A New Object Detection Method in Indoor Scenes Based on Spatial Distance Clustering

Dianyuan Wu[1, a], Kai Xu[1, b]

[1]School of Computer Science, National University of Defense Technology, Changsha 410000, China.

[a]wudianyuan10@nudt.edu.cn, [b]kevin.kai.xu@gmail.com.com

Abstract. 3D object detection is one of the important problems in machine vision. With the increasing popularity of depth camera, 3d object detection in point cloud has become a research hotspot of 3D vision. The irregular format of point cloud makes the traditional deep learning method based on image convolution unable to understand and analyze it. In recent years, the deep learning framework as PointNet[1] directly applied to the original structure of point cloud has greatly improved the ability of deep learning network to process point cloud data. This paper depends on the existing point cloud deep learning framework, proposes an object detection method based on semantic segmentation. We use the clustering method based on region growth to detect the objects in indoor scenes. This method has been tested on the Stanford large-scale 3D Indoor Spaces Dataset (S3DIS) Dataset[2] with good results.

Keywords: Deep learning, point cloud, object detection.

1. Introduction

At present, most researchers transform point cloud to regular 3D voxel grids[3,4,5] or collections of images[6,7,8]. The common practice of semantic 3D point cloud modeling research is to directly learn from the research idea of semantic classification of two-dimensional images, that is, to conduct point-by-point feature classifier training based on 3D geometric features, and to obtain semantic classification results through spatial context constraint optimization.

PointNet[1] is the pioneering work which takes point cloud as direct input. The key to the approach is the use of a single symmetric function, max pooling. Effectively the network learns a set of optimization functions that select interesting or informative points of the point cloud and encode the reason for their selection. The final fully connected layers of the network aggregate these learnt optimal values into the global descriptor for the entire shape as mentioned above (shape classification) or are used to predict per point labels (shape segmentation). Inspired by PointNet, this paper presents a neural network structure semantic segmentation of point cloud in three-dimensional scene; Then the region growing is used to cluster the point clouds with the same semantic label at different scales.

2. Method

2.1 Problem Definition

Taken point cloud data as input, our goal is to classify and localize objects in 3D scenes. In the basic setting each point is represented by just its three coordinates (x, y, z). Additional dimensions may be added by computing normals, color channels, or other local or global features. Each object is represented by a 3D bounding box which is Axially-aligned parameterized by two points P_{min} (x_{min}, y_{min}, z_{min}) and Pmax (x_{max}, y_{max}, z_{max}). We use the Intersection over Union (IOU) between predicted bounding box and Ground-truth in volume as the measure for evaluation.

2.2 Pipeline and Backbone for Semantic Segmentation

Our pipeline was shown in Fig. 1. The point cloud can be viewed as an N*D matrix, where N represents the number of points and D represents the information dimension of each point. Due to unordered input, we use some existing deep architectures for 3D point cloud[10] such as PointNet as

the bases of our 3D semantic segmentation network. Then, we use connected component with segmentation scores to get object proposals in scenes. Starting from a random point in the scene, we find its predicted label and use regional growth algorithm to search nearby points with the same label under certain conditions iteratively.

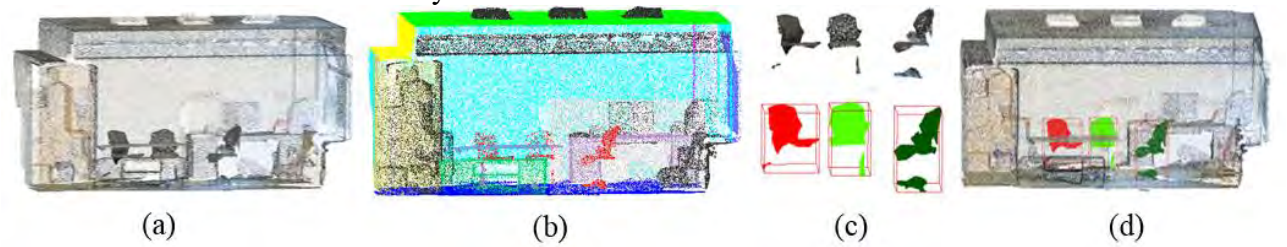

Fig. 1 The Pipeline of our method. (a) shows the original indoor scene. (b) shows the result of semantic segmentation. (c) shows the clustering of target objects. (d) shows the border box generated in the original scene.

