3D OBJECT DETECTION FROM MOBILE LIDAR POINT CLOUD WITH DEEP LEARNING

Muhammed Enes Atik[1] and Zaide Duran[1]
[1]Department of Geomatics Engineering, Istanbul Technical University (ITU),
Maslak, Istanbul, 34469, Turkey,
Email: atikm@itu.edu.tr; duranza@itu.edu.tr

KEY WORDS: Deep Learning, Object Detection, Point Cloud, YOLO, Mobile LiDAR

ABSTRACT: The usage area of LiDAR technology, which can be detected from the aerial, terrestrial and mobile, is expanding day by day. Especially for mapping and autonomous vehicles, mobile LiDAR offers very useful data. Mobile point clouds are a type of data obtained using laser scanners mounted on a moving vehicle. An accurate sense of space and precise positioning are crucial requirements for reliable navigation and safe driving of autonomous vehicles in complex dynamic environments. Recently, deep learning approaches have been preferred for the evaluation and information extraction of complex mobile LiDAR data. Although successful results have been obtained for camera-based solutions with deep learning, it may not be fast enough in inference paths due to convolution operations. There are improved methods for real-time performance in object detection. Single-shot detectors, like YOLO, are some of the best in this regard. In this study, moving or stationary vehicles, people and cyclists on the point cloud have been detected by deep learning. Vehicles, pedestrians and cyclists were detected with YOLO3D-YOLOv3 and YOLO3D-YOLOv4, which are the developed version of the YOLO algorithm applied to 2D images for 3D point clouds. KITTI benchmark dataset was used in this study. The point cloud is projected onto a grid mesh with a resolution of 0.1 m per pixel in the form of a bird's eye view. The range of a LiDAR patch is 30.4 meters to right and 30.4 meters to the left, and 60.8 meters forward. Input shape of 608x608 per channel is obtained by using this range with the resolution of 0.1 m per pixel. Average mean precision (mAP) results in this study were obtained within the mAP lower limit of 0.5 IoU for each object class. The mAP was obtained as 83.04%. with YOLO3D-YOLOv4 and 81.50% with YOLO3D-YOLOv3.

1. INTRODUCTION

Laser scanners acquire a large number of points within a given field of view in a very short time. Thus, due to the 3D point cloud of the object surface, a detailed, usable 3D model can be produced (Duran and Aydar, 2012; Akyol and Duran, 2014). Point clouds contain information such as 3D position information, density and color (Atik et al., 2021). Mobile point clouds are a type of data obtained using laser scanners mounted on a moving vehicle. An accurate sense of space and precise positioning are crucial requirements for reliable navigation and safe driving of autonomous vehicles in complex dynamic environments (Li et al., 2020). Performing these two tasks requires acquiring and processing highly accurate and information-rich data of real-world environments (Van Brummelen, 2018). Mobile LiDAR point clouds are used in two main areas: (1) real-time environment detection and processing for segmentation and object detection; (2) creating high resolution (HD) maps and urban models for highly accurate positioning and referencing (Levinson et al., 2011).

In general, object detection approaches can be divided into two groups: voxel-based and machine learning-based. In voxel-based approaches, points are not treated individually but as collections of points to be grouped with regular shapes (cube, sphere, prism, etc.). The most important parameter in voxel-based approaches is the size of the local point group to be determined. Voxel-based approaches utilize multiple point clouds obtained by mobile laser scanning (Yoon et al., 2019; Vallet et al., 2015). Deep learning approaches can also be used to detect moving objects directly. PointRNN (Fan and Yang, 2019) is the most popular method developed in this regard. PointRNN based on recurrent neural network (RNN) is one of the few methods developed for detecting moving objects. RNNs are a deep learning approach that involves combining networks in a loop. Each network takes information from the previous network as input, processes it and produces an output for the next network. RNN is a sequential learning model that connects hidden layers to nodes and can learn sequence property dynamically (Atik and Duran, 2021). The main disadvantage of this approach is the need for large data to train algorithms. A labor-intensive process is required to generate training and test data. Deep learning is a new field for object detection, classification and segmentation and still developing (Atik and Ipbuker, 2021).

