Object Detection From a Few LIDAR Scanning Planes

Zoltan Rozsa, *Member, IEEE*, and Tamas Sziranyi, *Senior Member, IEEE*

***Abstract*—LIDAR sensors enable object and free-space detection for intelligent transportation systems and vehicles. This paper proposes a recognition method for LIDARs based on only a few detection planes. This method is useful especially in the case when the angular resolution of the scan is sufficient, but in the vertical direction the planes are far from each other. We use Fourier descriptor to characterize a scan plane and Convolutional Neural Network for classification. Our method exploits both time varying shape information and contours from multiple scan planes if available. The method performs at least as well as the state of the art algorithms in case of near field, and it also expands the detection range. It was evaluated on tens of thousands of samples from large public datasets and we did separate evaluation for far field objects as well.**

***Index Terms*—LIDAR, autonomous vehicle, object classification, remote sensing.**

I. Introduction

IN AUTONOMOUS driving different sensors proved to be efficient for different tasks: cameras for object recognition or depth sensors (e.g. LIDARs) for free-space or object candidate detection. However, to ensure road safety, different sensor modalities have to work together [1], [2]. We have to maximize the efficiency of each distinct sensor modality for each task, in order to minimize the probabilities of accidents in all circumstances. In this work we improve the overall classification performance of LIDAR sensors by widening their application range for faraway cases. Most of the layer-object intersections (in case of multi-layer LIDARs) are cone-based ones instead of plane. However in the target range, the produced object segments can be approximated as planar segments, so we will refer to them with this term.

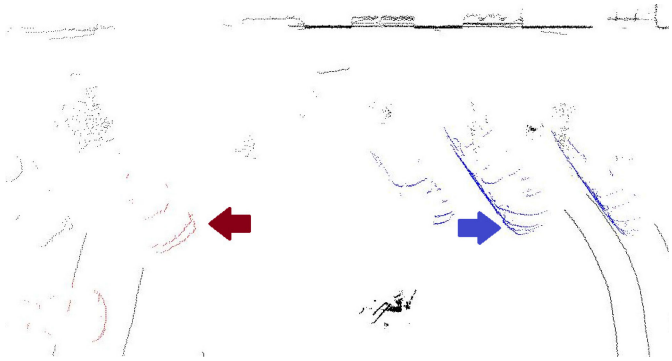


Fig. 1. Velodyne VLP16 sequence. Cars which are represented only with 2-3 detection planes (there is no extractable local surface information) marked with red points. Blue points correspond to objects, where local surface information can be extracted (point clouds with higher vertical resolution).

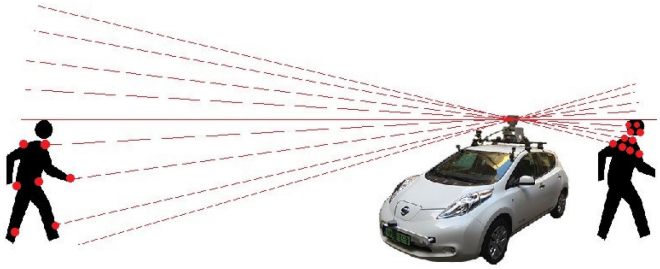


Fig. 2. Illustration of far object problem (Electric car of MTA SZTAKI equipped with GoPro cameras, Velodyne HDL-64 and VLP-16 sensors).

Intelligent vehicles are often equipped with LIDARs of relatively good vertical resolution (e.g. Velodyne HDL-64[1]), but experience shows that far objects will be represented with only a few planes and local surface features cannot be extracted (Figs. 1 and 2) thus, methods based on surface features cannot be used. Another frequent scenario is the case when a vehicle is equipped with LIDARs only providing a few detection planes (e.g. SICK LD-MRS[2] or Velodyne VLP-16[3]) or in some cases only one plane (e.g., SICK LMS5xx series[4]). In [3] we proposed a solution for recognition problem of Automated Guided Vehicles (AGVs),

Manuscript received May 14, 2018; revised December 11, 2019 and February 15, 2019; accepted March 11, 2019. Date of publication August 28, 2019; date of current version November 21, 2019. This work was supported in part by the Higher Education Excellence Program of the Ministry of Human Capacities in the frame of Artificial Intelligence research area of Budapest University of Technology and Economics under Grant BME FIKP-MI/FM and in part by the Hungarian Scientific Research Fund under Grants OTKA/NKFIH K_120499 and KH_125681. *(Corresponding author: Zoltan Rozsa.)*

The authors are with the Machine Perception Research Laboratory, The Hungarian Academy of Sciences Institute for Computer Science and Control (MTA SZTAKI), H-1111 Budapest, Hungary, and also with the Faculty of Transportation Engineering and Vehicle Engineering, Budapest University of Technology and Economics (BME KJK), H-1111 Budapest, Hungary (e-mail: rozsazoltan@sztaki.hu; sziranyi.tamas@sztaki.mta.hu).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIV.2019.2938109

[1]http://velodynelidar.com/hdl-64e.html

[2]https://www.sick.com/us/en/detection-and-ranging-solutions/3d-lidar-sensors/ld-mrs/c/g91913

[3]http://velodynelidar.com/vlp-16.html

[4]https://www.sick.com/us/en/detection-and-ranging-solutions/2d-lidar-sensors/lms5xx/c/g179651

where we proposed 3D reconstruction by fusing the different scanning planes of LIDARs producing only a low number of such planes. In case of autonomous vehicles, even faster decision is required than in case of AGVs, because of higher speed. Decisions have to be made based on the recent few scans; there is no time for the accumulation of more scans. This paper proposes a solution to this problem by handling all the object candidates as a set of plane curves. We will show that these plane curves are suitable for object recognition by considering their change over time and that increasing the number of scan planes increases the recognition probability.

Since recent works (e.g. [4], [5]) show good detection performance for a few categories (about 95% recall for three categories) on 2D LIDAR scans, we aim to enhance the recognition methods by expanding the application range to the limit of the scanning resolution.

There is a pressing need in transportation automation for object and situation recognition. Numerous methods are available for LIDAR based object detection and classification in the literature ([6], [7]) with high performance. However, all these methods operate in the "close" range. The exact range is method and sensor dependent, but generally the target objects are in the range of 10–20 m and not farther than 30 m. LIDAR-based recognition uses (at least 2.5D) surface features, and these features are not present in the strongly discontinuous point cloud segments that are further away at a larger distance. So, extending the processing range is practically impossible for methods that use surface features. Still, an increase in processing range would preferable to enable processing in higher speed scenarios. One possible solution is to fuse different sensor modalities. LIDAR sensors can still be used to detect points at larger distance, however without usable surface information. We propose a solution to exploit the available information as much as possible. The proposed method's performance is not influenced by range and comparable to the state of the art (2D and 2.5D).

Main contributions of the paper are:

- New approach for plane curve description;
- New object representation: as set of time varying plane curves;
- Possibility to evaluate plane curve groups (in volume or tracked through time);
- We propose a maximum likelihood scheme in order to increase the recognition probability by using all the available aggregated data;
- Offer solution to recognition problems caused by objects present only in a few layers in case of any LIDARs.

This paper is organized as follows: Section II surveys the literature about the current state of the art in the topic. Section III describes the proposed method in detail. In Section IV some sensor limitations are discussed and datasets used for evaluation are presented. Sections V and VI show our test results and compare them to other methods. First, our initial class prediction is introduced, which is superior to the compared methods, then results based on tracking are presented, which further increases our recognition performance. Finally, Section VII draws some conclusions about the topic.

II. Related Works

Methods working on 3D LIDARs like [6] and [7] use convolutional networks of coherent 3D point clouds as input, while in our case only separated 2D LIDAR segments are available. For this problem of plane curves, we propose a new approach, by statistically combining the information of the separated 2D curves. In the following, we investigate the literature of object classification for LIDARs having one or a few scanning planes.

Nowadays, 2D or 3D laser scanners are often used in autonomous driving and robotics. Algorithms which aimed object detection [8] and tracking [9] with laser range finders were already presented in the 2000s. In these early approaches, more than one object class was not considered; the primary goal was finding and tracking people. Today, with the development of computer vision algorithms and sensors it is common to aim for the recognition of multiple classes in such planar contour data. The authors of [4] were capable of differentiating four categories with good accuracy based on Euclidean distance, using the width of an obstacle and the measured intensity. By adding the range variance feature to the descriptor [10], they were able to increase the classification accuracy. Other approaches like the one presented in [11] converted the detected blobs to a 5×5 binary image and used SVM (Support Vector Machine) to classify the objects into two categories (vehicles or pedestrians); other proposed a distant-invariant feature for segmentation and detection of people without walking aids, people with walkers, people in wheelchairs and people with crutches [12].

