PointNet Evaluation for On-Road Object Detection Using a Multi-resolution Conditioning

Jose Pamplona[1(✉)], Carlos Madrigal[1], and Arturo de la Escalera[2]

[1] Artificial Vision and Photonics Lab, Instituto Tecnológico Metropolitano,
Calle 54 A #30-01, Medellín, Colombia
josepamplona212620@correo.itm.edu.co

[2] Intelligent Systems Lab, Universidad Carlos III de Madrid,
Avda. de la Universidad 30, 28911 Leganés, Spain
http://www.itm.edu.co, http://portal.uc3m.es

Abstract. On-road object detection is one of the main topics in the development of autonomous vehicles. Factors related to the diversity of classes, pose changes, occlusions, and low resolution make object detection challenging. Most of the object detection techniques which have been based on RGB images, have limitations because of the influence of environmental lighting conditions. Consequently, other sources of information have become interesting for undertaking this task. This paper proposes an on-road object detection method, which uses 3D information acquired by a LiDAR HD sensor. We evaluate a neural network architecture based on PointNet for multi-resolution 3D objects. To carry this out, a multi-resolution conditioning stage is proposed in order to optimize the performance of the PointNet architecture applied over LiDAR data. Both the training and evaluation processes are performed by using the KITTI dataset. Our approach uses low computational cost algorithms, which are based on occupancy grid maps for on-road object segmentation. The experiments show that the proposed method achieves better results than PointNet evaluated on a single resolution.

Keywords: Pedestrian detection · LiDAR · Deep learning · Resolution conditioning

1 Introduction

Traffic accidents are the first preventable death cause worldwide [8]. An emerging approach to solve this problem is autonomous driving [3]. In recent years, there has been an increasing interest in autonomous vehicles. Based on Scopus database, in 2017 more than 3,500 published papers show the interest of the researching community.

One of the greatest challenges on autonomous driving is object detection, which is needed to take driving decisions. Therefore, the autonomous vehicles have integrated diverse devices to sense their environment. Commonly, an

R. Vera-Rodriguez et al. (Eds.): CIARP 2018, LNCS 11401, pp. 513–520, 2019.
https://doi.org/10.1007/978-3-030-13469-3_60

autonomous vehicle has GPS, encoders, inertial measurement units, cameras, and LiDAR sensors [5]. The last two comprise an artificial vision system used for object detection applications.

Since 2010, autonomous vehicles have incorporated 3D vision systems. Those that stand out the most are LiDAR-based sensors, which reconstruct their surrounding environment in three dimensions by using a point cloud representation. Some studies examined LiDAR as an option for object detection in an autonomous driving environment by using a 2D LiDAR device [7,16]. Due to the low resolution of 2D LiDAR sensors, most of these papers use this device as a part of a sensor fusion scheme for object detection, where the main sensor is an RGB camera. This scheme is limited by hard environmental light conditions. Despite the 2D LiDAR low resolution, the papers that only use this device, achieve good results on vehicle detection, but the information is limited for detecting small objects. With the inclusion of a high definition LiDAR [14] in autonomous driving platforms, there is a growing number of researchers, who have exclusively used 3D information for on-road object detection tasks [1,6,17].

An early approach to the object detection on LiDAR data focused on handcrafted features. Those features are extracted by using different techniques [6]. However, as it is evident from object detection tasks based on RGB images, in recent years, Deep Learning started to dominate 3D data object detection, too.

Based on a KITTI 3D object detection benchmark [], there have been a number of studies, which involve Deep Learning that has reported the best performance on different street object detection. Some methods listed in the benchmark still use RGB images, even if these are only used for 3D region proposals [9]. However, there are explorations which exclusively use LiDAR data for object detection. One of the outstanding methods is VoxelNet. This method applies voxelization of the entire point cloud in order to use it as an organized representation on a Deep Net Architecture [17]. An approach more recently presented implements a multilayer bird's eye view representation of the 3D point cloud to use it into a convolutional network [1]. Despite its results, the representations used may lose information in the transformation process, which can be essential on object detection tasks.

