DISCRIMINATIVE LEARNING OF POINT CLOUD FEATURE DESCRIPTORS BASED ON SIAMESE NETWORK

Xuelun Shen[1]*, Cheng Wang*[1*]*, Chenglu Wen*[1]*, Weiquan Liu*[1]*, Xiaotian Sun*[1]*, Jonathan Li*[1,2]

[1]Fujian Key Laboratory of Sensing and Computing for Smart City, School of Information Science and Engineering, Xiamen University, Xiamen, Fujian361005, China
[2]GeoSTARS Lab, Department of Geography and Environmental Management, University of Waterloo, Waterloo, Ontario N2L 3G1, Canada
*Corresponding author: cwang@xmu.edu.cn

ABSTRACT

It is challenging to direct extract the feature descriptors of the object in the point cloud, although deep learning has been widely used with the classification and detection in the point cloud, those methods hidden feature presentation in the network. Since the point cloud scanned by the Laser Scanner usually have different point density, unordered and even the different occlusion, which go beyond the reach of hand-crafted descriptors, e.g. FPH, FPFH, VFH, ROPS. In this paper, we aim to direct extract the feature descriptors of the point cloud object through the raw point cloud. Inspired by the recent success of the Siamese networks[6], PointNet[7] and PointNet++[8], we propose a novel network to direct extract the feature descriptors of the whole point cloud object. We train our network with the Euclidean distance as the loss function which reflects feature descriptors similarity. The experiment object datasets were acquired by Mobile Laser Scanning (MLS) system which contains 6 categories. Experiment result shows that our network has a robust generalization, which can well direct extract the feature descriptors of the whole point cloud object.

Index Terms— Point cloud, feature description, mobile laser scanning, siamese network

1. INTRODUCTION

Mobile Laser Scanning (MLS) systems have been commonly used in many applications such as smart cities and intelligent transportation, because of a MLS system is flexible and can rapidly acquire high dense and accurate 3D point clouds with geometry, color and intensity information from object surfaces. In practical applications, such as intelligent transportation, the classification of point cloud objects is very important[1], so the feature descriptors of the objects are very necessary.

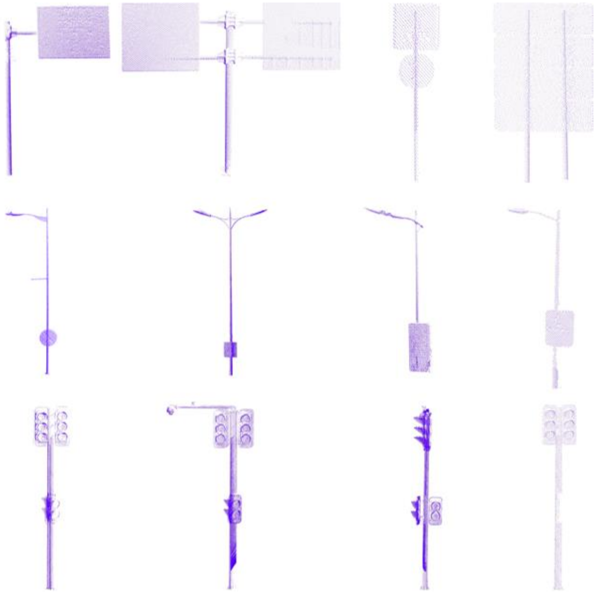

Figure 1. These objects in the same row are the same category but have different shapes. The top row is traffic sign, the middle row is street light, and the bottom is the traffic light.

For the feature description in the point cloud, almost of the traditional algorithms are based on hand-crafted local feature description, like PFH[2], FPFH[3], VFH[4], ROPS[5], and their variants, cannot extract the feature description with the whole point cloud object. For the MLS point cloud data, the point density of the same type object may be different, usually, there will be different occlusions, even their shapes are different, as shown in Figure 1. So, these are the big challenge for the traditional algorithms, where they are unable to extract the feature descriptors of the entire point cloud object.

In this paper, with the deep learning make dramatic progress in point cloud processing, and inspired by the Siamese networks[6], PointNet[7] and PointNet++[8], we propose a new framework based on the Siamese Network to direct extract the feature description from the raw point cloud with the entire objects. The Siamese network is a neural network with two branches network, the inputs of which are pairs of data. And the architecture in the two branches can be the same or different. The weights in the two branches could be shared or not shared, in this paper, we share the weight in the Siamese network. PointNet was proposed by Qi et al[7].