Defining simple geometric features and building a strong classifier from them (e.g. with AdaBoost) is a common approach in the classification of 2D contour data. The authors of [13] predict human shape by detecting different body parts at different heights using more than 10 features acquired from the scans and AdaBoost algorithm. A similar approach is presented by [14] using multiple laser rangefinders and by [15] using a multi-layer one. Motion characteristics also can be used to identify humans with baby cart, shopping cart or wheel chairs [16]. [17] used time-varying plane curve representation to solve the stereo-vision based multiple object tracking problem.

A. Properties and Disadvantages of Available Methods

To summarize, the properties of currently available methods for classification based on one or a few planar scans:

- In most cases they are based on geometrical features and either Adaboost method or neural network ([5], [18]) methods to build a strong classifier.
- Time-variant information is rarely used [19], and information via multiple planes is rarely exploited (only searching for specific body parts), and to the best of our knowledge, these two have never been utilized simultaneously.
- The above methods are mostly designed for the classification of indoor (e.g. industrial hall) objects scanned with indoor sensors with limited range (they depend on the sensor's range and angular resolution).
- Only a few classes are considered for detection.

- The methods are tested on few thousands samples at best [5].

B. Advantages Offered by the Proposed Method

There is a need for robust solution for the problem of recognizing far field outdoor objects, for application in high-speed autonomous vehicles. Even in best case scenarios (Velodyne VLS-128[5] with very high vertical resolution) in range of 150 m (half of its official range, on the highway a 130 km/h vehicle approaches the object in about 4 s) the LIDAR cannot see a 1.5 m height object in more than five planes, and it only sees it discontinuously (methods invented for 2.5D or 3D point clouds could not work).

We would like to address the above issues with the advantages listed below:

- We propose a method that extends the classification possibility of LIDARs with a few layers and far field classification of 3D LIDARs by utilizing both time-varying shape and multiple plane information.
- Our method designed for outdoor objects is invariant from the sensor type.
- It is model-free; we do not restrict it by assuming any relation between the sensor planes.
- We evaluate test results via tens of thousands of samples.
- Our model is able to evaluate the range from single layer scanning to tracked and/or multiple layers' scanning; and it has the same training framework for the feature-level (one plane), object-level (multiple planes in a frame) and tracked object-level recognition.

III. The Proposed Method

Preliminary results were presented about object representation as set of plane curves in a volume in [20]. The main contribution of this paper is the multiple level classification procedure, but we enclose the full pipeline we used in our test sequences and propose it to use for the whole process. In the following, we assume to work with a LIDAR sensor with a few scanning planes. First, the proposed preprocessing steps, then the classification method are presented. In the following, we will refer to the point cloud acquired during one full 3D scan period of the LIDAR sensor as one frame.

A. Preprocessing of LIDAR Clouds

Our point cloud processing pipeline is based on simple and robust methods. Known point cloud preprocessing methods from the literature are listed here in the order of processing that we used in our experiments to segment objects of scanning planes:

1) Registering consecutive frames: Iterative Closest Point (ICP) [21].
2) Ground detection: M-estimator SAmple Consensus (MSAC) Plane fitting [22].
3) Object detection: Euclidean cluster extraction [23] with distance varying neighborhood radius.
4) Objects are separated into plane curves, which are continuously matched in the consecutive frames (using features from simple geometry).

These methods are frequently applied together for similar purposes, e.g. [24]. In case of moving objects the preprocessing has to be extended with change detection (determining points with significant change by algorithm proposed in [25] called M3C2); moving object detection (objects which have many points with significant change); and tracking (gathering location, extension, velocity and orientation to a vector, James Munkres's variant of Hungarian assignment algorithm [26] is used to match the candidates). These processing steps are shown in Fig. 3. Note that: Stationary objects' shape is changing along the viewpoint change of a moving LIDAR sensor.

B. Descriptor of Planar Object Segment

Assuming the above preprocessing steps have been performed, here the objects are represented by plane curves tracked through several frames. We use a $f * (n + 6)$ matrix as a descriptor of a LIDAR segment. Here, f indicates the number of frames the segment is tracked and n represents the number of Fourier descriptor components [27] and 6 is the number of statistical measures we compare, explained in Section III-B2. If n is higher the representation is more precise, however it will require segments composed of a higher (minimum) number of points [28]. The composition of this descriptor is explained in the following:

1) Fourier Descriptor: The Fourier descriptor can be used to reconstruct the exact curve, so we use this property instead of extracting geometric features [29]. We consider the segment as a closed contour (we construct an ordered 2D point cloud, by copying the original points in reverse order to the end of the original point cloud). Translation and rotation invariance is achieved by removing the mean from the 2D point cloud and by using absolute values in Fourier space. We found that this representation shows robustness against varying point density as well. The kth frequency component can be calculated by the Discrete Fourier transform (DFT) of the contour (2D point cloud) with center of gravity in the origin, defined as:

$$C(k) = \frac{1}{N}\sum_{m=0}^{N-1}(c(m) - c_0)e^{\frac{-i2\pi mk}{N}} \tag{1}$$

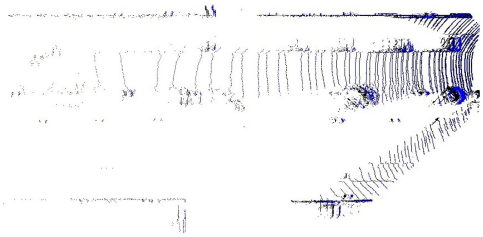
where N is the number of contour points, $c(m) = x(m) + iy(m)$, $x(m)$ and $y(m)$ are the x and y complex coordinates of the mth contour point, $c_0 = \frac{1}{N}\sum_{m=1}^{N} c(m)$, and i is the imaginary number. The Fourier descriptor we use, utilizes the first n frequency component: $FD(j) = |C(j)|$ $(j = 1, \ldots, n)$

2) Statistical Measures: Our descriptor also contains the mean (μ) and standard deviation (σ) values of altitude, distance to the sensor and intensity values along the planar curve.

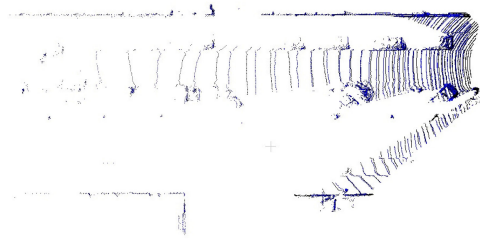
$$\mu_A = \frac{1}{N}\sum_{m=1}^{N} A_m \tag{2}$$

$$\sigma_A = \sqrt{\frac{1}{N-1}\sum\nolimits_{m=1}^{N} |A_m - \mu_A|^2} \tag{3}$$

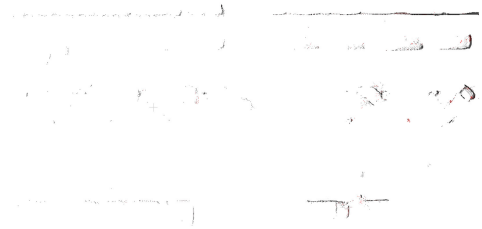
[5]https://velodynelidar.com/vls-128.html



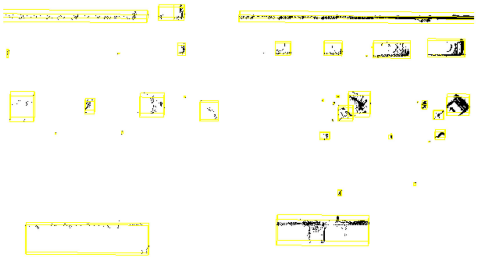
(a) Two consecutive frames without registration



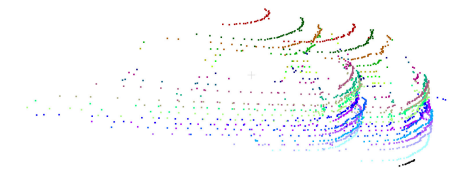
(b) Two consecutive frames registered



(c) Change detection (points with significant changes marked with red, there is no significant change in object level) on one frame without ground



(d) Detected objects



(e) Segments of a car are matched on two consecutive frames (Same color indicates the match). Note that: for illustration purposes we have chosen an object with several segments, however the method was designed primarily for objects with only a few of those.

Fig. 3. Example of preprocessing steps on KITTI tracking database.

where A substitutes the measures: altitude ($z(m)$), distance to the sensor in the $x-y$ plane ($r(m) = \sqrt{x(m)^2 + y(m)^2}$) and measured intensity values ($I(m)$).

3) Shape Variation Through Time: The values extracted in the previous subsections are stored through consecutive frames and stored as different rows in the descriptor matrix. This is illustrated in Table I.

