Deep Multi-level Feature Learning on Point Sets for 3D Object Recognition

Yang Xiao, Yanxin Ma, Qianlan Huang and Jun Zhang

ABSTRACT

In recent years, deep learning has become an important method on point cloud for 3D object recognition. PointNet is the first neural network which could directly consume point cloud as input. However, the PointNet couldn’t capture the local features. In this work, we introduce a multi-level feature extraction neural network which extracts the characteristics of the multi-level structure in PointNet. Experiments are conducted on the ModelNet40 dataset with several state-of-the-art methods. The proposed method achieves a higher accuracy on 3D object recognition with 89.4%. Experimental results have demonstrated the superior performance of the proposed multi-level feature learning network.

INTRODUCTION

In recent years, the study of 3D object recognition has important research significance and has always been a research topic in the computer vision and computer graphics areas. With the rapid development of 3D imaging technology, low-cost miniaturization of three-dimension sensors such as Kinect and Google Tango, have a large number of emerging and gradually equipped with mobile devices in recent years. Simultaneously, it has numerous applications such as human-machine interaction, automatic driving system, intelligent robots and city monitoring system. Because of the numerous tech development, it makes the development of deep learning based 3D object recognition algorithms possible. And
then, convolution neural networks (CNNs) have been increasingly applied for 3D object recognition systems.

Due to point clouds is not in a regular format, most researchers transform this data to regular 3D voxel grids or collections of images before sending them to a deep network architecture. The method of feature learning for 3D object recognition in depth learning can be roughly divided into three methods which including Multi-view based [1,2], volumetric representation based [3,4] and based on point cloud. The multi-view based method is to project the three-dimension shape into the two-dimension image space, and then use the method of depth learning to extract the two-dimension image. This method can make full use of the two-dimension image field superior performance of the network architecture, and the existence of massive image data for the depth of learning model for pre-training. However, this method couldn’t get more spatial structure information. Besides the multi-view method just provides a 2D contour representation of the 3D object and does not include sufficient geometric information as a 3D representation because some details information are not encoded. Therefore, 3D volumetric networks give a new direction. Voxel convolution neural network is a kind of neural network structure which considers the three-dimension shape as the probability distribution in the three-dimension voxel grid, and thus expresses it as a two-dimension tensor or two-dimension tensor. The groundbreaking work in this area was made by Wu et al. [3] In 2015. They proposed a convolution deep belief network for 3D object recognition and builded the 3D ModelNet dataset. Subsequently, the number of 3D volumetric convolution neural networks have been proposed to apply to 3D shapes. Qi et al. [4] proposed a volumetric 3D CNN by subvolume supervision to address overfitting. Maturana and Scherer [5] designed a convolution neural network for real-time recognition of 3D object recognition volumetric convolution neural network which named VoxNet [6]. Soon after, Qi et al. proposed PointNet [7] to solve the previous three-dimension voxel cannot directly deal with unstructured point cloud data problems which can directly deal with the depth of point cloud data learning network structure and efficiently complete the three-dimension point of the transport office identification and segmentation tasks. Although these works have been got a great achievement, most of existing deep learning models’ cost are high and their frameworks are complex. In this paper, we study the invariance and completeness of 3D point cloud feature extraction in depth learning, and start it based on convolution neural network in depth learning technology and then apply it to 3D object recognition.

We propose a multi-level feature extraction convolution neural network (namely, MFECNN) based on PointNet by extracting multi-level structure to get more local information and carry out feature learning to improve the category object recognition accuracy and efficiency.

The main contributions of our work are as follows:

(i) We design a deep CNN suitable for inputing unordered point to 3D object recognition. It utilizes multi-level information to get great improvements as
compared to some existing models including Spherical Harmonic descriptor (SPH) [8], 3DshapeNets [3], VoxNet [5], subvolume [4], Light Field descriptor (LFD) [7] and PointNet [8].

(ii) We show how to extract more multi-level structure’s local information to improve the recognition rate.

(iii) Comparative experiments have been conducted on the ModelNet40 dataset [3]. The experimental results show that the proposed model provides a basic structure for 3D object recognition tasks with disorder point clouds processing.

This paper is structured as follows. Section 2 gives a literature review of 3D object recognition methods, especially the CNN based method. Section 3 presents our model and introduces the method to utilize multi-level information and the implementation details. Section 4 the recognition experiments conducted on ModelNet40 dataset. Finally, Section 5 concludes this paper.

RELATED WORK

The core of 3D object recognition is to extract the 3D shape features of discernibility, simplicity and low dimension. The classical approach is designed to function according to specific tasks and domain knowledge. The main purpose of these methods is to design 3D shape features with good discriminating ability, robustness, invariance and computability by extracting the spatial distribution of geometric attributes of 3D object space or using statistical histogram. On the contrary, deep learning through training to automatically obtain data characteristics, to avoid human intervention [9].

