ESANN 2021 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 6-8 October 2021, i6doc.com publ., ISBN 978287587082-7.
Available from http://www.i6doc.com/en/.

Cross-modal verification for 3D object detection

Haodi ZHANG, Alexandrina ROGOZAN and Abdelaziz BENSRHAIR *

LITIS LAB - INSA de ROUEN - Normandie Universite
786 Avenue de l'Universite 76000 Rouen- France

Abstract. To overcome the deficiency in the single modality of LiDAR point cloud, we propose a cross-modal verification (CMV) model for reducing 3D object detection false positives. The abundant color and texture information in image modality allow the classification of the projection region of 3D bounding box proposal in the image plane. Three 3D object detectors are adopted as backbone and eight evaluation metrics are used to fully investigate the proposed model. The experiment results show that the proposed CMV model removes more than 50% of false positives in 3D object detection proposals and significantly improves the performance of 3D object detection.

1 Introduction

3D object detection is a key component of the autonomous vehicle perception system. Based on this component, the object shape and spatial position information are obtained. Furthermore, image scale variation and perspective transformation problems can be easily avoided in 3D object detection. Therefore, 3D object detection has become a key research area.

However, due to the difference in data modalities, the detection algorithms based on the image modality need to be modified to adapt to the point cloud modality. To overcome the shortage of feature extraction from point cloud modality, [1] and [2] propose the point-wise and the voxel-wise feature extraction modules, respectively. Based on these two feature extraction modules, many LiDAR-based 3D object detectors have been proposed [3, 4, 5].

For the image modality, there are two approaches to implement 3D object detection. 1) [6] [2] attempt to locate 3D objects only using image modality. Since the depth information in the image modality is non-explicit, the detection precision of the image-based method is significantly lower than that of the LiDAR-based method. 2) The other approach is to fuse images and LiDAR. With two data modalities, it is able to detect 3D object robustly and accurately [7].

We notice that there are still a large number of false positives (FP) in the results of state-of-the-art multi-modal fusion 3D detection methods. Hence, we propose a cross-model verification (CMV) model that uses image modality to filter the detection results of LiDAR-based 3D detectors. The proposed CMV model is a novel method to address result-level fusion. All 3D detection proposals are projected onto image plane. The proposed CMV model removes FPs based

*This work was supported by the China Scholarship Council (CSC) under grant 201806960147. The authors would like to thank the Centre Regional Informatiqueet d'Applications Numeriques de Normandie (CRIANN) for providing computational resources.

ESANN 2021 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 6-8 October 2021, i6doc.com publ., ISBN 978287587082-7.
Available from http://www.i6doc.com/en/.

on image classification results. The experimental results show that the CMV model can significantly reduce the number of FPs by up to 50%.

2 Related Work

2.1 LiDAR-based 3D Object Detection

The data structure of LiDAR point cloud is irregular and orderless. A traditional CNN designed for dense image modality could not operate properly. To accommodate the data format required for CNN, some detection methods project the LiDAR point cloud into a bird's-eye view (BEV) image. In order to extract features from orderless point clouds, PointNet [1] propose to use symmetry function. This sophisticated design has been referenced and used in many 3D object detectors. Due to the large number of LiDAR point clouds, VoxelNet [2] proposes to use Voxel Feature Extractor (VFE) for feature extraction of the voxelized point cloud. This kind of method usually yields better results.

2.2 Image-based Object Classification

Image classification is a fundamental research in computer vision where deep neural networks have achieved great success. Some milestones like the LeNet-5, YOLO and SSD continue to improve the image recognition capabilities. A typical image object classifier consists of two modules: CNN-based feature extraction and class regression based on the fully connected layer.

2.3 Multi-modal Fusion 3D Object Detection

Multi-modal fusion is used to compensate for the instability of a single modality. LIDAR provides accurate depth information and is not affected by lighting conditions. Images can provide object color and texture detail information. The multi-modal 3D object detection methods can be categorized as result level, feature level and multi level depending on the fusion level. Current result-level fusion methods focus on fusing images with LiDAR detection results based on projection and geometric constraints.