TABLE I
EXAMPLE DESCRIPTOR (TRANSPOSE MATRIX) ABOUT THE CAR IN FIG. 5 USING $n = 5$ FOURIER DESCRIPTOR COMPONENTS. FDx INDICATES THE XTH FOURIER DESCRIPTOR COMPONENT, z IS THE ALTITUDE, r IS THE DISTANCE TO THE ORIGIN, I MEANS INTENSITY AND l IS THE FRAME NUMBER

Frame	l	$l-1$	$l-2$	$l-3$	$l-4$
FD(1)	0.5882	0.5885	0.6228	0.6396	0.5616
FD(2)	0.3791	0.3778	0.4107	0.4078	0.3325
FD(3)	0.2693	0.2662	0.2831	0.2712	0.2376
FD(4)	0.1794	0.1763	0.1755	0.1639	0.1649
FD(5)	0.0096	0.0953	0.0911	0.0944	0.1111
μ_z	0.6807	0.6946	0.7112	0.7247	0.6374
σ_z	0.0611	0.0606	0.0656	0.0648	0.0484
μ_r	12.8307	13.0329	13.2751	13.4702	13.5801
σ_r	0.8878	0.8818	0.9543	0.9434	0.7830
μ_I	0.4256	0.4243	0.4178	0.5896	0.4195
σ_I	0.1069	0.1147	0.1169	0.3084	0.2418

Algorithm 1: Plane Curve Description.

Data: Set of organized point clouds $\overline{PC}$
Result: $F[f][n+6]$ - 2D array (descriptor)
$PC_k = \{X_1, ..X_i.., X_m, X_m.., X_i, ..X_1\}$;
Transform PC_k to PC_k^{xy} indicating the x, y plane in Cartesian coordinate system.;
for $k = 1 : f$ **do**
  for $j = 1 : n$ **do**
    $F[k][j] = FD[j]$;
  end
  $F[k][j+1] = \mu_{zk}$;
  $F[k][j+2] = \sigma_{zk}$;
  $F[k][j+3] = \mu_{rk}$;
  $F[k][j+4] = \sigma_{rk}$;
  $F[k][j+5] = \mu_{Ik}$;
  $F[k][j+6] = \sigma_{Ik}$;
end

4) Pseudo-Code for Plane Curve Description: Using the descriptor elements introduced above, we provide a pseudo-code for our new approach for plane curve description in Algorithm 1. The input data is an object either represented as a set of time varying plane curves (plane curve tracked through f frames): $\overline{PC} = \{PC_1, .., PC_k, .., PC_f\}$, where the timestamps $t_f > t_{f-1}$ and $PC_f = \{X_1, .. X_i.., X_m\}$, where $X_i = \{\phi_i, r_i, z_i\}$ (LIDAR plane points in polar coordinates) $\phi_i > \phi_{i-1}$, or a set of spatially neighboring plane curves (initial estimation): $\overline{PC} = \{PC_1, .., PC_k, .., PC_h\}$ where the mean altitudes $\mu_{zh} > \mu_{zh-1}$.

C. Classification of Planar Segments

For the object classification based on the introduced descriptor, we use Convolutional Neural Network as a classifier.

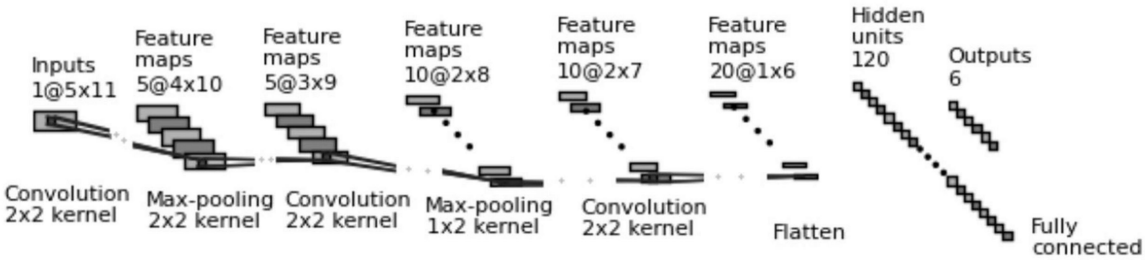


Fig. 4. Network architecture: all convolutional layers are followed by ReLUs and the fully-connected layer is followed by a softmaxlayer not illustrated in the scheme. There were 0 padding, both stride and dilation factor 1 (with both directions) are applied (in case of all layers).

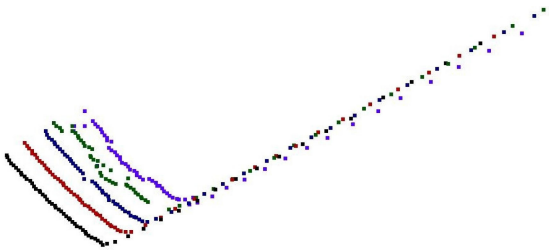


Fig. 5. Example of a vehicle segment from 5 consecutive frames ($f = 5$). Segments of a car (Purple: Frame l, Green: Frame $l-1$, Blue: Frame $l-2$, Red: Frame $l-3$, Black: Frame $l-4$).

CNNs are frequently used for different purposes (detection, segmentation, classification, etc.) in image processing, a survey about them can be found in [30]. CNN was selected for classification because we found its prediction probabilities to be superior to other classifiers we tried (e.g. Naive Bayes Classification or Support Vector Machine [31]). CNN is commonly applied in various areas ([32], [33]) where a convenient description can be made by a 2D matrix like in our case. 2D feature map structure is used because time (or space) are assumed to be locally related features. The typical structure of these types of networks are based on consecutive convolution, ReLU (rectified linear unit) and max-pooling layers. These elements form the basis of our architecture as well, where the layer sizes were adjusted to our descriptor size. Our network architecture is shown in Fig. 4. The input of our CNN is a 5×11 matrix (Fig. 5, Table I as in our experiments $f = 5$), while the output has the dimension equal to the number of categories, where each of the six values indicates the probability of the corresponding category. Parameters of the network training: Environment: Matlab; Learning rate: 0.01; Max epochs: 100 (Iterations/epoch: 1078); Solver: Stochastic Gradient Descent with Momentum; Validation frequency: 300 iterations; Hardware resource: Single GPU (Nvidia GTX 1080). This CNN is the common framework for both feature and object based recognition.

D. Maximum Likelihood for Object Level Decision

In the case when an object was built up from or tracked through multiple planar curves we applied maximum likelihood method. The result of our CNN for a single planar curve is a number between 0 and 1 in case of each category (summing up these number for all categories we get 1, because of the definition of softmaxlayer). We consider the resulting numbers as the probability of a curve of belonging to a category. We estimate the probability of the whole object to belonging to a category by:

$$P(\overline{X}, \lambda_j) = \prod_{i=1}^{N} P(X_i, \lambda_j), \tag{4}$$

where: $P(X_i, \lambda_j)$ is the probability of the ith planar curve of an object of corresponding to the jth category and N is the number of curves building up the object.

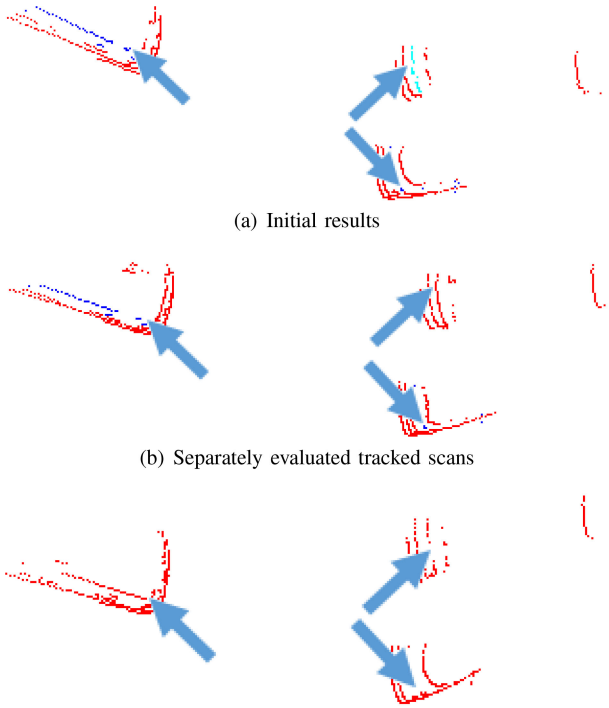
We predict the category for the whole object by selecting the one with the maximum probability.

E. Object Classification Pipeline

In this subsection, we will show some results on LIDAR sequences (raw data without annotation). We used the pipeline described in III-A to illustrate how the proposed system should work on a continuous data stream and recognize moving objects.

After finding moving objects (e.g. pedestrians, cyclists and cars) and tracking them for five frames, we evaluated our classification method on feature and object level as well. We have chosen one frame (Fig. 6) to illustrate the classification result both with and without using the proposed maximum likelihood scheme (also, an initial result - based on segments without tracking - is shown). In this figure one can see that some parts of cars are categorized cyclists or truck. In case of tracking of a planar curve only the reasonable fault remained (one curve of a car classified as truck). By using the maximum likelihood method and making decision on object level, the prediction for the whole cars will be correct.