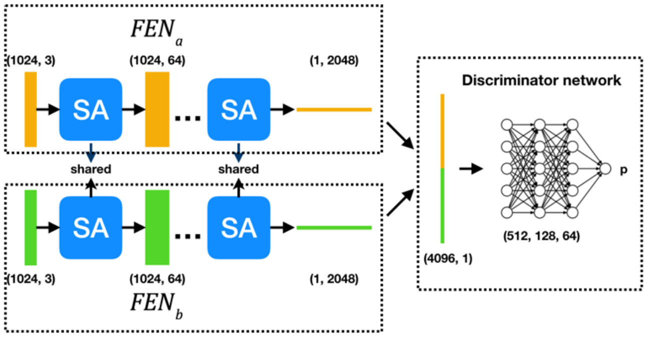

Figure 2. The network framework. Feature Encode Network feeds in the unordered point cloud and outputs feature vector. Discriminator network feeds in two feature vector and outputs a probability that two point clouds are in the same class. In the Feature Encode Network, the SA stands for the set abstraction layer whose detail structure is shown in Figure 3.

which could input unordered point cloud and output a feature vector for this point cloud. It uses the max-pooling operator eliminate the inference of the order of point clouds. In the improved version PointNet++[8], it proposed a hierarchical neural network. It applies PointNet recursively on a nested partitioning of the input point set, which enables it to learn local features with increasing contextual scales.

Furthermore, we use the Euclidean distance as the loss function of our network, for which it is more intuitive to measure the feature descriptors extracted with the whole point cloud object in a unified metric space. In addition, we collected 2,000 MLS point cloud object which contains bus station, fire hydrant, street light, traffic light, traffic sign, and trashcan. By the different occlusion and point density, the shape of the point cloud objects are different. So, we constructed 50,000 pairs of point cloud objects as the training data and testing data.

In summary, the main contribution of this paper is that we construct a novel network to discriminate learning the feature descriptors from the raw point cloud, and directly extract the feature descriptors with the whole point cloud object, rather than local features. We use a deeper network to extract features and experiment result show that our network has a robust generalization, which can well direct extract the feature descriptors of the whole point cloud object.

2. OVERVIEW

Our network has two branches, as shown in Figure 2, each one is Feature Encode Network (FEN) which to extract the feature of the point cloud object. The parameters are shared between FEN_a and FEN_b. FEN feeds in unordered point cloud and outputs correspond feature descriptor. In order to train our network to learn the meaningful descriptor, we add another network to restrict the previous network which called discriminator network. The input of the discriminator network is two feature descriptors from the previous network and its output is the probability p of inputs are same class.

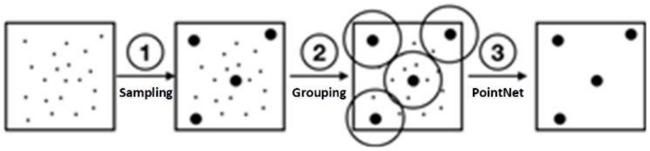

Figure 3. The Set Abstraction layer structure. After sampling and grouping, the point cloud gets a sparse point with a high-dimensional feature representation. The first operation is Sampling, the second operation is Grouping and the third part operation is done in the PointNet.

After our network trained, the FEN will be exactly the feature extractor we want.

3. NETWORK STRUCTURE

Feature Encode Network (FEN) consists of hierarchical Set Abstraction (SA) layer which designed in PointNet++[8], as shown in Figure 2.

One Set Abstraction layer consists of sampling layer, grouping layer, and density adaptive PointNet layer.

3.1 Sampling layer

Sampling layer selects a set of points from a point cloud which defines the centroids of point cloud local regions. Given inputs points $\{x_1, x_2, \dots, x_m\}$, use iterative farthest point sampling to choose a subset of points $\{x_{i_1}, x_{i_2}, \dots, x_{i_m}\}$, such that x_{i_j} is the most distant point from set $\{x_{i_1}, x_{i_2}, \dots, x_{i_{j-1}}\}$ with regard to the rest points. This layer will sample N' centroids from the input N points, as the first operation shown in Figure 3.