Three different approaches have been developed for 3D object detection depending on the sensor type: LIDAR-only approaches, camera-only approaches and the LIDAR-camera fusion approaches (Ali et al., 2018). Methods using only Lidar differ in their preprocessing steps. Some approaches project the point cloud onto the 2D plane (Li et al.,

2016; Chen et al., 2016). Some other approaches also convert points to voxels (Li, 2107; Zhou and Tuzel, 2018). Camera-only approaches try to create 3D bounding boxes over images using prior knowledge of objects (Mousavian et al., 2017). There are methods that try to detect 3D objects with stereo images (Chen et al., 2017). The LiDAR-camera fusion approach (Qi et al., 2018; Chen et al., 2016) aims to combine the advantages of both sensors. LiDAR provides rich geometric information, while the camera provides visual features.

In this study, moving or stationary vehicles, people and bicycles on the point cloud have been detected by using YOLO3D-YOLOv3 and YOLO3D-YOLOv4 method. Object detection was performed using mobile laser scanning data with the KITTI dataset.

2. MATERIAL AND METHODS

2.1 the KITTI Dataset

KITTI benchmark dataset (Geiger et al., 2012) (Figure 1) was used in this study. The point cloud was projected in 2D space as a bird view grid map with a resolution of 0.1m per pixel. The range represented from the LiDAR space by the grid map is 30.4 meters to right and 30.4 meters to the left, and 60.8 meters forward. Using this range with the above-mentioned resolution of 0.1 results in an input shape of 608x608 per channel. The height in the LiDAR space is clipped between +2m and -2m and scaled to be from 0 to 255 to be represented as pixel values in the maximum height channel.

Figure 1. Example from KITTI dataset.

2.2 YOLO3D

YOLO3D (Ali et al., 2018), which is the version of the You Only Look Once (YOLO) (Redmon et al., 2016) algorithm developed for 2D images, is used for point clouds. YOLO is an open source object detection algorithm based on convolutional neural networks. You Only Look Once (YOLO) is among the most well-known deep learning algorithms, and it stands out with its speed thanks to its single-stage detection architecture. Yolo can detect objects in an image at a glance. With a single neural network, it marks all objects in the image with bounding boxes (Cepni et al., 2020). In addition, being CPU-based and open source code stand out as the advantages of YOLO.

YOLO3D is a method developed by extending YOLO architectures. The 3D point cloud converted to bird's eye view is used as input. The network architecture was created by redefining YOLO in accordance with the LiDAR structure. The estimation output is 9 values: Object Bounding Box (OBB) center in 3D (x, y, z), the 3D dimensions (length, width and height), the orientation in the bird-view space, the confidence, and the object class label (Ali et al., 2018).

In this study, YOLOv3 and YOLOv4 architectures. The main difference between YOLO-v3 and YOLO-v4 algorithms is the backbone structure they use. Darknet53 is used for YOLO-v3, while CSPDarknet53 is used for YOLO-v4. Accordingly, YOLOv3 consists of a total of 106 layers, while YOLO-v4 contains a total of 161 layers (Bochkovskiy et al., 2020). The YOLOv4 architecture uses a modified Path aggregation network, a modified spatial attention module, and a modified spatial pyramid pool to gather information that improves accuracy.

3. METHODOLOGY

The YOLO3D-YOLOv3 model is trained in an end-to-end fashion. Stochastic gradient descent with a momentum of 0.9, and a weight decay of 0.0005 were used. The model was trained the network for 60 epochs, with a batch size of 2. The weights of the pre-trained model were used as initial weights. Similar training parameters were used for YOLO3D-YOLOv4. In YOLO3D-YOLOv4, the epoch number is 93. The early stopping method was used for optimization. Before reaching the determined number of epochs, training was stopped in an epoch with high accuracy.

Since the test labels are not shared, the labeled data is divided into training and testing. The dataset consists of sequential mobile LiDAR frames. From the training set of 7481 images, 6000 images are used for training and remaining 1481 images are used for validation. The mAP results reported in this project are evaluated into this valid set with custom mAP evaluation script with 0.5 IoU for each object class. Google Colab was used as the development environment. Google Colab is a cloud service that provides GPU support by Google. The cloud system has NVIDIA Tesla P100 GPU and 25 GB RAM.

3. RESULTS AND DISCUSSION

As a result of the study, YOLO3D-YOLOv4 and YOLO3D-YOLOv3 methods were compared. 3D object detection was made in 1481 samples used for the test. Precision, recall, F1-score and average precision (AP) were used as accuracy metrics. YOLO3D-YOLOv4 has a higher value than YOLO3D-YOLOv3 in all evaluation metrics. The mAP obtained with YOLO3D-YOLOv4 is 83.04, while the mAP obtained with YOLO3D-YOLOv4 is 81.50. When class metrics are evaluated, recall values are higher than precision values. That is, even if not all objects in the point cloud can be detected, almost all of the detected objects are assigned to the correct class. Because there are many examples in the dataset, the car metrics belonging to the car class are larger than the cyclist and pedestrian. The precision, recall, F1 score and AP values of the car class in YOLO3D-YOLOv4 are 85.20, 98.66, 91.44 and 97.80, respectively. YOLO3D-YOLOv4 has more layers than YOLO3D-YOLOv3. Thus, the weights can be calculated more accurately. Also, YOLO3D-YOLOv4 has higher accuracy on small objects.