The measured average running times of the prototype pipeline on the configuration of Intel Core i7-4790K @ 4.00GHz processor, 32 GB RAM, Nvidia GTX 1080 GPU, Windows 10 64 bit operating system are the following: preprocessing - 92 ms, the proposed algorithm (description - 10 ms, CNN - 4 ms, object level decision - 2 ms) - 16 ms. The process cycle time through the whole pipeline is about 108 ms. The preprocessing steps can be speed by GPU implementation [34]. Furthermore, the reported time corresponds to the whole scene, containing all object candidates, this time can be significantly decreased by using Region of Interests. The proposed classification scheme based on planar segments is fast enough to support intelligent vehicles.



(a) Initial results

(b) Separately evaluated tracked scans

(c) Object level decision with maximum likelihood from tracked scans

Fig. 6. Examples of the proposed pipeline, ground was segmented (Colormap: Red - Car and Van, Blue - Truck, Cyan - Cyclist).

IV. Sensor Resolution and Datasets

In the following, we will show results on two datasets. The first one is the KITTI dataset [35], which is commonly used to test algorithms for autonomous driving; the second one is from a recent work [7]. The tests contain near, medium distance and far objects (individual evaluation for far objects is the topic of the VI-C subsection). The number of plane segments for an object (in the following denoted as m) varies from 1 to cc. 20 (latter indicates a very close object). The first part of the tests (Section V) will show results of our initial evaluation (without tracking) and later (Section VI) we present the results on tracked frames. The former one is important in order to be able to make decisions at the beginning. It is also comparable to current state of the art methods, while results based on tracking are a new contribution of the paper.

In the following Tables, we will use the forthcoming notation to distinguish the different cases in the table captions:

Tracking: T - evaluation based on tracked segments; I - initial result (without tracking the plane segment).

Maximum likelihood: S - plane curves or plane tracked curves are separately evaluated (without maximum likelihood); M - maximum likelihood is used on the result of separately evaluated curves or tracked curves; G - grouped evaluation of plane curve of the same object with CNN.

Number of frames used in maximum likelihood decision: 0: maximum likelihood in the evaluation is not used; 1-X: the given number of frames are used; A: all the available frames are used.

Database: K - KITTI dataset; B - Budapest dataset [7].

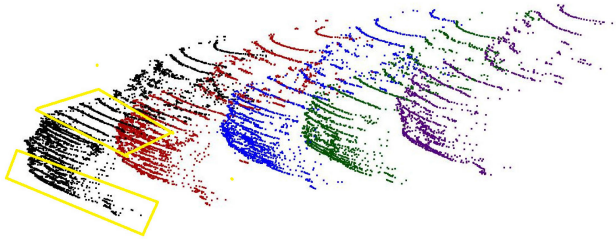


Fig. 7. A car represented in 5 frames (Purple: Frame l, Green: Frame $l-1$, Blue: Frame $l-2$, Red: Frame $l-3$, Black: Frame $l-4$); Local curve groups indicated with yellow quadrants.

Object distance: C - Evaluating all the annotated object segments (complete set, including near/medium/far objects); F - Evaluation only for far objects.

The different cases are explained by using Fig. 7.

- In case of initial estimation, previous occurrences of the object are not used, so the classification is based only on the curves of the actual frame (indicated with same color in Fig. 7). CNN with single curve at the input is used for classification. If maximum likelihood is applied in this initial case, we use all the classified curves of the current frame (indicated with same color) to gain an object level decision (e.g. IM1KF-Table XIV). Without maximum likelihood we acquire an independent classification result in feature level for each curve (e.g. IS0KC-Table III).
- Another case of initial estimation (evaluation at the first appearance of the object without the possibility of tracking), is the case when curve groups are classified simultaneously with our CNN, where maximum 5 curves can be fed into the input layer (e.g. IG0KF-Table XV). Two curve groups of 5 are indicated with yellow quadrants in Fig. 7. This method proved to be more efficient than maximum likelihood in case of far objects, where only a few layers are present.
- In case of evaluation of tracked curves, we categorize the plane curves grouped with their matched representation on previous frames (as shown in Fig. 5). After the classification, we use maximum likelihood on different number of curves from different number of frames (explained in details in Section VI). For example, in case of Fig. 7 we can use all the $5 * m$ curves (e.g. TM5BC-Table XIII) of each 5 occurrence of the car (assuming we have tracked the object at least from the frame l-9). Our test results will show that five frames contain sufficient information for the nearly best performance.

In the evaluations we use the average F-measure: $\overline{F} = \frac{1}{k}\sum_{i=1}^{k} F_i$, with the usual definition of single category F-rate:

$$F = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \tag{5}$$

where $\text{precision} = \frac{\text{true positive samples}}{\text{positive samples}}$ and $\text{recall} = \frac{\text{true positive samples}}{\text{relevant samples}}$. F-measure weighted by the sample numbers of each category is denoted by $\overline{F_w}$:

$$\overline{F_w} = \sum_{i=1}^{k} F_i \cdot \frac{n_i}{N} \tag{6}$$

Where F_i is the F-measure of ith category, k is the number of categories, n_i is the cardinality of ith category, and N is number of all the samples.

A. Sensor Resolution

The detectable far field objects and their definition as a far field object highly depends on the actual LIDAR sensor. As an example, we examine the scanning efficiency for Velodyne HDL-64E sensor planes and the detection range for pedestrians. Addressing the scanning resolution issue, we assume a human with 60 cm width and 160 cm height. Based on the sensor's vertical resolution, which is about 0.4 degrees, in an optimal situation it is possible to detect the human in two scan planes even about in 110 m range (the official sensor range is 120 m). However, calculating with its angular resolution of around 0.2 degree (it can vary from 0.08 to 0.35 degrees depending on its scanning frequency), in order to find at minimum 5 points on a scan plane (in our experiments we used minimum 5 points in a planar scan to represent an object), the human cannot be farther than 34 m. Although, lowering the scanning frame-rate enables 0.08 degree angular resolution, and one can find other LIDARs in the market with even higher angular resolution (SICK LD-LRS[6] -0.0625 degree, Quanergy M8[7] -0.03 degree). Consequently, it is possible to detect pedestrians even at a large distance. These values have to be checked for each sensor.

B. KITTI Dataset

We conducted proof of concept tests on the training set of the KITTI tracking database [35]; this set is annotated, so here it is guaranteed that we can have information about an object through several frames. In this set, labeled objects are annotated through different number of frames in 21 sequences. It allows us to investigate our classification algorithm independently of the quality of the preprocessing, both for our initial estimation and for tracked cases. Object tracking is one of the preprocessing steps in the latter case, its performance could affect the classification accuracy (best reported Multiple Object Tracking Accuracy - MOTA in this database is about 90% for cars). However, we need only a few frames (5 in these tests) for evaluation, so we assume that the classification is not influenced significantly by the tracking performance. We do not use tracking for the initial evaluation (showing already good performance). In these tests we gathered all the not occluded and not truncated objects from 8 categories (car, van, truck, pedestrian, person sitting, cyclist, tram, miscellaneous) which can be tracked through at least 5 frames and have at least 1 plane in each of these frames with minimum 5 points. These objects were cut out based on their annotated 3D bounding boxes and then they were divided into segments based on the scanner planes. This resulted in 197,256 samples, which we divided into training (70%), validation (15%) and test (15%) sets. It is worth mentioning that segmentation algorithms like [36] work with about 90% F-rate, so they can influence the final classification results. The categories of *car* and *van* and also *pedestrian* and *person sitting* are combined, because they are 'neighboring' categories.

[6]https://www.sick.com/ag/en/detection-and-ranging-solutions/2d-lidar-sensors/ld-lrs/c/g91912

[7]http://quanergy.com/m8/

C. Budapest Dataset

This dataset is presented in [7]. The authors segmented the objects as in [36]. Intensity data is not provided, so it was left out from our descriptor. There are four object categories in this urban data, namely: vehicle, street furniture, pedestrian and facade.

V. Comparison in Case of Spatial Series (Initial Estimation From One Frame)

In this section, we compare the proposed approach to state of the art classification methods. In order to accomplish adequate comparison (which is based on the same amount of information) we used here our initial estimation from one frame without tracking. To the best of our knowledge, there are no other methods applying tracked frames as part of their descriptor, so comparison is presented only in this section. When an object first appears in the far field, we do not have information about it from previous frames. Until we gather information about how its shape varies, we propose to make a decision as follows. Fill the different rows of the input descriptor matrix of the CNN with locally neighboring planes of the same object (instead of temporal neighbors), and if we have less than five, use the actual one more than once. The issue of instant decision is very important, because acquiring five frames can be time-consuming (depending on the frame-rate of the LIDAR sensor). Even with 20 Hz scanning frequency, a vehicle moving with 90 km/h on a highway can move about 6 m. To summarize, it is essential to make a fast decision as our initial estimation, even if it has a lower precision.

