DNN Based Camera and Lidar Fusion Framework for 3D Object Recognition

K Zhang1,*, S J Wang2, L Ji3 and C Wang1

1E/E department, Brilliance Automobile Engineering Research Institute, Shenyang 110141, China
2School of Mechanical Engineering, Shenyang University of Technology, Shenyang, 110023, China
3School of Mechatronics Engineering, Shenyang Aerospace University, Shenyang, 110136, China

*hit_zhangkun@126.com

Abstract. A 3-stages deep neural network (DNN) based camera and lidar fusion framework for 3D objects recognition is proposed in this paper. First, to leverage the high resolution of camera and 3D spatial information of Lidar, region proposal network (RPN) is trained to generate proposals from RGB image feature maps and bird-view (BV) feature maps, these proposals are then lifted into 3D proposals. Then, a segmentation network is used to extract object points directly from points inside these 3D proposals. At last, 3D object bounding box instances are extracted from the interested object points by an estimation network followed after a translation by a light-weight TNet, which is a special supervised spatial transformer network (STN). Experiment results show that this proposed 3d object recognition framework can produce considerable result as the other leading methods on KITTI 3D object detection datasets.

1. Introduction

With the development of deep neural network, great progress has been made in vision based 2D object recognition [1,2] and segmentation [3], among which the art car detection methods have achieved an average precision (AP) of over 90%. Autonomous vehicles (AV) require precise 3D understanding to achieve accurate perception and smart planning, but the 3D recognition result was not that good as visual based recognition methods, which is still a great challenge. The AV perception system is usually designed with redundancy, which includes various kinds of sensors, like camera, lidar, radar, ultrasonic sensor, etc.. While the ultrasonic sensors are mainly used for near range object detection such as parking assist, and radar is usually supplied as a black box, which are considered out of the scope of this paper. According to the requirement of AV planning and control systems, the tasks of 3D perception include bounding box estimation (box position, size and orientation), type classification (car, truck, pedestrian or cyclist) and moving state estimation (speed and acceleration). While the moving state is usually estimated by extended Kalman filter, this paper mainly focus on the 3D object bounding box estimation and type classification [4].

Lidar has the advantage of accurate depth measurements, while the RGB image acquired by camera contains more semantic information, its believed the fusion of lidar and camera will achieve better results. Lidar points are commonly represented as 3D points cloud, which is much sparse than RGB...
image. Inspired by the achievements that neural networks made in 2D perception, many works have been done to utilize the DNN based methods in 3D point cloud procession. State-of-the-art 3D object recognition methods either leverage the mature 2D convolutional neural network (CNN) framework either by projecting the point cloud to BV[5-7] or FV (Front View)[8, 9], or down sample the point cloud as 3D voxels[10, 11], these projections or down sampling, however, may lost the depth or spatial information of 3D point cloud. XiaoZhi Chen raised a MV3D method, which first represented the sparse 3D point cloud as a synthetic BV image, directly generated 3D proposals from this image, then pooled information from BV, Front View (FV) and raw image feature maps using those proposals as region of interest (ROI), and then used a region-based network to classify multiclass objects and estimated their 3D boxes. Qi raised a neural network called PointNet to directly fulfil jobs as feature extraction, object classification and semantic point-wise segmentation from raw point cloud [12, 13]. In order to extend PointNet to instance level 3D object recognition function, in their following works, 2D proposals are first generated from mature CNN based 2D image detectors and are the raised to a 3D frustum, the points inside which are the then segmented and the 3D bounding boxes are estimated using a bounding box regression network [14]. Inspired by the region proposal network (RPN) proposed by Faster-RCNN[15], Jason et al. proposed a 3D RPN using predefined 3D anchor boxes and multimodal feature fusion of BV feature maps and original image feature maps, the feature ROIs are then cropped, resized and fused to utilize 3D detection[5]. However, the effect of little objects detections like pedestrians or cyclists, which are commonly occluded by other objects, was not that good as vehicle detections.

In contrast to previous works, we believe that every object lives in 3D space, so the 3D spatial information is as important as the RGB information in 3D object detection. Thus, a fusion 3D object recognition framework is proposed in this paper, which is consisted of 3 stages: region proposal generation, 3D object points segmentation and 3D bounding box estimation & classification, as is shown in figure 1. First, two 2D RPNs are constructed to generate proposals from RGB image feature maps and BV feature maps separately, the 2D proposals are lifted into 3D frustum or cuboid, fused through non-maximum suppression (NMS) to generate 3D proposals, the CNN features of these two RPNs are also shared by the segmentation network and estimation network; then, a segmentation network is used to extract object points directly from points inside these 3D proposals; at last, 3D object bounding box instances are extracted by an estimation network followed after a light-weight spatial transformer network.