We find that various 3D object detectors have plenty of false positives in their detection results. By using the cross-modal verification (CMV) model, these false positives can be easily identified and removed. The proposed CMV model is plug-and-play and can be easily integrated into various 3D object detectors to help improve detection performance.

3 Methods

3.1 Cross-modal verification 3D object detector

The fundamental basis for cross-modal verification of LiDAR with images is that the LiDAR could be projected onto image plane. For the KITTI 3D object detection dataset [8], the projected point $\mathbf{pt}_{img} = (u, v)$ in the image plane of

ESANN 2021 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 6-8 October 2021, i6doc.com publ., ISBN 978287587082-7.
Available from http://www.i6doc.com/en/.

the spatial point $\mathbf{pt}_{pc} = (x, y, z)$ can be obtained by $\mathbf{pt}_{img} = \mathbf{P}_{rect}\mathbf{R}_{rect}\mathbf{pt}_{pc}$. Where $\mathbf{P}_{rect} \in \mathbb{R}^{3\times4}$ is the projection matrix after rectification. $\mathbf{R}_{rect} \in \mathbb{R}^{4\times4}$ is the expanded rectifying rotation matrix.

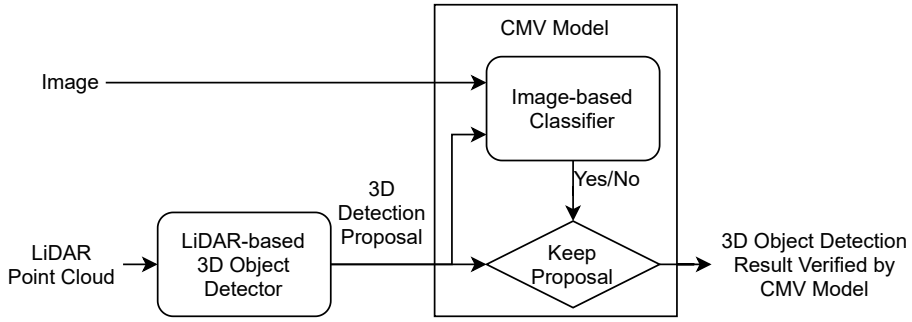

Fig. 1: Cross-modal verification for 3D object detection.

The proposed cross-modal verification 3D object detector has two modules which are the LiDAR-based 3D object detector and the cross-modal verification (CMV) model as shown in Fig. 1. All existing 3D object detectors can be employed to generate 3D bounding box proposals. The CMV model has an image-based classifier to output confidence scores which indicate the selection of 3D bounding box proposals. The deep neural network is used to achieve the image object classification. Four convolutional layers with 3×3 kernel are deployed for feature extraction. After each convolutional layer, the feature maps are downsampled by a factor of 2 using maxpooling. The output of the class regression for the cascaded two fully connected layers are '**Foreground**' and '**Background**'. The class '**Foreground**' includes *Cars*, *Pedestrians* and *Cyclists*. The class '**Background**' consists of streets, skies, trees and road signs.

3.2 Autonomous Driving Object Recognition Dataset

The autonomous driving object recognition dataset is build on the KITTI 3D object detection dataset. Based on the ground truth, we crop and keep the object regions in the images as the **Foreground** class elements. To increase the diversity of the samples, the object images with an overlap of 0.7 with ground truth are cropped at each of the four vertices of the bounding box. The objects which are too small have been removed based on two criteria, $max(h, w) > 15$ pixels and $min(h, w) > 10$ pixels. Where h and w are the height and width of the object images.