3.2 Grouping layer

Grouping layer feeds in a point set of size $\mathrm{N} \times (\mathrm{d} + \mathrm{C})$ and a set of centroids of size $\mathrm{N}' \times \mathrm{d}$. the output is groups of point sets of size $\mathrm{N}' \times \mathrm{K} \times (\mathrm{d} + \mathrm{C})$, where each group corresponds to a local region with K points in the neighborhood of centroid points, as the second operation shown in Figure 3.

3.3 Density adaptive PointNet

Density adaptive PointNet layer is inputted N' a local region of points with size $\mathrm{N}' \times \mathrm{K} \times (\mathrm{d} + \mathrm{C})$. It abstracts high dimension feature vector with a size of $\mathrm{N}' \times (\mathrm{d} + \mathrm{C}')$. In another word, these K points are extracted as a feature vector, as the third part shown in Figure 3.

The origin PointNet is given an unordered point set $\{x_1, x_2, \dots, x_n\}$ with $x_i \in \mathbb{R}^d$, define a set function $\mathrm{f}: \mathrm{X} \rightarrow \mathbb{R}$ that maps a set of points to a vector:

$$f(x_1, x_2, \dots, x_n) = \gamma\left(\max_{i=1,\dots,n} \{h(x_i)\}\right) \qquad (1)$$

Where γ and h are implemented by multi-layer perceptron (MLP) network, as shown in Figure 4.

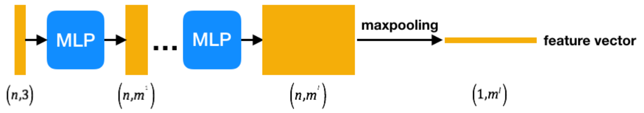

Figure 4. The PointNet structure. PointNet feeds in the unordered point cloud and outputs a feature vector for this point cloud. The max-pooling operator will eliminate the inference of the order of point clouds.

The set function f in Eq.1 is invariant to input point permutations and can arbitrarily approximate any continuous set function.

In implement, the PointNet layer use a density adaptive strategy, it outputs feature of a region at some level L_i is a concatenation of two features. One feature is summarizing the features from the lower level L_{i-1} using the set abstraction layer. The other feature is the feature directly extract from the input $\mathrm{N}' \times \mathrm{K} \times (\mathrm{d}+\mathrm{C})$ point cloud using a single PointNet.

3.4 Discriminator network

Discriminator network feeds in two features descriptor from two branches. We concatenate the two feature descriptor and put them into 3-layer Multi-Layer Perceptron (MLP) network with drop-out and batch-normalization, as shown in Figure 2. For the non-linear activate function in the MLP, different with most articles all use the Relu[9] or Tanh[10], we adopt Relu or the first 2 layers and Tanh for the last layer. The output of the discriminator network is a scalar which is the probability of inputs are the same class.

4. LOSS FUNCTION

The loss function consists of two parts. One part is hinge loss which optimizes the prediction accuracy. The other part is l2 regularization term to avoid over-fitting. The whole loss function could be a formula as:

$$\min_{\omega} \frac{1}{2} \|\omega\|_2 + \sum_{i=1}^{N} \max(0, 1 - y_i \cdot O_i) \qquad (2)$$

Where ω is the weight of the network. O_i is the output of the network for the i_{th} pair of point cloud, and $\mathrm{y}_i \in \{-1, 1\}$ is the label of the training data. When two point clouds class are same, $\mathrm{y}_i = 1$, otherwise $\mathrm{y}_i = -1$.

At the beginning, we tried using the $\mathcal{L}_2$ Euclidean Distance as the loss function after the features is extracted from the two branches without discriminator network. The $\mathcal{L}_2$ Euclidean Distance is more intuitive for the metric of the two descriptors, which minimize $\mathcal{L}_2$ Euclidean Distance of same class point cloud and increase the $\mathcal{L}_2$ Euclidean Distance of the different class point cloud. However, this loss function leads the network to degradation, as the network will not learn any meaningful features with the train data, the feature description is all zero. So, we think we need another network to constrain the features extracted in previous. The auto-encoder network is the best choice, but reconstruct unordered point cloud from a simple feature vector is too difficult. So, we chose a discriminator network instead of doing it.