A novel approach for point cloud object classification and segmentation has been proposed by Qi et al. in two architectures: PointNet [10] and PointNet++ [12], which exclusively and directly use raw point clouds. This architecture is mainly based on a symmetric function to transform the point cloud into an orderly invariant space. The PointNet architecture has shown good results on object classification in benchmarks, such as ModelNet [15]. However, this architecture has a limitation as for the number of points to be processed. This architecture uses a fixed number for input points, which makes inefficient its use on data with variable resolution. The data captured by a LIDAR sensor have different resolution for different distances to object. This entails the necessity to improve the PointNet architecture to use it in autonomous vehicle applications.

In this paper, a PointNet evaluation for on-road object detection is presented, which uses LiDAR data. Our work has 3 main blocks. The first one, a 3D object generator, which segments a LIDAR point cloud scene into 3D on-road objects. The second one, a multi-resolution conditioning stage, which adjusts each object point cloud to a specific-resolution. And the third one, six PointNet models at six different resolutions were trained.

This paper is structured as follows: Sect. 2 describes the methodology used to design our method. And Sect. 3 contains the experimental results and conclusions.

2 Framework Overview

Figure 1 represents the global architecture of our approach. Here, block 1 presents the segmentation process, block 2 represents a resolution conditioning process, and block 3 represents a multiresolution approach based on PointNet.

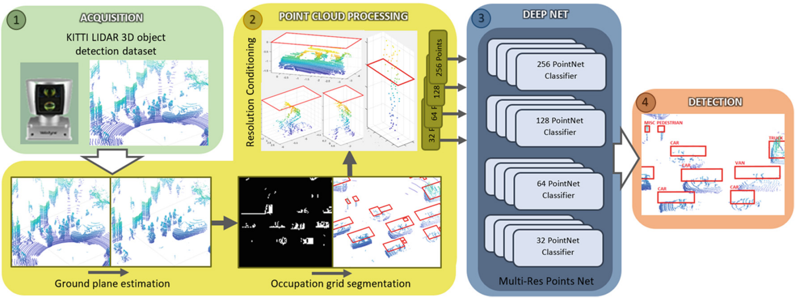

Fig. 1. General scheme of our traffic detection method by using a multi-resolution PointNet architecture.

In the following sub-sections, each block on the Fig. 1 and some tools and techniques needed for the process will be explained.

2.1 Segmentation

The elimination of the ground points is a necessary step for non-ground element segmentation. For this purpose, a modeling of ground plane is done by an algorithm named Random Sample Consensus (RANSAC) [13], which meets the model of the plane that has the higher number of points into a defined distance. The RANSAC algorithm is restricted by 3° from the plane defined by the Z axes and the sensor position height. The result of applying the algorithm is presented in Fig. 2b.

The occupation grid is an approach in 2 dimensions, which has been efficiently used on pedestrian detection [6]. Starting with the point cloud without the ground plane points, a mesh over the XY plane is established. The mesh is defined by 20 cm by 20 cm^2, taking 20 cm as a prudent distance in order to

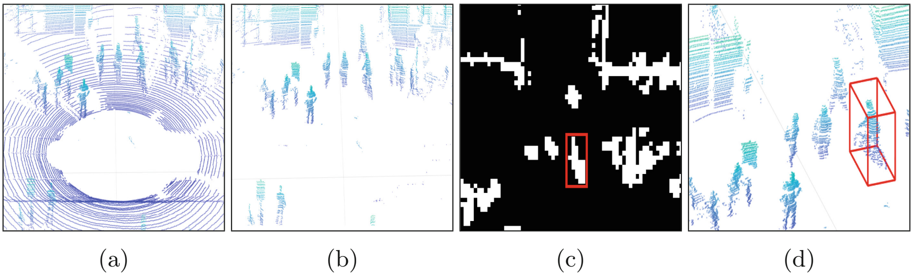

Fig. 2. Segmentation process. (a) Original LiDAR point cloud frame. (b) LiDAR point cloud frame after ground extraction. (c) Occupation grid representation with an object bounding box. (d) 3D bounding box from occupation grid.

establish two point clouds as different objects. All the points on each square of the mesh are projected into a single plane in order to use the number of points into each square as input of the occupation array. To ignore very low resolution elements, squares with less than 4 points are removed.