There are six classes in KITTI 3D object detection dataset. The most mentioned classes are *Car*, *Pedestrian* and *Cyclist*. These three classes are considered as the **Foreground** class in our autonomous driving object recognition dataset. For the **Background** class, we use an algorithm to randomly crop background images in KITTI dataset. To avoid making some objects appears in the **Background** class images, the algorithm checks the overlap between the generated

ESANN 2021 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 6-8 October 2021, i6doc.com publ., ISBN 978287587082-7.
Available from http://www.i6doc.com/en/.

background bounding boxes with every ground truth. There are 114342 **Foreground** class images and 94052 **Background** class images for training. As for testing, there are 122198 **Foreground** class images and 99676 **Background** class images.

4 Experiments

4.1 Methodology

3D object detection. Three representative detectors are selected as the baseline models for evaluation, which are SECOND [3], PointPillars [4], and PartA2 [5]. Since the proposed cross-modal verification model is plug-and-play, these three detectors can be added directly to obtain performance improvements with minimal modifications. There are 7481 training frames and 7518 test frames with both modalities in KITTI 3D object detection dataset. The annotations of training data are available for public access. While for the testing split, only data and calibration files are provided. To evaluate the performance of the detector, we divide the training data into two splits according to [3], which are used for training and testing.

Cross-modal verification Model. The classifier for the cross-modal verification is begin with four convolutional layers, followed by two fully connected layers. The convolution kernel size is 3×3 with a 2×2 maxpooling. The input image is resized to 64×64 pixels. After four convolutional layers, the output feature map size is 4×4 pixels with the batch size of 256. The first fully connected layer reduces the number of features from $4 \times 4 \times 256$ to 128. Then the last fully connected layer output the scores of two classes. The autonomous driving object recognition dataset are used to train and test the CMV model.

4.2 Experiment Results

For 3D object detection task, the most adopted metric is the average precision (AP). The precision and recall rate are calculated from true positive (TP), false positive (FP) and false negative (FN). We also calculate the F1-score and F2-score based on the precision and recall in order to compare the improvement effects.

The experiment results for 3D object detection of three classes are shown in Table 1. To be classified as a true positive (TP), the detection result bounding box should at least has an intersection over union (IoU) more than 0.7 with the ground truth bounding box. The false positives (FP) of 3D object detection results have been greatly reduced due to the cross-modal verification (CMV) model. As a result, the precision has a greatly improved by up to 18.81%. In contrast, the recall rate is slightly decreased since the CMV module only removes FPs from the detection results and does not bring in new proposals. The effectiveness of CMV should be quite obvious when we compare F1-scores and F2-scores. All results are improved for F1-score, where precision and recall

ESANN 2021 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 6-8 October 2021, i6doc.com publ., ISBN 978287587082-7.
Available from http://www.i6doc.com/en/.

are considered equally important. Even for the F2-score, where recall is more emphasized, all scores have been increased.

Class	Model	TP	FP	FN	PRE	REC	F1	F2	AP
CAR	SECOND	6611	6999	1175	48.57	84.90	61.79	73.85	79.68
	SECOND-CMV	6546	**3788**	1241	**63.34**	84.06	**72.24**	**78.90**	**79.93**
PED	SECOND	1220	3363	519	26.62	50.93	34.96	43.06	56.92
	SECOND-CMV	1146	**1549**	593	**42.52**	48.01	**45.10**	**46.80**	55.19
CYC	SECOND	456	2158	88	17.44	27.95	21.48	24.95	65.35
	SECOND-CMV	422	**795**	122	**34.67**	25.37	**29.30**	**26.81**	64.07
CAR	PointPillars	6423	7367	1366	46.57	82.46	59.53	71.45	75.99
	PointPillars-CMV	6348	**4243**	1441	**59.93**	81.49	**69.07**	**76.02**	**76.25**
PED	PointPillars	1126	6555	613	14.65	45.18	22.13	31.89	44.58
	PointPillars-CMV	1071	**2252**	668	**32.22**	42.63	**36.70**	**40.04**	**46.69**
CYC	PointPillars	405	2144	137	15.88	22.86	18.75	21.02	59.77
	PointPillars-CMV	383	**721**	160	**34.69**	20.99	**26.16**	**22.79**	**60.74**
CAR	PartA2	6770	5142	1027	56.83	86.82	68.69	78.53	83.12
	PartA2-CMV	6749	**3908**	1048	**63.32**	86.55	**73.14**	**80.64**	**83.25**
PED	PartA2	1256	4109	483	23.41	55.01	32.84	43.31	54.54
	PartA2-CMV	1216	**2146**	523	**36.16**	53.71	**43.22**	**48.96**	**54.82**
CYC	PartA2	468	1060	75	30.62	31.30	30.96	31.16	73.33
	PartA2-CMV	454	**587**	89	**43.61**	30.22	**35.70**	**32.20**	73.17

