 3D Point Clouds Using Point Feature Histograms,” in *IROS*, 2009, pp. 994–1000.

[10] M. Himmelsbach, F. v. Hundelshausen, and H.-J. Wuensche, “Fast Segmentation of 3D Point Clouds for Ground Vehicles,” in *IV*, 2010, pp. 560–565.

[11] A. Petrovskaya and S. Thrun, “Model Based Vehicle Detection and Tracking for Autonomous Urban Driving,” *AuRo*, vol. 26, no. 2-3, pp. 123–139, 2009.

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

[13] 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.

[14] K. Klasing, D. Wollherr, and M. Buss, “Realtime Segmentation of Range Data Using Continuous Nearest Neighbors,” in *ICRA*, 2009, pp. 2431–2436.

[15] B. Douillard, J. Underwood, N. Kuntz, V. Vlaskine, A. Quadros, P. Morton, and A. Frenkel, “On the Segmentation of 3D LIDAR Point Clouds,” in *ICRA*, 2011, pp. 2798–2805.

[16] 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.

[17] M. Himmelsbach and H.-J. Wuensche, “Tracking and Classification of Arbitraty Objects with Bottom-Up/Top-Down Detection,” in *IV*, 2012, pp. 577–582.

[18] 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.

[19] S. N. Parizi, J. Oberlin, and P. F. Felzenszwalb, “Reconfigurable Models for Scene Recognition,” in *CVPR*, 2012, pp. 2775–2782.

[20] R. A. Jacobs, M. I. Jordan, S. Nowlan, and G. E. Hinton, “Adaptive mixtures of local experts,” *Neural Comp.*, vol. 3, pp. 1–12, 1991.

[21] S. Prince, *Computer Vision: Models, Inference and Learning*. Cambridge University Press, 2012.

[22] M. Everingham, L. V. Gool, C. K. I. Williams, J. Winn, and A. Zisserman, “The PASCAL Visual Object Classes (VOC) Challenge,” *IJCV*, vol. 88, no. 2, pp. 303–338, 2010.

[23] 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, pp. 4391–4398.

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

[25] 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.

[26] K. Chatfield, V. Lempitsky, A. Vedaldi, and A. Zisserman, “The devil is in the details: an evaluation of recent feature encoding methods,” in *BMVC*, 2011, pp. 76.1–76.12.

[27] 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.

[28] A. Geiger, C. Wojek, and R. Urtasun, “Joint 3D Estimation of Objects and Scene Layout,” in *NIPS*, 2011.