After the many of convolution neural networks have been used to process 2D images and the CNN begin to be applied to 3D object recognition. Shi et al. [10] proposed a convolution neural network (namely, DeepPano) [11] for 3D object recognition. It was based on the panoramic images of objects. Considering that although neural networks have a certain robustness to translation, the 2D projection will change when the object is rotated, which will have significant influence on the features extracted by convolution neural network. To overcome this negative effect, Shi et al. first project each 3D object into a panoramic image around the principal axis. The panoramic image is extracted by convolution neural network and the object is classified. The experimental results show that this method can preserve the shape information of 3D objects to a certain extent through transformation, but the transformation process itself changes the local and global structures of 3D shapes, resulting in the decrease of feature discrimination. Meanwhile, Su et al. [2] proposed multi-view convolution network structure (Multi-View CNN, MVCNN) [13]. These authors use the multi-view 2D projection of the 3D object to extract a concise 3D feature descriptor for the classification and retrieval of 3D shapes. In this paper, first of all, the projection of 3D shape under twelve different viewpoints is obtained, and then the characteristics of projection images under various viewpoints are studied by
using VGG-M convolution neural network. Finally, the multi-view features are pooled and sent to the next CNN network to get the final shape features. The experimental results show that multi-view images can get better performance than single-viewpoint images and has state-of-the-art performance in 3D object recognition. Johns et al. [12] proposed a convolution neural network using multi-view 2D images without camera trajectory. In this method, the input images are combined in pairs and put into the convolution neural network along with their relative poses. This method combines a gray-scale image, a depth image, or both as an input to a network and paired. The input image is processed by two convolution neural networks and cascaded together before being input to the fully connected layer. The result of this method based on ModelNet40 is better than 3DShapNets [3] and MVCNN [13].

MODEL

In this section, our model will be introduced. Point clouds are a collection of massive points at the surface characteristics of the object. It can be used to keep a more complete structural information when we deal with 3D object recognition and scene segmentation tasks. For most of 3D CNNs, the input of the model should be voxels, occupancy grids. Besides, 3D data represented by mesh or point clouds must be converted to volumetric data for most of CNNs before being recognized.

Specially, 3D point cloud data is expressed in the form of 3D coordinates, which contains a direct vector of spatial information. Due to the existence of massive point cloud data for an object, all cloud point data needs to be prevented from exploding when dealing with point cloud data, and its processing efficiency has become a key factor for judging the network's goodness. Simultaneously, before inputting this data into the network training, point cloud need to be converted to the regularly input form for most of CNNs. In our CNN, point cloud is directly as input and then put it into our network (Fig.1).

![Figure 1. The framework of the proposed multi-level CNN which extract more local feature information.](image-url)
In this network, the input data is disorderly point cloud: $B \times N \times 3$. The T-Net (3) is the key to solve this problem. It uses a symmetric function to align point cloud data. After that, the multilayer Perceptron (MLP) achieves shared parameters and the part of 1,2,3 layers of labels achieve the feature extraction and these features are sent to the max pooling layers. And then, the features of three layers are connected by concat function. At last, the feature data form is $B \times N \times 1 \times 1152$ and it is sent to two fully connected layers train. The data output can be calculated as Eq. 1.

$$\text{outputsize} = \frac{n - f + p}{\text{stride}}, \quad p = \text{valid}$$  \hspace{3cm} (1)

### EXPERIMENTS AND RESULTS

Our MFECNN model was tested in ModelNet datasets. And the datasets are most of popular 3D object recognition dataset. Meanwhile, Princeton ModelNet is a large 3D repository of CAD models (shapes) without noise and the commonly used datasets are ModelNet10 and ModelNet40 which consist of ten and forty categories respectively. The ModelNet10 dataset contains 4,899 CAD models and the ModelNet40 dataset contains 12,311 CAD models. In this paper, our model was tested in ModelNet40 and our model achieves good performance in spite of the number of various orientations.

Our network achieves 3D object recognition through multi-level relationships. Our model was evaluated on the ModelNet40 shape classification benchmark. And the ModelNet40 consists of 9,843 CAD models for training and 2,468 for testing. Our model was trained on ModelNet40 from scratch and the experimental results show a higher recognition accuracy. And then the results of our model was compared with the state-of-the-art 3D object recognition models on the ModelNet40 dataset, as shown in Table I. As we can see, the two indicators: accuracy average class and accuracy overall.

<table>
<thead>
<tr>
<th>Method</th>
<th>Input</th>
<th>Accuracy average class (%)</th>
<th>Accuracy overall (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SPH</td>
<td>mesh</td>
<td>68.2</td>
<td>-</td>
</tr>
<tr>
<td>3DShapeNets</td>
<td>volume</td>
<td>77.3</td>
<td>84.7</td>
</tr>
<tr>
<td>VoxNet</td>
<td>volume</td>
<td>83.0</td>
<td>85.9</td>
</tr>
<tr>
<td>Subvolume</td>
<td>volume</td>
<td>86.0</td>
<td>89.2</td>
</tr>
<tr>
<td>LFD</td>
<td>image</td>
<td>75.5</td>
<td>-</td>
</tr>
<tr>
<td>MVCNN</td>
<td>image</td>
<td>90.1</td>
<td>-</td>
</tr>
<tr>
<td>PointNet</td>
<td>point</td>
<td>86.2</td>
<td>89.2</td>
</tr>
<tr>
<td>PointNet++</td>
<td>point</td>
<td>-</td>
<td>91.9</td>
</tr>
<tr>
<td>ours</td>
<td>point</td>
<td>87.5</td>
<td>89.4</td>
</tr>
</tbody>
</table>
It can be observed from Table I that our model’s accuracy achieves good results and the method obtain a compelling accuracy of 90.1%. The results also shows that our model is able to achieve 3D object recognition which improves the accuracy compare to PointNet. Compared to PointNet++ [14] model (with an accuracy of 91.9%), their model achieves state-of-the-art results. This is because PointNet++ added hierarchical structure and build a hierarchical grouping of points and progressively abstract larger and larger local regions along the hierarchy. Simultaneously, we observed in our experiments that our model has a lower calculation time and the efficiency has a good performance.

**CONCLUSION**

In this paper, we proposed a multi-level CNN for 3D object recognition. The proposed network input paid attention to 3D point cloud which directly consume point cloud and could provide more 3D information. Simultaneously, our multi-level CNN also gets better result on point cloud for 3D object recognition. Experimental results show that the proposed network achieves a comparable recognition accuracy with several simple operations and may get good results which applies to semantic segmentation.

**REFERENCES**