A. 2D LIDAR

In [4] and [10] a method is proposed for pedestrians, vehicles, 2 wheel vehicles and rubber cones classification in 2D LIDAR scans. In [4] 100%, 88.38%, 99.38 and 100% accuracy was reported for the above categories, but only for a few thousand test samples. We tested the method on the KITTI database. In this test a nearest neighbor classification was performed based on Euclidean distance of width, range variance and intensity data of the curves as [4] proposed. Their results can be seen in Table II which can be compared to the results of our initial estimation in Table III on the same database. The difference between the measures on KITTI and reported in [4] can be explained by different category numbers, types, cardinality and type of the data. In the test of Table III we gained information for our method from only a single planar curve without tracking it, in this case all the five rows of our descriptor matrix were filled with the same data. Our method outperforms [4] in each aspect except the recall of tram and miscellaneous category, but the overall performance difference is significant. The comparison is shown in Table IV.

TABLE II
CONFUSION MATRIX FOR PLANAR CURVES BY THE METHOD PROPOSED IN, DATABASE2D, [10] *in* KITTI DATASET. (IS0KC) (1: CAR AND VAN, 2: TRUCK, 3: PEDESTRIAN AND PERSON SITTING, 4: CYCLISTS, 5: TRAM, 6: MISC)

	1	2	3	4	5	6	Precision (%)
1	**73682**	2298	2274	986	162	1705	**90.9**
2	2315	**5757**	36	39	59	320	**67.5**
3	2471	39	**79742**	8473	0	679	**87.2**
4	944	30	8328	**2827**	1	215	**22.9**
5	164	50	0	2	**65**	8	**22.6**
6	1691	321	618	211	7	**737**	**20.6**
Recall (%)	**90.7**	**67.8**	**87.6**	**22.6**	**22.1**	**20.1**	$\overline{F}$:**0.519** $\overline{F_w}$:**0.825**

TABLE III
CONFUSION MATRIX FOR PLANAR CURVES (PART OF INITIAL ESTIMATION) BY USING THE PROPOSED METHOD IN KITTI DATASET. (IS0KC) (1: CAR AND VAN, 2: TRUCK, 3: PEDESTRIAN AND PERSON SITTING, 4: CYCLISTS, 5: TRAM, 6: MISC)

	1	2	3	4	5	6	Precision (%)
1	**78914**	2346	1577	1200	148	2235	**91.3**
2	1128	**6069**	1	27	100	400	**78.6**
3	744	28	**82366**	5277	0	516	**92.6**
4	369	30	6993	**5998**	0	321	**43.8**
5	1	0	0	0	**46**	0	**97.9**
6	111	22	61	36	0	**190**	**45.2**
Recall (%)	**97.1**	**71.4**	**90.5**	**47.8**	**15.7**	**5.2**	$\overline{F}$:**0.571** $\overline{F_w}$:**0.874**

B. 3D LIDAR

The authors of [7] present the results for the Budapest database presented in Table V. In [7] the object classification is based on one frame without tracking. In order to compare our methods, we built up the descriptor of object slices from 5 neighboring scan planes of an object in each frame (instead of 5 different occurrences of the same plane segment in consecutive frames). After separate evaluations of curve groups, we used the proposed maximum likelihood method on the five-curve-groups to make a concatenated decision on the object level. Results of [7] for this database: Precision: 90%, Recall: 87%, F-rate: 89%, detailed evaluation can be found in [7]. Our precision and recall values are higher (Table VI) in case of each category except the vehicle (Table V). This can be explained by the fact the authors used contextual labeling refinement after classification to distinguish the similar categories of vehicle and facade. By applying tracking and object level evaluation as we propose (Table XIII) we outperform [7] in all the measures. Note that method of [7] cannot be used in case of small number of scan planes.

TABLE IV
METHOD COMPARISON (PROPOSED AND [4], [10]) IN CASE OF 2D CONTOURS (IS0KC)

Categories	Precision (%)		Recall (%)		F-rate	
	proposed	[4], [10]	proposed	[4], [10]	proposed	[4], [10]
Car and Van	**91**	91	**97**	91	**0.94**	0.91
Truck	**79**	68	**71**	68	**0.75**	0.68
Pedestrian and Person sitting	**93**	87	**91**	88	**0.92**	0.88
Cyclists	**44**	23	**48**	23	**0.46**	0.23
Tram	**98**	23	16	**22**	**0.28**	0.23
Misc	**45**	21	5	**20**	0.09	**0.21**
Average	**75**	52	**55**	52	**0.57**	0.52

TABLE V
METHOD COMPARISON (PROPOSED AND [7]) IN CASE OF 2.5D OBJECTS WITHOUT TRACKING (IM1BC)

Categories	Precision (%)		Recall (%)		F-rate	
	proposed	[7]	proposed	[7]	proposed	[7]
Vehicle	**99**	98	97	**99**	0.98	**0.99**
Street Furniture	**99**	92	90	**97**	**0.94**	0.94
Pedestrian	**89**	78	**100**	78	**0.94**	0.78
Facade	**95**	93	**89**	77	**0.92**	0.84
Average	**96**	90	**94**	87	**0.94**	0.89

TABLE VI
CONFUSION MATRIX FOR PLANAR CURVES BY USING PROPOSED METHOD WITH MAXIMUM LIKELIHOOD SCHEME FOR m CURVES OF ONE FRAME IN BUDAPEST D (IM1BC). (1: VEHICLE, 2: STREET FURNITURE, 3: PEDESTRIAN, 4: FACADE)

	1	2	3	4	Precision (%)
1	**4540**	0	0	47	**99.0**
2	31	**4020**	0	0	**99.2**
3	0	433	**4593**	109	**89.4**
4	105	24	9	**2809**	**95.3**
Recall (%)	**97.1**	**89.8**	**99.8**	**89.4**	$\overline{F}$:**0.943** $\overline{F_w}$:**0.955**

VI. CLASSIFICATION RESULTS OF TIME SERIES (FROM INITIAL TO TRACKING BASED)

Four type of tests are evaluated in case of both datasets:

- Classification of each planar segment as part of different object as data acquired from a single-layer LIDAR (depending on previous occurrence of this segment on previous frames, but independent from another segments of the same object).

TABLE VII
CONFUSION MATRIX FOR TRACKED PLANAR CURVES WITH THE PROPOSED METHOD IN KITTI DATASET. (TS0KC) (1: CAR AND VAN, 2: TRUCK, 3: PEDESTRIAN AND PERSON SITTING, 4: CYCLISTS, 5: TRAM, 6: MISC)

	1	2	3	4	5	6	Precision (%)
1	**79969**	1851	544	781	66	1664	**94.2**
2	394	**6429**	1	2	13	89	**92.8**
3	300	21	**87382**	4564	0	766	**93.9**
4	270	15	2915	**7111**	0	84	**68.4**
5	13	49	0	0	**215**	0	**77.6**
6	321	130	156	80	0	**1061**	**60.7**
Recall (%)	**98.4**	**75.7**	**96.0**	**56.7**	**73.1**	**29.0**	$\overline{F}$:**0.752** $\overline{F_w}$:**0.918**

TABLE VIII
CONFUSION MATRIX FOR TRACKED PLANAR CURVES BY USING THE PROPOSED METHOD WITH MAXIMUM LIKELIHOOD SCHEME FOR m TRACKED CURVES OF ONE FRAME IN KITTI DATASET. (TM1KC) (1: CAR AND VAN, 2: TRUCK, 3: PEDESTRIAN AND PERSON SITTING, 4: CYCLISTS, 5: TRAM, 6: MISC)

	1	2	3	4	5	6	Precision (%)
1	**81141**	338	14	11	8	1245	**98.1**
2	88	**8101**	0	0	0	0	**98.9**
3	2	0	**90529**	1846	0	846	**97.1**
4	8	0	402	**10674**	0	56	**95.8**
5	9	56	0	0	**286**	0	**81.5**
6	27	0	53	7	0	**1517**	**94.6**
Recall (%)	**99.8**	**95.4**	**99.5**	**85.2**	**97.3**	**41.4**	$\overline{F}$:**0.815** $\overline{F_w}$:**0.972**

TABLE IX
CONFUSION MATRIX FOR TRACKED PLANAR CURVES BY USING THE PROPOSED METHOD WITH MAXIMUM LIKELIHOOD SCHEME FOR $5 * m$ TRACKED CURVES OF FIVE FRAMES IN KITTI DATASET. (TM5KC) (1: CAR AND VAN, 2: TRUCK, 3: PEDESTRIAN AND PERSON SITTING, 4: CYCLISTS, 5: TRAM, 6: MISC)

	1	2	3	4	5	6	Precision (%)
1	**81167**	385	0	0	0	1423	**97.8**
2	96	**8110**	0	0	0	0	**98.8**
3	0	0	**90650**	1974	0	874	**97.0**
4	0	0	329	**10564**	0	45	**96.6**
5	0	0	0	0	**294**	0	**100.0**
6	4	0	19	0	0	**1322**	**98.3**
Recall (%)	**99.9**	**99.6**	**99.7**	**84.3**	**100.0**	**36.1**	$\overline{F}$:**0.899** $\overline{F_w}$:**0.972**

- Based on the result of the above classification, decision is made on object level. 3D objects are segmented based on the sensor rings. If an object is built up of more than one segment ($m > 1$) in the actual frame, we use the proposed maximum likelihood scheme on the CNN output to make an object level decision.
- Based on the result of independent classification of each planar segment, decision is made on object level based on the last $f = 5$ frames. If the tracked object is built up of m segments in each of the last 5 frames, we use the proposed maximum likelihood scheme on the CNN output to make an object level decision based on $5 * m$ segments.
- Based on the result of the independent classification of each planar segment, decision is made on object level based on the last f frames. If the tracked object is built up of m segments in each of the last f frames, we use the proposed maximum likelihood scheme on the CNN output to make on object level decision based on $f * m$ segments.