![Figure 1](attachment:image.png)

Figure 1. camera and lidar fusion 3D object recognition framework

Experiment results on KITTI dataset show that this proposed fusion 3D object recognition framework can produce considerable result as other leading methods. The main contributions of the study are: (1) region proposal generation from both RGB image feature maps and bird view point image feature maps. (2) object point segmentation and bounding box estimation using 3D point cloud and RGB image fused features. (3) CNN feature sharing between RPN, segmentation network and estimation network.
2. Fusion 3D object recognition framework
As is illustrated in figure 1, the architecture of the proposed fusion 3D object detection framework is consisted of three stages: 2D region proposal generation, 3D object point segmentation and 3D bounding box estimation & classification, as is shown in figure 2. The detailed design of these 3 stages are introduced below.

2.1. Problem definition
Given a lidar point cloud and a corresponding RGB camera image (front view in this paper), the goal of 3D object recognition is to classify the types of the interested objects (vehicle, pedestrian and cyclist for this paper) in the field of view (FOV) of the RGB image and estimate their 3D bounding boxes. Each object bounding box is parameterized by a class type (one of \(n\) types), box center \((c_x, c_y, c_z)\), box size \((h, w, l)\) and its orientation \((\theta)\), which is yaw angle for this paper.

2.2. Region proposal generation
Inspired by the RPN designed by Ren and Girshick, two 2D RPNs are built to generate proposals, one from RGB image and the other from BV point image. The designed RPN network is consisted of an encoder-decoder form feature extractor and 2 conv layers for region proposal generation, as is shown in figure 2.

![Figure 2. Structure of the proposed RPN](image)

For the encoder, a modified version of VGG16 is used, with reduction the number of channels by half, and abnodation conv5 and its following layers. Although the output feature map has rich features, their resolution is 1/8 less then the original input, which would not be enough to extract proposals, especially for the sparse BV image. To solve this problem, a feature pyramid network (FPN) decoder is designed, like [16], which can generate feature maps with a relative high resolution through a deconvolution and concatenation operation. This feature volume are also shared by the segmentation network and estimation network. Then, anchor boxes with sizes of \(\{16, 32, 64, 128\}\) and aspect ratio of \(\{1/3, 1/2, 1, 2, 3\}\) are used to predict region proposals.

For the bird view, a synthetic image representation like AVOD is used, which is generated from point cloud with height, intensity and density information. Since BV is based on orthogonal projection, the proposals generated by RPN is raised to a 3D space as a cuboid, with max/min point height as top and bottom surface. On the other hand, the raw RGB image proposals are lifted to a 3D frustum as PointNet. The proposals are then aggregated through non-maximum suppression (NMS).

2.3. Object point segmentation
As occlusion and background clutter are common in road scenarios, the points inside these proposed 3D proposals may contain non-interest foreground or background clutters, which would severely distract the bounding box estimation performance if not taken care of. Intuitively, both high density image information and spatial 3D point cloud information are useful to cluster foreground, interested object and background, a segmentation network is designed to extract object points directly from points inside these 3D proposals plus 2D image features.

The structure of segmentation network is illustrated in figure 3, which is similar to PointNet proposed by Charles R Q with some modifications. It includes two shared point embedding multilayer perceptron (MLP) layers, a max pooling layer and a point-wise classification MLP layer. In comparison to original PointNet, the ROI pooled pyramid features from the raw RGB image and BV
image are also fused to the point-wise features, to predict the probability of every point as interested object point or not.

![Figure 3. Structure of the segmentation network](image)

After the object segmentation, the points belonging to the object are extracted and the non-interested point clouds are discarded, which may be ground, road side trees, or part point from other objects.

### 2.4. Bounding box estimation and classification

As is analyzed, there is still a quite large distance between the segmented object point clusters and the estimated bounding box center, so a light weight TNet is used to estimate the translation, whose structure is shown in figure 4, includes one MLP layer, a max pooling layer and 3 FC layer.

![Figure 4. Structure of STN](image)

The proposed TNet is a special case of STN with explicitly supervision, whose structure is designed as the one proposed by Charles R Q. It includes one MLP layer, a max pooling layer and 3 FC layer, with some feature fusion modifications.

![Figure 5. Structure of estimation network](image)

The structure of the estimation network is illustrated in figure 5, which includes one MLP layer, a max pooling layer and 3 FC layer. It’s a modified version of PointNet, which fused the ROI pooled pyramid features from the raw RGB image and BV image to the point cloud global feature to estimate 3d bounding box parameters.

### 3. Experiments and results

In order to test the effect of proposal generation and 3D bounding box estimation, tests are done on KITTI 3D object detection dataset.
3.1. Implementation details
The KITTI 3D object detect provide synchronized lidar point clouds and corresponding FOV images, at the same time the ground truth bounding boxes are also provided. The training set contains 7481 frames, which are separated to 2 subsets, one set with 3717 frames is used for the network training and the other one with 3769 sets is used for validation. Those points lied inside the given ground truth bounding boxes are labelled as the segmentation ground truth. Choose the batch size as 32, the initial learning rate as 0.001, which is decayed per 60000 iterations, it takes nearly 30 hours for the network to be trained for 100 epochs on a single GTM980M GPU.