Table 1. Results of the methods.

	Precision		**Recall**		**F1 Score**		**AP**	
Class	**YOLOv3**	**YOLOv4**	**YOLOv3**	**YOLOv4**	**YOLOv3**	**YOLOv4**	**YOLOv3**	**YOLOv4**
Car	81.64	**85.20**	**98.85**	98.66	89.42	**91.44**	97.58	**97.80**
Pedesterian	35.73	**41.83**	92.01	**93.29**	51.57	**57.76**	66.39	**67.13**
Cyclist	42.81	**57.24**	93.77	**94.14**	58.78	**71.19**	80.54	**84.18**

Although the detection rate of pedestrians and cyclists is low, the classes of those detected are correct. Very successful results were obtained in the detection of cars. Almost all of the cars were correctly identified. With the results obtained in a dynamic data, YOLO3D is particularly suitable for use for car detection from point cloud. The accuracy of these classes can be increased by increasing the number of pedestrians and cyclists in the training data. Images of the results are shown in Figure 2. The detection speed of the YOLO3D from the point cloud varies between

35 and 60 ms for a frame. As in 2D images, the YOLO algorithm can quickly detect objects in 3D point clouds. There is no difference between the versions in terms of estimation time.

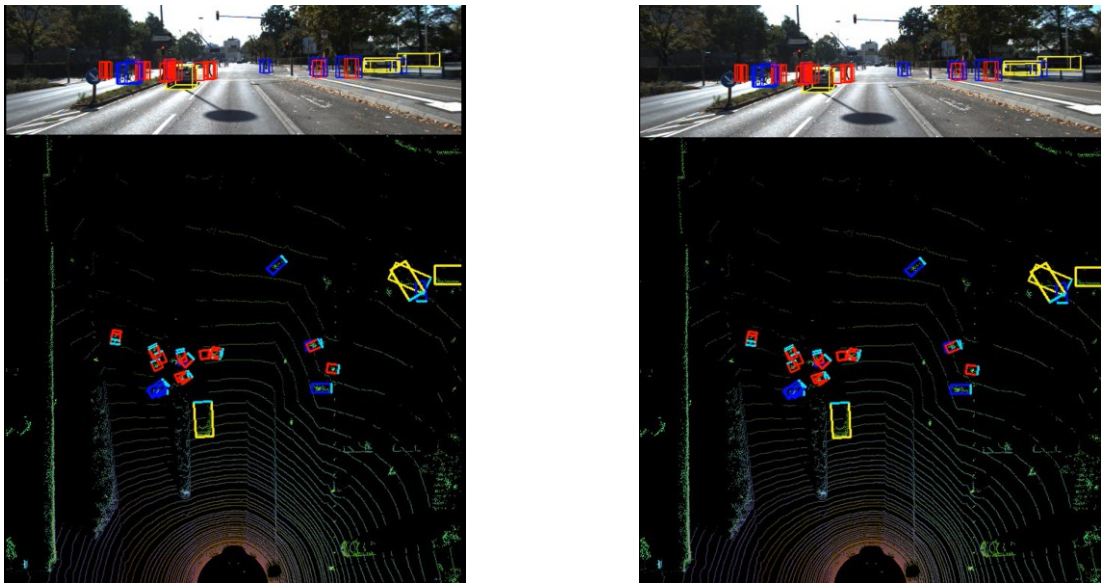

Figure 2. Result scenes. Yolo3D-Yolov4 on the left and YOLO3D-YOLOv3 on the right.

4. CONCLUSION

In this study, 3D object detection was performed on mobile lidar data using different versions of YOLO3D. Training time and amount of data depend on hardware power. Therefore, training can be developed with high-performance hardware. In addition, training data for the cyclist and pedestrian classes can be added. In future studies, other datasets and algorithms in the literature can be examined. In addition, the classes to be determined in the dataset can be multiplied. Moving objects should be treated as a separate problem. Special solutions should be produced for these. Deep learning has great potential for object detection in point clouds.