2.3 Clustering Method Based on Euclidean Distance and Semantic Score

Our paper designs a clustering method based on Euclidean distance and semantic score in order to distinguish the different instance in the same category foundation. However, all clustering methods depend on the selection of suitable clustering conditions[9]. In our paper, the clustering algorithm adopts the method of regional growth. Two critical Parameters of clustering are spatial distance threshold and noise conditions. In Fig. 2, we show the flow chart of Clustering algorithm.

Input: point cloud U with the same semantic label, and which score is greater than θ ($\theta = 0.7$)

Output: a number of point cloud clusters with different instance segmentation label

Step 1: The seed point Q with a confidence greater than 0.9 was randomly selected to calculate the distance between the rest points and the seed point. Taking score into account, the formula to measure Euclidean distance in space is.

$$d = \frac{\sqrt{(x_i - x)^2 + (y_i - y)^2 + (z_i - z)^2}}{\theta} \tag{1}$$

Step 2: Select the points that metric distance d less than τ. And remove them from point cloud U and merge them into point cloud cluster k_i.

Step 3: Determines whether the point cloud cluster k_i is noise. If it is noise, it is discarded. Otherwise, judged as the detection object.

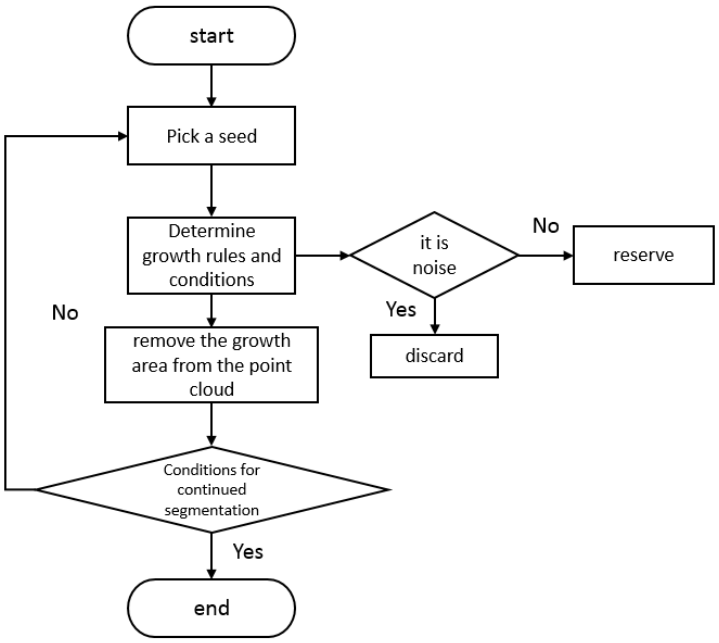

Fig. 2 The flow chart of clustering algorithm

3. Experiments

We experiment on the Stanford 3D semantic parsing dataset [2]. The dataset, which includes scan data for 271 rooms in six scenarios, covers buildings of more than 6,000 square meters and covers more than 215 million points, each of which specifies a segmentation category and an instance category.

3.1 Preprocessing of Data

In order to standardize the training data, the scene was divided into blocks of 1m*1m, and the scene was moved in accordance with the step size of 0.5m in the XY direction of the spatial coordinate system. The point cloud in each block was randomly resampled to 4096 points.

3.2 Results on Semantic Segmentation and Object Detection in Scenes

S3DIS data set are divided into 13 categories, including ceiling、floor、wall、beam、column、window、door、table、chair、sofa、bookcase、board、clutter. The Results on semantic segmentation in scenes are shown in Table 1. Metric is average precision over 13 classes and classification accuracy calculated on points. In Fig. 3, we show qualitative segmentation results. In Fig. 3, we show the precision-recall curves for object detection compared with PointNet.