A. KITTI Dataset

The accuracy ($\frac{\text{Number of correct predictions}}{\text{Total number of predictions}}$) of the CNN for all the objects is about 92% on the training, validation and test sets as well without the proposed maximum likelihood scheme. Confusion matrices for all the samples are shown in Tables VII, VIII, IX and X. In Table VII all the planar segments were evaluated independently from the other curves building up the same object. In Table VIII, IX and X the planar curves were evaluated at object level with the proposed maximum likelihood scheme based on the indicated number of frames.

The confusion matrix in Tables VII and VIII show that even one 2D contour in one frame can produce good initial results and Tables VIII-X show that the maximum likelihood scheme we propose is effective to increase accuracy. Conclusion for each category:

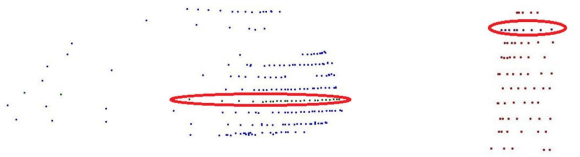
- The precision and recall values of car (and van) and pedestrian (and person sitting) categories are high.
- Trucks are frequently categorized as car or van, which is reasonable. The overall performance is sufficient.
- Cyclists are also miss-categorized frequently (as Pedestrian or Person Sitting). The performance measurements in case of this category is lower, but it has to be noted there were much less samples and maximum likelihood scheme raised these values significantly.
- The results on tram class are sufficient, however it should be noted it is not representative because of the very small number of samples.
- Finally, in case of misc category our proposed method did not performed well because of the variety of the diverse objects hard to be identified based on sometimes under-sampled 2D contours. Although as Table X shows, good precision can be achieved.

By examining the results for Table IX and X, one can conclude that there is small improvement by using more than 5 frames, so the assumption which stated enough using 5 tracked frames is proved to be right. Fig. 8 shows examples of categorized plane curves from two objects.

The results are convincingly sufficient considering that pedestrian detection methods are robust against up to 30% occlusion [37] on 2D images, and in a similar dataset [35] best detection results using both vision and LIDAR data [38] is about 90%.

TABLE X
CONFUSION MATRIX FOR TRACKED PLANAR CURVES BY USING THE PROPOSED METHOD WITH MAXIMUM LIKELIHOOD SCHEME FOR $f * m$ TRACKED CURVES OF ALL THE AVAILABLE FRAMES IN KITTI DATASET. (TMAKC) (1: CAR AND VAN, 2: TRUCK, 3: PEDESTRIAN AND PERSON SITTING, 4: CYCLISTS, 5: TRAM, 6: MISC)

	1	2	3	4	5	6	Precision (%)
1	**81164**	288	0	0	0	1842	**97.4**
2	103	**8207**	0	0	0	0	**98.8**
3	0	0	**90998**	342	0	841	**97.0**
4	0	0	0	**12196**	0	0	**100.0**
5	0	0	0	0	**294**	0	**100.0**
6	0	0	0	0	0	**981**	**100.0**
Recall (%)	**99.9**	**96.6**	**100.0**	**97.3**	**100.0**	**26.8**	$\overline{F}$:**0.893** $\overline{F_w}$:**0.975**



(a) Car (b) Pedestrian

Fig. 8. Examples of the KITTI database: Points of miss-categorized plane curves are circled with red. Note that: for illustration purposes we have chosen the above objects with a dense series of scanner-plane segments, however the method was designed primarily for objects with only a few of those.

TABLE XI
CONFUSION MATRIX FOR TRACKED PLANAR CURVES WITH THE PROPOSED METHOD IN BUDAPEST DATASET. (TS0BC) (1: VEHICLE, 2: STREET FURNITURE, 3: PEDESTRIAN, 4: FACADE)

	1	2	3	4	Precision (%)
1	**4484**	95	4	100	**95.8**
2	96	**4344**	4	1	**97.7**
3	9	36	**4573**	5	**98.9**
4	87	2	21	**2859**	**96.3**
Recall (%)	**95.9**	**97.0**	**99.4**	**96.4**	$\overline{F}$:**0.972** $\overline{F}$:**0.973**

B. Budapest Dataset

Our quantitative results for this dataset are in Tables XI, XII and XIII. Table XI shows all the planar curves evaluated independently from the other curves building up the same object, while in Tables XII and XIII the planar segments were evaluated at object level with the proposed maximum likelihood scheme based on the indicated number of frames. The results have high precision and recall values for each category and they are increasing as we increase the number of scan planes used in the decision process. If we make a decision using 5 frames even 1.0 F-measure can be achieved. Vehicle and facade also vehicle and street furniture categories are mixed up in the beginning, however this miss-categorization is not significant at all.

TABLE XII
CONFUSION MATRIX FOR TRACKED PLANAR CURVES BY USING THE PROPOSED METHOD WITH MAXIMUM LIKELIHOOD SCHEME FOR m TRACKED CURVES OF ONE FRAME IN BUDAPEST DATASET. (TM1BC) (1: VEHICLE, 2: STREET FURNITURE, 3: PEDESTRIAN, 4: FACADE)

	1	2	3	4	Precision (%)
1	**4620**	0	0	0	**100.0**
2	0	**4477**	0	0	**100.0**
3	0	0	**4602**	0	**100.0**
4	56	0	0	**2965**	**98.2**
Recall (%)	**98.8**	**100.0**	**100.0**	**100.0**	$\overline{F}$:**0.996** $\overline{F}$:**0.997**

TABLE XIII
CONFUSION MATRIX FOR TRACKED PLANAR CURVES BY USING THE PROPOSED METHOD WITH MAXIMUM LIKELIHOOD SCHEME FOR $5 * m$ TRACKED CURVES OF FIVE FRAMES IN BUDAPEST DATASET. (TM5BC) (1: VEHICLE, 2: STREET FURNITURE, 3: PEDESTRIAN, 4: FACADE)

	1	2	3	4	Precision (%)
1	**4676**	0	0	0	**100.0**
2	0	**4477**	0	0	**100.0**
3	0	0	**4602**	0	**100.0**
4	0	0	0	**2965**	**100.0**
Recall (%)	**100.0**	**100.0**	**100.0**	**100.0**	$\overline{F}$:**1.000** $\overline{F_w}$:**1.000**

C. Far Field Objects

Up to now, we have shown that our method is superior in near/medium distance objects' cases. Now, we demonstrate that in case of far objects. Here, the definition of far objects is the following: its distance from the sensor is several ten meters, resulting in a few scan planes (m) and a few points per planes (minimum 5), relatively far from each other, the point cloud is vertically undersampled.

In the Budapest dataset, far field objects are not present, but in the KITTI dataset objects far from the LIDAR sensor are annotated too, so we can manage to show recognition results on this dataset. Fig. 11 shows examples of car and pedestrian located relatively far from the LIDAR. Tables XIV and XVI present evaluation on objects built up from maximum four scan planes. In this case the average distance of the center of gravity from the sensor is about 41 m. In these Tables Car, Van and Truck categories are merged, because there were very small number of truck objects in this evaluation, and they are similar to the other category. Examining Table XVI (tracked curves and object level evaluation for far objects) we can state that the results are similar to Table VIII (tracked curves and object level evaluation) in sense of precision, recall and F-measure, the range does not influence significantly the recognition. Figs. 9 and 10 shows

TABLE XIV
CONFUSION MATRIX FOR PLANAR CURVES (INITIAL ESTIMATION) (FOR FAR OBJECTS) BY USING THE PROPOSED METHOD WITH MAXIMUM LIKELIHOOD SCHEME FOR m CURVES OF ONE FRAME IN KITTI DATASET (IM1KF) (1: CAR, VAN AND TRUCK, 2: PEDESTRIAN AND PERSON SITTING, 3: CYCLISTS, 4: TRAM, 5: MISC)

	1	2	3	4	5	Precision (%)
1	**5197**	2	19	35	49	**98.0**
2	4	**898**	102	0	0	**89.4**
3	0	0	**68**	0	0	**100.0**
4	0	0	0	**0**	0	**0.0**
5	0	4	4	0	**8**	**50.0**
Recall (%)	**99.9**	**99.3**	**35.2**	**0.0**	**14.0**	$\overline{F}$:**0.534** $\overline{F_w}$:**0.956**

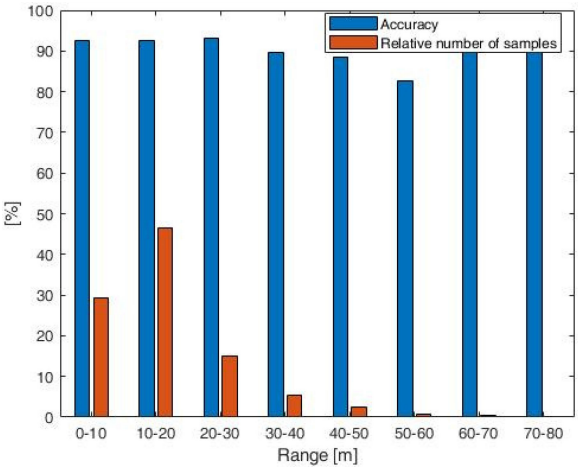


Fig. 9. Overall accuracy and relative number of samples with respect to the plane segment's (TM1KC) average distance to the sensor. Sample numbers of different range bins vary between a few tens and tens of thousands.