REFERENCES

Akyol, O., Duran, Z. 2014. Low-cost laser scanning system design. Journal of Russian Laser Research 35(3), 244-251.

Ali, W., Abdelkarim, S., Zidan, M., Zahran, M., El Sallab, A. 2018. Yolo3d: End-to-end real-time 3d oriented object bounding box detection from lidar point cloud. In Proceedings of the European Conference on Computer Vision (ECCV) Workshops, 1-12.

Atik, M. E., Duran, Z. (2021). Classification of Aerial Photogrammetric Point Cloud Using Recurrent Neural Networks. Fresenius Environmental Bulletin 30(4 A), 4270-4275.

Atik, M. E., Duran, Z., Seker, D. Z. (2021). Machine Learning-Based Supervised Classification of Point Clouds Using Multiscale Geometric Features. ISPRS International Journal of Geo-Information 10(3), 187.

Atik, S. O., Ipbuker, C. (2021). Integrating Convolutional Neural Network and Multiresolution Segmentation for Land Cover and Land Use Mapping Using Satellite Imagery. Applied Sciences 11(12), 5551.

Bochkovskiy, A., Wang, C. Y., Liao, H. Y. M. 2020. Yolov4: Optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934.

Cepni, S., Atik, M. E., Duran, Z. 2020. Vehicle detection using different deep learning algorithms from image sequence. Baltic Journal of Modern Computing 8(2), 347-358.

Chen, X., Kundu, K., Zhu, Y., Ma, H., Fidler, S., Urtasun, R. 2017. 3d object proposals using stereo imagery for accurate object class detection. IEEE transactions on pattern analysis and machine intelligence 40(5), 1259-1272.

Chen, X., Ma, H., Wan, J., Li, B., Xia, T. 2017. Multi-view 3d object detection network for autonomous driving. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 1907-1915.

Duran, Z., Aydar, U. 2012. Digital modeling of world's first known length reference unit: The Nippur cubit rod. Journal of cultural heritage 13(3), 352-356.

Fan, H., Yang, Y. 2019. PointRNN: Point recurrent neural network for moving point cloud processing. arXiv preprint arXiv:1910.08287.

Geiger, A., Lenz, P., Urtasun, R. 2012. Are we ready for autonomous driving? the kitti vision benchmark suite. In 2012 IEEE conference on computer vision and pattern recognition pp. 3354-3361. IEEE.

Levinson, J., Askeland, J., Becker, J., Dolson, J., Held, D., Kammel, S., Thrun, S. 2011. Towards fully autonomous driving: Systems and algorithms. In 2011 IEEE Intelligent Vehicles Symposium (IV), 163-168.

Li, B. 2017. 3d fully convolutional network for vehicle detection in point cloud. In 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) pp. 1513-1518. IEEE.

Li, B., Zhang, T., Xia, T. 2016. Vehicle detection from 3d lidar using fully convolutional network. arXiv preprint arXiv:1608.07916.

Li, Y., Ma, L., Zhong, Z., Liu, F., Chapman, M. A., Cao, D., Li, J. 2020. Deep learning for LiDAR point clouds in autonomous driving: a review. IEEE Transactions on Neural Networks and Learning Systems 99, 1-21.

Mousavian, A., Anguelov, D., Flynn, J., Kosecka, J. 2017. 3d bounding box estimation using deep learning and geometry. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition pp. 7074-7082.

Qi, C. R., Liu, W., Wu, C., Su, H., Guibas, L. J. 2018. Frustum pointnets for 3d object detection from rgb-d data. In Proceedings of the IEEE conference on computer vision and pattern recognition pp. 918-927.

Redmon, J., Divvala, S., Girshick, R., Farhadi, A. 2016. You only look once: Unified, real-time object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition pp. 779-788.

Vallet, B., Xiao, W., Brédif, M. 2015. Extracting mobile objects in images using a velodyne lidar point cloud. ISPRS annals of the photogrammetry, remote sensing and spatial information sciences 2(3), 247.

Van Brummelen, J., O'Brien, M., Gruyer, D., Najjaran, H. 2018. Autonomous vehicle perception: The technology of today and tomorrow. Transportation research part C: emerging technologies 89, 384-406.

Yoon, D., Tang, T., Barfoot, T. 2019. Mapless online detection of dynamic objects in 3d lidar. In 2019 16th Conference on Computer and Robot Vision (CRV), 113-120.

Zhou, Y., Tuzel, O. 2018. Voxelnet: End-to-end learning for point cloud based 3d object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition pp. 4490-4499.