Table 1 Results on semantic segmentation in scenes.

Ceiling	floor	wall	beam	column	window	door	table	chair	sofa	bookcase	board	clutter	avg class
0.92	0.99	0.89	0.76	0.64	0.92	0.90	0.82	0.76	0.21	0.69	0.53	0.72	0.752

We observe that in some rooms such as office lots of objects (e.g. bookcases) are close to each other, where connected component would fail to correctly segment out individual ones.

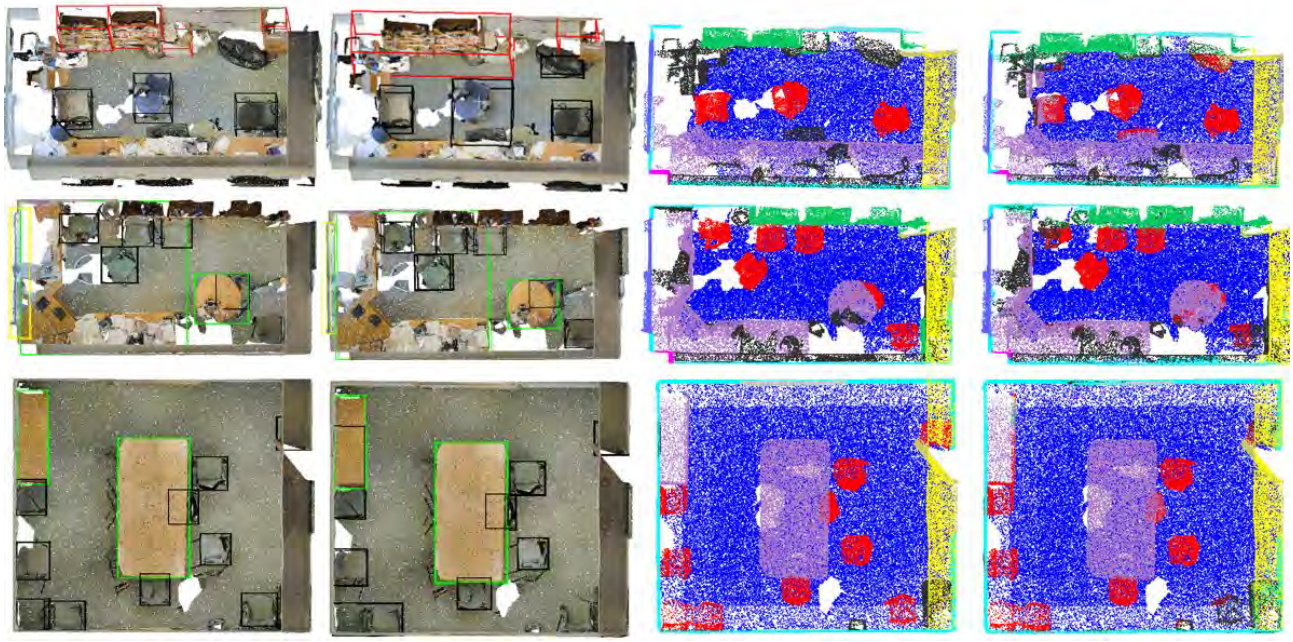

Fig. 3 Results

The first column is input point cloud with ground truth of box. The second column is ground truth for bounding box. The third column is the ground truth for semantic segmentation. The fourth columnis segmentation result (on points) displayed in the same camera viewpoint as input.