Table 1: Results of cross-modal verification for 3D object detection.

We notice that even false positives have been removed more than half, the average precision do not improve significantly. The reason for this evaluation result is that the average precision does not consider FP to be as important as TP. In the process of calculating the average precision, all detection results are sorted by the given class confidence scores. Then N interpolation intervals are applied to integrate the area under curve (AUC) of the precision-recall (PR) curve. Due to the low confidence scores of most FPs, they only have an impact in the last interpolation interval, which accounts for just 1/N of the overall average precision. At the same time, the reduction of TP may lead to a decrease in the number of interpolation intervals. This eventually leads to a significant degradation of average precision. Therefore, since the average precision calculation takes TP more importantly, it leads to the results in Table 1.

5 Conclusion

In this paper, a cross-modal verification (CMV) model is proposed for reducing 3D object detection false positives. The extensive color and texture information in the image modal is used to complement the deficiencies of the LiDAR point cloud modal. The proposals obtained from the 3D object detector are verified by the CMV model and the false positives are discarded. To train the CMV model, an autonomous driving object recognition dataset is build. Three 3D object detectors and eight metrics are adopted to fully investigate the proposed model. All experimental results show the enhancement of the proposed CMV

ESANN 2021 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 6-8 October 2021, i6doc.com publ., ISBN 978287587082-7.
Available from http://www.i6doc.com/en/.

model for 3D object detection.

Our future work will focus on embedding CMV model into 3D object detectors to obtain an end-to-end architecture. In this way, the embedded model has the possibility to process the proposal of the original proposals. Another possible improvement would be to perform alternative metrics for 3D object detection to obtain a comprehensive evaluation of TP and FP. The object classifier in the experiments can be replaced by other classifiers to obtain better FP reduction performance.

References

[1] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 652–660, 2017.

[2] Yin Zhou and Oncel Tuzel. Voxelnet: End-to-end learning for point cloud based 3d object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4490–4499, 2018.

[3] Yan Yan, Yuxing Mao, and Bo Li. Second: Sparsely embedded convolutional detection. *Sensors*, 18(10):3337, 2018.

[4] Alex H Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. Pointpillars: Fast encoders for object detection from point clouds. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 12697–12705, 2019.

[5] Shaoshuai Shi, Zhe Wang, Jianping Shi, Xiaogang Wang, and Hongsheng Li. From points to parts: 3d object detection from point cloud with part-aware and part-aggregation network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[6] Di Feng, Christian Haase-Schuetz, Lars Rosenbaum, Heinz Hertlein, Claudius Glaeser, Fabian Timm, Werner Wiesbeck, and Klaus Dietmayer. Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges. *IEEE Transactions on Intelligent Transportation Systems*, 2020.

[7] Yaodong Cui, Ren Chen, Wenbo Chu, Long Chen, Daxin Tian, Ying Li, and Dongpu Cao. Deep learning for image and point cloud fusion in autonomous driving: A review. *IEEE Transactions on Intelligent Transportation Systems*, 2021.

[8] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013.