Figure 2c shows an occupation array where the white points represent 20 by 20 squares with more than 3 points. This array is treated by a connected region algorithm to determine those groups, which compose an object [2]. From the connected regions, the bounding boxes are extracted and these are transformed into 3D bounding boxes giving an extra margin to avoid the loss of points as it is shown in the Fig. 2d. The point cloud into each 3D bounding box is stored with the label information, into an H5 format, which will be used by the Deep Net architecture.

2.2 Resolution Conditioning

In order to take advantage of the parallel computing, the Deep Net architectures use groups of information to compute it simultaneously by reducing the processing time. This batch should be made up of information of the same size or shape. By the nature of the point cloud 3D representation, each object in a given point cloud has a different quantity of points and a group of different objects, which will have a variable shape, for that reason, a resolution conditioning is required to unify a batch size for a Deep Net training.

As our approach is based on multi-resolution, the data prepared to train our method should be separated depending on the number of points, which compose each object. As it will be explained on the deep architecture subsection, our approach uses 6 different resolutions (16, 32, 64, 128, 256, and 512 points). For objects with more than 512 points, a sub-group of 512 points is randomly selected. For objects with points between 256 and 512, 256 points randomly selected and the same is true for the rest of ranges except for objects that have less than 16 points. In this case, points are aggregated to achieve the 16 points resolution, which is made up by the most simple up-sampling procedure. This is performed repeating the existing points in the object until the desired number of points is reached. Since the objective of this paper is to evaluate the performance of the PointNet architecture, it is unwanted to improve the data, which is used in it.

2.3 PointNet Architecture for Multiple Resolution Objects

Due to the multi-resolution of the objects represented by a LiDAR point cloud, a method which can handle a highly variable point quantity is required. For this purpose, a parallel architecture of point nets, which process separately point clouds with six ranges of resolutions is established. This architecture learns specific features for each resolution, which takes advantage of the available information on each object. Besides, this approach avoids the excessive down-sampling or up-sampling, which is required by a single resolution PointNet.

The evaluation approach is made up of six PointNets, where each one will be trained with different batch shapes. Finally, in the evaluation process, the trained model will be selected depending on the quantity of points, which represents the object to be classified. This ensures that the evaluation process will not have more computation cost than a single resolution PointNet.

3 Experiments

3.1 Dataset

The data that is used on our approach was obtained from the KITTI vision benchmark suite. The dataset has 22 sequences of LiDAR data, which represents 39.2 Km in diverse autonomous driving environments. From all the data on the dataset, the 3D object detection benchmark selected 12,000 frames of tridimensional information, which was acquired by an HD LiDAR sensor [4]. From these 29 Gb of LiDAR data, over 40,000 individual objects are extracted and stored in an H5 format with its respective labels.

More than 30,000 objects in the dataset are small vehicles, which are labeled as Car-Van. This implies a highly unbalanced database, which can hinder training process. To improve this issue, new examples from the existent data were generated. It was performed by a sub-sampling process in order to simulate different acquisition distances for the same object.

The existent examples, which have enough points, are randomly sub-sampled in order to generate new objects simulating acquisitions distances of up to 70 m. This process is performed on pedestrians, cyclists, and miscellaneous elements. These classes have a low number of examples compared to the vehicles on the dataset. At the end of the process, the data set is made up of 58,566 objects on multiple resolutions. The dataset is divided in order to get 80% for the training process and 20% for the evaluation.