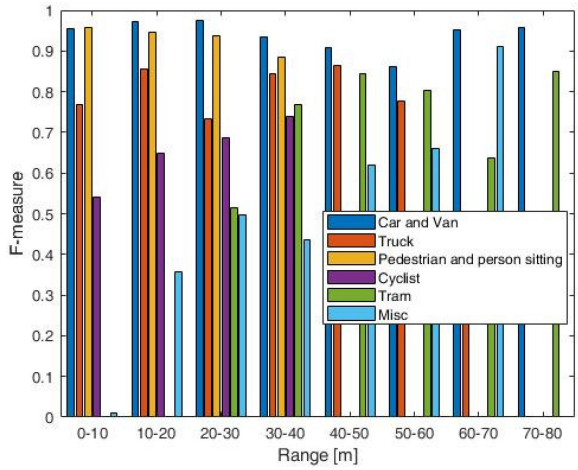


Fig. 10. F-measure values of each category with respect to the plane segment's (TM1KC) average distance to the sensor. Note that: If a bar is missing in a range that indicates that in the given range interval there were no object sample of the given category.

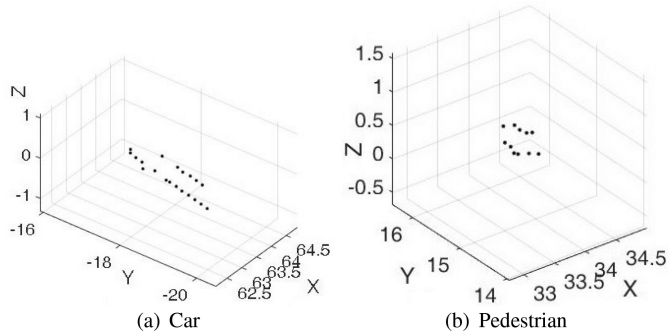


Fig. 11. Examples of far objects (the car is about 66 m and the pedestrian is about 37 m far from the LIDAR).

TABLE XV
CONFUSION MATRIX FOR PLANAR CURVES (INITIAL ESTIMATION) (FOR FAR OBJECTS) BY USING THE PROPOSED METHOD FOR SIMULTANEOUS EVALUATION (CNN) OF 5 CURVES OF ONE FRAME IN KITTI DATASET. (IG0KF) (1: CAR, VAN AND TRUCK, 2: PEDESTRIAN AND PERSON SITTING, 3: CYCLISTS, 4: TRAM, 5: MISC)

	1	2	3	4	5	Precision (%)
1	**5192**	2	18	35	36	**98.3**
2	4	**902**	70	0	1	**92.3**
3	0	0	**102**	0	1	**99.0**
4	0	0	0	**0**	0	**0.0**
5	5	0	3	0	**19**	**70.4**
Recall (%)	**99.8**	**99.8**	**52.9**	**0.0**	**33.3**	$\overline{F}$:**0.615** $\overline{F_w}$:**0.964**

overall accuracy and F-measure of categories varying through range.

Tables XIV (object level evaluation with maximum likelihood for our initial estimation in case of far objects) and XV (object level evaluation with CNN for our initial estimation in case of far objects) shows us teaching and evaluating the curves together improves the classification.

VII. CONCLUSION

In the paper, a 2D-based 3D recognition method is proposed which utilizes time varying shape information and 3D information if it is available. We designed this method in order to solve the recognition problem of far objects from dense LIDAR point clouds or the general recognition problem for few layer LIDARs. Our method can be used for outdoor objects being invariant of the sensor. We have proposed a 3D training method based on 2D planar curves, where in the same framework we can process feature based and object based recognition, making a robust system for initial guess, object based and object tracking based evaluation. Our proposed method is novel and it is independent of object models. We demonstrated that it is capable to categorize noisy 2D point clouds in a large public database. However, we compared it to a recent method used for object detection in 2D

TABLE XVI
CONFUSION MATRIX FOR TRACKED PLANAR CURVES (FOR FAR OBJECTS) BY USING THE PROPOSED METHOD WITH MAXIMUM LIKELIHOOD SCHEME FOR m TRACKED CURVES OF ONE FRAME IN KITTI DATASET. (TM1KF) (1: CAR, VAN AND TRUCK, 2: PEDESTRIAN AND PERSON SITTING, 3: CYCLISTS, 4: TRAM, 5: MISC)

	1	2	3	4	5	Precision (%)
1	**5184**	0	5	8	31	**99.1**
2	2	**849**	25	0	0	**96.9**
3	0	28	**156**	0	0	**84.8**
4	2	0	0	**27**	0	**93.1**
5	13	27	7	0	**26**	**35.6**
Recall (%)	**99.7**	**93.9**	**80.8**	**77.1**	**45.6**	$\overline{F}$:**0.804** $\overline{F_w}$:**0.977**

LIDAR point clouds and to another method uses 3D recognition in one frame; in both cases our method is proved to be superior.

- In case of 2D contours, the proposed method ($\overline{F}$:**0.571**, $\overline{F_w}$:**0.874** - Table III) outperformed the state of the art ($\overline{F}$:0.519, $\overline{F_w}$:0.825 - Table II) [4], [10].
- For 3D objects, our separate evaluation of planar curves with maximum likelihood aggregation ($\overline{F}$:**0.943** - Table VI) proved to be more efficient than the method in [7] ($\overline{F}$:0.890 - Table V).
- We proved that using grouped evaluation of planar curves for far objects (only have a few segments) by CNN ($\overline{F}$:**0.615**, $\overline{F_w}$:**0.964** - Table XV) is even better than combining separate evaluation and maximum likelihood method ($\overline{F}$:0.534, $\overline{F_w}$:0.956 - Table XIV), and our method perform just as well in case of far objects ($\overline{F}$:**0.804**, $\overline{F_w}$:**0.977** - Table XVI) as in case of near ones ($\overline{F}$:**0.815**, $\overline{F_w}$:**0.972** - Table VIII).
- One can achieve the best result evaluating tracked curves with our CNN and using maximum likelihood to aggregate the results from the segments on object level. As we increase the number of frames used for maximum likelihood (until cc. 5) the performance increases as well ($\overline{F}$:**0.899**, $\overline{F_w}$:**0.972** - Table IX) for KITTI database and ($\overline{F}$:**1.0**, $\overline{F_w}$:**1.0** - Table XIII) for Budapest database).

We recommend to use our method as an extension to 3D recognition methods ([6], [7]) in environments they cannot process (far field); it expands the detection range. We proposed a recognition system where in the same framework and training system we start with good initial guessing, then tracking the object with continuously increasing efficiency, as the visibility and the number of scanning slices are also increasing. The paper proves the rationality of representation of LIDAR objects as a set of time-varying plane curves. We would later examine how other frameworks can benefit from this representation e.g. Recurrent neural network, or Long short-term memory (LSTM) networks in case of continuous parallel traffic of fare vehicles.

REFERENCES

[1] C. Hubmann, J. Schulz, M. Becker, D. Althoff, and C. Stiller, "Automated driving in uncertain environments: Planning with interaction and uncertain maneuver prediction," *IEEE Trans. Intell. Veh.*, vol. 3, no. 1, pp. 5–17, Mar. 2018.

[2] B. Okumura *et al.*, "Challenges in perception and decision making for intelligent automotive vehicles: A case study," *IEEE Trans. Intell. Veh.*, vol. 1, no. 1, pp. 20–32, Mar. 2016.

[3] Z. Rozsa and T. Sziranyi, "Obstacle prediction for automated guided vehicles based on point clouds measured by a tilted LIDAR sensor," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 8, pp. 2708–2720, Aug. 2018.

[4] M. Lee, S. Hur, and Y. Park, "Obstacle classification method based on 2D LIDAR database," *Pattern Recognit. Lett.*, vol. 8, no. 8, pp. 1442–1446, 2014.