3.2. KITTI test results
The average precision (AP) of this designed camera and lidar information fused 3D object recognition is shown in Table 1, for moderate car/pedestrian/cyclist detection job, the AP is 69.21/43.23/55.34, although there is still a long way to go before real-life usage, the result is still considerable to the result of the other leading detection methods.

| Method     | car easy | car moderate | car hard | pedestrian easy | pedestrian moderate | pedestrian hard | cyclist easy | cyclist moderate | cyclist hard |
|------------|----------|--------------|----------|-----------------|---------------------|-----------------|--------------|-----------------|--------------|
| Point-Frustum | 81.20    | 70.39        | 62.19    | 51.21           | 44.89               | 40.23           | 71.96        | 56.77           | 50.39        |
| AVOD       | 81.94    | 71.88        | 66.38    | 50.80           | 42.81               | 40.88           | 64.00        | 52.18           | 46.61        |
| Ours       | 80.25    | 69.21        | 60.15    | 51.04           | 43.23               | 40.77           | 67.26        | 55.34           | 45.26        |

The precision-recall curves for cars, pedestrians and cyclists are shown in figure 6, which is also considerable to the result of the other popular detection methods.

Figure 6. Structure of the estimation network
The qualitative results of 3D detection for cars, pedestrians and cyclists are shown in figure 7.
As is illustrated in figure 7 (zoom in and view in colour mode would get better effect), the proposed camera and lidar fused 3D object recognition method can obtain 3D bounding boxes with much accurate positions and sizes.

4. Conclusion

A three stages camera and lidar fused 3D object recognition framework is proposed in this paper. First, two RPNs are constructed to generate 2D proposals from RGB image feature maps and BV feature maps separately, which are then fused to generate 3D proposals. Then, a segmentation network is built to extract object points directly from points lied inside these 3D proposals fused by the shared features.
from image and BV. At last, an estimation network is used to generate 3D object bounding boxes. Experiment are conducted on KITTI 3D object detection dataset, and the result is considerable to the other state of the art 3D object detection methods, although there is still a long way to go before real-life usage.

**Acknowledgments**

This work is supported by the China National key research and development plan key special project of intergovernmental international science and technology innovation cooperation development project “automotive passive and active collaborative protection technology for pedestrian safety” (No. 2018YFE0192900).

**References**

[1] Girshick R 2015 Fast r-cnn In *Proceedings of the IEEE international conference on computer vision* pp 1440–48.
[2] Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu C Y and Alexander C B 2016 SSD: Single shot multibox detector In *European conference on computer vision* pp 21–37.
[3] He K, Gkioxari G, Doll´ar P, and Girshick R 2017 Mask r-cnn In *Computer Vision (ICCV), 2017 IEEE International Conference on* pp 2980–88.
[4] Cho H, Seo Y W, Kumar B V and Rajkumar R R 2014 A multi-sensor fusion system for moving object detection and tracking in urban driving environments In *Robotics and Automation (ICRA), 2014 IEEE International Conference* pp 1836–43.
[5] Ku J, Mozifian M, Lee J, Harakeh A and Waslander S L 2017 Joint 3d proposal generation and object detection from view aggregation. CoRR, abs/1712.02294.
[6] Liang M, Yang B, Wang S L and Urtasun R 2018 Deep continuous fusion for multi-sensor 3d object detection In *Proceedings of the European Conference on Computer Vision (ECCV)* pp 641–56.
[7] Yang B, Luo W J and Urtasun R 2018 Pixor: Realtime 3d object detection from point clouds In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* pp 7652–60.
[8] Chen X Z, Ma H M, Wan J, Li B and 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 1907–15.
[9] Xu B and Chen Z Z 2018 Multi-level fusion based 3d object detection from monocular images In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* pp 2345–53.
[10] Song S and Xiao J X 2016 Deep sliding shapes for amodal 3d object detection in rgb-d images In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* pp 808–16.
[11] Zhou Y and Tuzel O 2017 Voxelnet: End-to-end learning for point cloud based 3d object detection CoRR, abs/1711.06396.
[12] Charles R Q, Su H, Mo K and Guibas L J 2017 Pointnet: Deep learning on point sets for 3d classification and segmentation *Proc. Computer Vision and Pattern Recognition (CVPR)* 1(2):4.
[13] Charles R Q, Yi L, Su H and Guibas L J 2017 Pointnet++: Deep hierarchical feature learning on point sets in a metric space In *Advances in Neural Information Processing Systems* pp 5099–5108.
[14] Charles R Q, Liu W, Wu C, Su H and Guibas L J 2017 Frustum pointnets for 3d object detection from RGB-D data. CoRR, abs/1711.08488.
[15] Ren S, He K, Girshick R and Sun J 2015 Faster r-cnn: Towards real-time object detection with region proposal networks In *Advances in neural information processing systems* pp 91–99.
[16] Lin T Y, Doll’ar P, Girshick R, He K, Hariharan B and Belongie S 2016 Feature pyramid networks for object detection arXiv preprint arXiv:1612.03144.