3.2 Training Process

The model, which is described in Subsect. 2.3, is trained in 6 stages and defined by resolution groups. In order to individually establish the models, which would process each resolution of point clouds, the training process was done by each PointNet architecture separately. The 46,853 objects in the training dataset were conditioned following the procedure described in Sect. 3.2. Since there are some objects, which appear with an extremely low resolution, a threshold of eight points was defined to leave out of the training process, that is to say, those elements with less than eight points. To increase the training data, some of the shelf methods to generate new point clouds from the existent data are applied [11]. From each training object, a new object was created through a random angle rotation over its Z axis, and other were created through adding a random level of noise moving their points from 1 cm to 4 cm. The random noise added to the new objects also gives some robustness in environments like mild frog or rainy conditions. As our approach uses 6 PointNets, the training time increases compared to a single resolution PointNet. The training time of our proposal was near to the double of a single resolution PointNet trained with 512 points, about 45 h in a basic computer with a single GTX650Ti GPU.

3.3 Evaluation Process

To assess the performance of the proposed multi-resolution approach, the PointNet with all the testing dataset, in a single resolution architecture, was tested. In order to keep the original information of the objects with a resolution under the architecture defined, an up-sampling process by using repeated points was applied. The evaluation process of the single resolution PointNet was performed in six different resolutions, separately, in order to see the behavior of the detection across of each resolution. The results are presented in the first six rows of Table 1. As it can be seen, the low resolution objects negatively affect the performance of the PointNet object detector when it is used with 512 and 256 points. Nevertheless, the performance of this architecture is outstanding for the data used for training and evaluation. Interestingly, the results shown in Table 1 indicate that the resolution where each class achieves the better results are not the same. This means that if a model with a specific resolution is chosen, some classes would decrease their score. As far as the experimental evaluation of the multi-resolution architecture, the last row of Table 1 illustrates the performance of the proposed method.

Taken together, these results show that even when the single resolution PointNet can have lightly better results on certain classes, our multi-resolution adaption improves the general performance in 3.5% over the 32 points resolution single PointNet, which is the best result for the single-resolution architecture.

To illustrate the performance over the KITTI dataset, in the Fig. 3 two frames are shown with the detection results.

Table 1. Classification results of PointNet on KITTI 3D object detection data set

Resolution\classes	Average precision					
	Pedestrian	Cyclist	Car-van	Truck-Tram	Misc	General
512 points	0,718	0,156	0,888	0,631	0,186	0,717
256 points	0,887	0,447	0,862	0,632	0,413	0,797
128 points	0,949	0,517	0,828	**0,828**	0,438	0,811
64 points	0,988	0,624	0,884	0,73	**0,441**	0,861
32 points	**0,993**	0,623	**0,965**	0,519	0,146	0,881
16 points	0,987	0,643	0,94	0,63	0,153	0,872
Multiple Res.	0,9899	**0,7563**	0,9507	0,7958	0,3742	**0,9068**

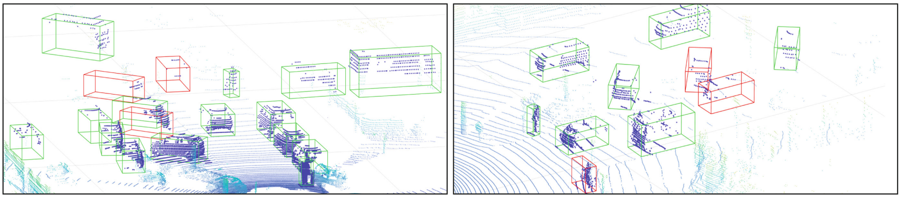

Fig. 3. Performance over KITTI 3D object detection dataset. Green boxes contain the good predictions (Color figure online)

4 Conclusions

The PointNet architecture was evaluated on the detection of on-road 3D objects using multiple resolutions of point cloud data, but without using any RGB information. A proposed resolution conditioning of 3D objects allowed us to use the point clouds representations of these objects on multiple specific resolution nets. This proposed approach using multiple resolutions on PointNet architecture yielded an improvement of 3.5% in general AP compared with the best performance of a single resolution approach. On the pedestrian class, the average precision of detection was 99%, but it is worth to note that this class is easily differentiable from the other objects because of its size characteristics. Future work will entail labeling process in order to increase the number of classes of the dataset, for example including classes like road signs and trees. With this, we attempt to find a more robust detector over the most common on-road objects. The labeling process should also have a data balancing purpose in order to improve the training process.