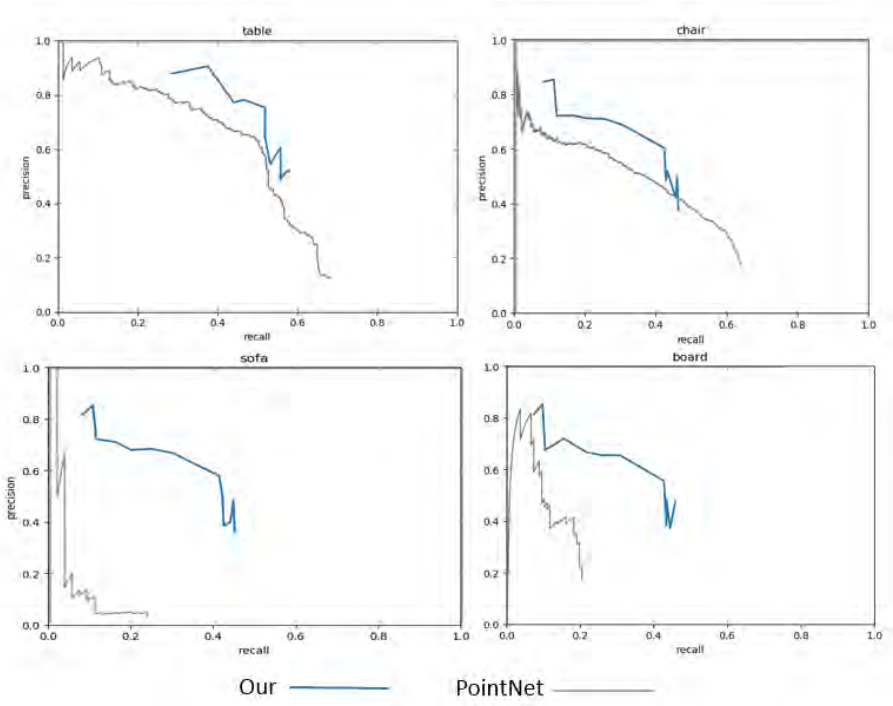

Fig. 4 Precision-recall curves for object detection in 3D point cloud.

The same with PointNet, we evaluated for four categories: table, chair, sofa and board. IoU threshold is 0.5 in volume.

4. Conclusion

We present a clustering algorithm for 3D object detection on point clouds. With the introduction of the bounding box as our output representation, group proposals with class predictions can be easily generated from the segmentation network. Experiments show that our algorithm can achieve good performance on 3D object detection and facilitate instance segmentation for various 3D scenes.

References

[1]. Charles R Q, Hao S, Mo K, et al. PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation. IEEE Conference on Computer Vision & Pattern Recognition 2017.

[2]. I. Armeni, O. Sener, A. R. Zamir, H. Jiang, I. Brilakis, M. Fischer, and S. Savarese. 3d semantic parsing of large-scale indoor spaces. In Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, 2016. 6, 7.

[3]. S. Song and J. Xiao. Sliding shapes for 3d object detection in depth images. In Computer Vision–ECCV 2014.2.

[4]. Zhirong Wu et al., "3D ShapeNets: A deep representation for volumetric shapes," 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 2015, pp. 1912-1920.

[5]. D. Maturana and S. Scherer. Voxnet: A 3d convolutional neural network for real-time object recognition. In IEEE/RSJ International Conference on Intelligent Robots and Systems, September 2015. 2, 5, 10, 11.

[6]. Charles R Q, Su H, Niebner M, et al. Volumetric and Multi-view CNNs for Object Classification on 3D Data. [IEEE 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) - Las Vegas, NV, USA, 2016.6.27-2016.6.30.

[7]. B. Li. 3d fully convolutional network for vehicle detection in point cloud. arXiv preprint arXiv:1611.08069, 2016. 2, 5, 6.

[8]. Joerg Liebelt C S. Multi-View Object Class Detection with a 3D Geometric Model. Computer Vision & Pattern Recognition. IEEE, 2010.

[9]. Ling H, Jacobs D W. Shape Classification Using the Inner-Distance[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2007, 29(2):286-299.

[10]. Qi, Charles R., et al. "PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space." arXiv preprint arXiv:1706.02413, 2017.