[5] L. Kurnianggoro and K. H. Jo, "Object classification for LIDAR data using encoded features," in *Proc. 10th Int. Conf. Human Syst. Interact.*, 2017, pp. 49–53.

[6] D. Maturana and S. Scherer, "VoxNet: A 3D convolutional neural network for real-time object recognition," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2015, pp. 922–928.

[7] A. Borcs, B. Nagy, and C. Benedek, "Instant object detection in LIDAR point clouds," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 7, pp. 992–996, Jul. 2017.

[8] K. O. Arras, O. M. Mozos, and W. Burgard, "Using boosted features for the detection of people in 2D range data," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2007, pp. 3402–3407.

[9] A. Fod, A. Howard, and M. A. J. Mataric, "A laser-based people tracker," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2002, vol. 3, pp. 3024–3029.

[10] M. Lee, S. Hur, and Y. Park, "An obstacle classification method using multi-feature comparison based on 2D lidar database," in *Proc. 12th Int. Conf. Inform. Technol. New Gener.*, 2015, pp. 674–679.

[11] F. Galip, M. H. Sharif, M. Caputcu, and S. Uyaver, "Recognition of objects from laser scanned data points using SVM," in *Proc. 1st Int. Conf. Multimedia Image Process.*, 2016, pp. 28–35.

[12] C. Weinrich, T. Wengefeld, M. Volkhardt, A. Scheidig, and H.-M. Gross, *Generic Distance-Invariant Features for Detecting People with Walking Aid in 2D Laser Range Data*. Cham, Switzerland: Springer, 2016, pp. 735–747.

[13] A. Carballo, A. Ohya, and S. Yuta, "People detection using range and intensity data from multi-layered laser range finders," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2010, pp. 5849–5854.

[14] O. M. Mozos, R. Kurazume, and T. Hasegawa, "Multi-part people detection using 2D range data," *Int. J. Social Robot.*, vol. 2, no. 1, pp. 31–40, Mar. 2010.

[15] L. Spinello, K. O. Arras, R. Triebel, and R. Siegwart, "A layered approach to people detection in 3D range data," in *Proc. 24th AAAI Conf. Artif. Intell.*, 2010, pp. 1625–1630.

[16] Z. Yücel, T. Ikeda, T. Miyashita, and N. Hagita, "Identification of mobile entities based on trajectory and shape information," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2011, pp. 3589–3594.

[17] A. Vatavu, R. Danescu, and S. Nedevschi, "Stereovision-based multiple object tracking in traffic scenarios using free-form obstacle delimiters and particle filters," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 1, pp. 498–511, Feb. 2015.

[18] L. Beyer, A. Hermans, and B. Leibe, "Drow: Real-time deep learning-based wheelchair detection in 2-D range data," *IEEE Robot. Autom. Lett.*, vol. 2, no. 2, pp. 585–592, Apr. 2017.

[19] B. Qin *et al.*, *A Spatial-Temporal Approach for Moving Object Recognition With 2D LIDAR*. Cham, Switzerland: Springer, 2016, pp. 807–820.

[20] Z. Rozsa and T. Sziranyi, "Street object classification via lidars with only a single or a few layers," in *Proc. 3rd IEEE Int. Conf. Image Process., Appl. Syst.*, 2018, pp. 156–161.

[21] P. Besl and N. D. McKay, "A method for registration of 3-D shapes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 14, no. 2, pp. 239–256, Feb. 1992.

[22] P. Torr and A. Zisserman, "MLESAC: A new robust estimator with application to estimating image geometry," *Comput. Vis. Image Understanding*, vol. 78, no. 1, pp. 138–156, 2000.

[23] R. B. Rusu, "Semantic 3D object maps for everyday manipulation in human living environments," Ph.D. dissertation, Comput. Sci. Dept., Technische Universitaet Muenchen, Munich, Germany, Oct. 2009.

[24] A. Asvadi, C. Premebida, P. Peixoto, and U. Nunes, "3D Lidar-based static and moving obstacle detection in driving environments: An approach based on voxels and multi-region ground planes," *Robot. Auton. Syst.*, vol. 83, pp. 299–311, 2016.

[25] D. Lague, N. Brodu, and J. Leroux, "Accurate 3D comparison of complex topography with terrestrial laser scanner: Application to the Rangitikei canyon (n-z)," *ISPRS J. Photogrammetry Remote Sens.*, vol. 82, no. Supplement C, pp. 10–26, 2013.
[26] M. L. Miller, H. S. Stone, and I. J. Cox, "Optimizing Murty's ranked assignment method," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 33, no. 3, pp. 851–862, Jul. 1997.
[27] J. Cooley, P. Lewis, and P. Welch, "The finite Fourier transform," *IEEE Trans. Audio Electroacoust.*, vol. AE-17, no. 2, pp. 77–85, Jun. 1969.
[28] C. E. Shannon, "Communication in the presence of noise," *Proc. IRE*, vol. 37, no. 1, pp. 10–21, Jan. 1949.
[29] A. Licsar and T. Sziranyi, "User-adaptive hand gesture recognition system with interactive training," *Image Vision Comput.*, vol. 23, no. 12, pp. 1102–1114, 2005.
[30] J. Gu *et al.*, "Recent advances in convolutional neural networks," *Pattern Recognit.*, vol. 77, pp. 354–377, 2017.
[31] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, 2nd ed. New York, NY, USA: Springer, 2008.
[32] N. M. Rad and C. Furlanello, "Applying deep learning to stereotypical motor movement detection in autism spectrum disorders," in *Proc. IEEE 16th Int. Conf. Data Mining Workshops*, 2016, pp. 1235–1242.
[33] Y. Du, Y. Fu, and L. Wang, "Skeleton based action recognition with convolutional neural network," in *Proc. 3rd IAPR Asian Conf. Pattern Recognit.*, 2015, pp. 579–583.
[34] L. Christino and F. Osório, "GPU-services: GPU based real-time processing of 3D point clouds applied to robotic systems and intelligent vehicles," in *Robotics*, F. Santos Osório and R. Sales Gonçalves, Eds., Cham, Switzerland: Springer, 2016, pp. 152–171.
[35] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the KITTI vision benchmark suite," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2012, pp. 3354–3361.
[36] A. Börcs, B. Nagy, and C. Benedek, "Fast 3-D urban object detection on streaming point clouds," in *Computer Vision - ECCV 2014 Workshops*, L. Agapito, M. M. Bronstein, and C. Rother, Eds., Cham, Switzerland: Springer, 2015, pp. 628–639.
[37] D. Varga and T. Sziranyi, "Robust real-time pedestrian detection in surveillance videos," *J. Ambient Intell. Humanized Comput.*, vol. 8, no. 1, pp. 79–85, Feb. 2017.
[38] X. Chen, H. Ma, J. Wan, B. Li, and T. Xia, "Multi-view 3D object detection network for autonomous driving," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6526–6534.

Zoltan Rozsa (S'16–M'18) received the M.Sc. degree in mechanical engineering modeling from the Budapest University of Technology and Economics, Budapest, Hungary. He is currently working toward the Ph.D. degree at the Department of Material Handling and Logistic Systems of Budapest University of Technology and Economics and as Research Assistant with the Computer and Automation Research Institute, Hungarian Academy of Sciences, Budapest, Hungary. His research interests include construction logistics, material handling machines, machine vision, and 3D recognition.

Tamas Sziranyi (M'90–SM'04) received the Ph.D. and D.Sci. degrees in 1991 and 2001, from the Hungarian Academy of Sciences, Budapest, Hungary. He was appointed as a Full Professor in 2001 at Pannon University, Veszprem, Hungary, and, in 2004, at the Peter Pazmany Catholic University, Budapest, Hungary. He is currently a Full Professor with the Budapest University of Technology and Economics. He has been a Research Scientist with the Institute for Computer Science and Control (MTA SZTAKI), Hungarian Academy of Sciences, Budapest, Hungary, since 1992, where he leads the Machine Perception Research Laboratory since 2006. He has authored and coauthored more than 270 publications including 50 in major scientific journals, and several international patents. His research activities include machine perception, pattern recognition, texture and motion segmentation, Markov random fields and stochastic optimization, remote sensing, surveillance, intelligent networked sensor systems, graph-based clustering, and digital film restoration. With his research laboratory he has participated in several prestigious international (ESA, EDA, FP6, FP7, OTKA) projects. Dr. Sziranyi was the Founder and Past President (1997–2002) of the Hungarian Image Processing and Pattern Recognition Society. He was an Associate Editor of the IEEE TRANSACTIONS ON IMAGE PROCESSING (2003–2009), and is the Associate Editor of *Digital Signal Processing* (Elsevier) since 2012. He was the recipient of the Master Professor Award in 2001, the Szechenyi Professorship, and the ProScientia (Veszprem) Award in 2011 and by the Officers' Cross by the President of Hungary in 2018. He is a Fellow of both the International Association of Pattern Recognition and the Hungarian Academy of Engineering since 2008.