References

1. Beltran, J., Guindel, C., Moreno, F.M., Cruzado, D., Garcia, F., de la Escalera, A.: Birdnet: a 3D object detection framework from LiDAR information. arXiv preprint arXiv:1805.01195 (2018)
2. Carson, C., Belongie, S., Greenspan, H., Malik, J.: Blobworld: image segmentation using expectation-maximization and its application to image querying. IEEE Trans. Pattern Anal. Mach. Intell. **24**(8), 1026–1038 (2002)
3. Fagnant, D.J., Kockelman, K.: Preparing a nation for autonomous vehicles: opportunities, barriers and policy recommendations. Transp. Res. Part A Policy Pract. **77**, 167–181 (2015)
4. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? The kitti vision benchmark suite. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3354–3361. IEEE (2012)
5. Hwang, S., Kim, N., Choi, Y., Lee, S., Kweon, I.S.: Fast multiple objects detection and tracking fusing color camera and 3D LiDAR for intelligent vehicles. In: 2016 13th International Conference on Ubiquitous Robots and Ambient Intelligence (URAI), pp. 234–239. IEEE (2016)
6. Kidono, K., Miyasaka, T., Watanabe, A., Naito, T., Miura, J.: Pedestrian recognition using high-definition LiDAR. In: 2011 IEEE Intelligent Vehicles Symposium (IV), pp. 405–410. IEEE (2011)
7. Lin, B.Z., Lin, C.C.: Pedestrian detection by fusing 3D points and color images. In: 2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS), pp. 1–5. IEEE (2016)
8. Global status report on road safety 2015. World Health Organization (2015)
9. Qi, C.R., Liu, W., Wu, C., Su, H., Guibas, L.J.: Frustum pointnets for 3D object detection from RGB-D data. arXiv preprint arXiv:1711.08488 (2017)
10. Qi, C.R., Su, H., Mo, K., Guibas, L.J.: Pointnet: deep learning on point sets for 3D classification and segmentation. In: Proceedings of the Computer Vision and Pattern Recognition (CVPR), vol. 1, no. 2, p. 4. IEEE (2017)
11. Qi, C.R., Su, H., Nießner, M., Dai, A., Yan, M., Guibas, L.J.: Volumetric and multi-view CNNs for object classification on 3D data. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5648–5656 (2016)
12. Qi, C.R., Yi, L., Su, H., Guibas, L.J.: Pointnet++: deep hierarchical feature learning on point sets in a metric space. In: Advances in Neural Information Processing Systems, pp. 5099–5108 (2017)
13. Raguram, R., Frahm, J.-M., Pollefeys, M.: A comparative analysis of RANSAC techniques leading to adaptive real-time random sample consensus. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008. LNCS, vol. 5303, pp. 500–513. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-88688-4_37
14. Schwarz, B.: LiDAR: mapping the world in 3D. Nat. Photonics **4**(7), 429 (2010)
15. Wu, Z., et al.: 3D shapenets: a deep representation for volumetric shape modeling. In: CVPR, vol. 1, p. 3 (2015)
16. Xue, J.r., Wang, D., Du, S.y., Cui, D.x., Huang, Y., Zheng, N.n.: A vision-centered multi-sensor fusing approach to self-localization and obstacle perception for robotic cars. Front. Inf. Technol. Electron. Eng. **18**(1), 122–138 (2017)
17. Zhou, Y., Tuzel, O.: Voxelnet: end-to-end learning for point cloud based 3D object detection. arXiv preprint arXiv:1711.06396 (2017)