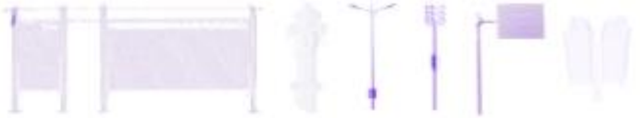

Figure 5. The 6 categories point cloud data in our dataset, from left to right are bus station, fire hydrant, street light, traffic light, traffic sign, and trashcan respectively.

Test	Train	Real-valued descriptors	Binary descriptors
Own Dataset	Own Dataset	87.18%	89.23%
	ModelNet40	83.70%	85.75%

Table 1. The accuracy of the real-valued descriptors searched and binary descriptors on the testing dataset.

5. EXPERIMENT

For all experiments, we use the Adaptive Moment Estimation (Adam) Optimizer with a learning rate of 0.001 for training. We use TensorFlow and one Titan X for training. The training process stops when the loss function converges. All the layers are implemented in CUDA to run in GPU.

Our proposed network to direct extract the point cloud object feature descriptor was tested on the objects with both sides of city road. The experiment testing datasets were acquired by a RIGEL VMX-450 system. The testing dataset contains 6 categories that are bus station, fire hydrant, street light, traffic light, traffic sign, and trashcan respectively, as shown in Figure 5.

There are two training datasets for this network. One training dataset is same as testing data from VMX-450, but they do not intersect. Another training dataset is ModelNet40[11] benchmark. There are 12,311 CAD models from 40 man-made object categories. But these 40 categories do not contain these 6 categories in our testing data. Even so, with the training dataset is ModelNet40, we can see that our network has a good performance on the testing datasets that it has not seen, as shown in Table 1.

In detail, we conducted two kinds of the test program. One is the real-valued descriptors search test. We randomly grab 10 target data and then grab 1 source data to be searched, which has the same class to one of the 10 target data but not the same object. Letting the network directly extract the features of these 11 data examples and then search for the Euclidean distance, if the closest target data and source data are the same category, it is regarded as correct, otherwise is wrong, and we call this search accuracy. Another is just inputted a pair point cloud objects data from testing dataset into the network which the output with the binary descriptors. The accuracy of the testing data is shown in Table 1.

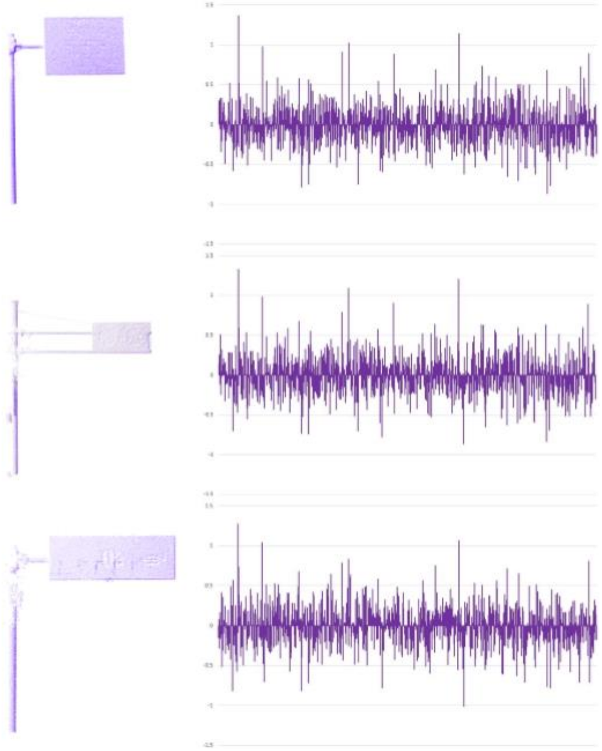

Figure 6. The feature descriptors histogram of the traffic sign.

As shown in Table 1, our network trained in the ModelNet40 and tested in the point cloud objects from VMX-450 get a good performance, this proved that our network has a robust generalization. In addition, the accuracy of the binary descriptors outperform the real-valued descriptors, there are two reasons for this. First, binary descriptors are more representative. Second, the discriminator network combines the two feature descriptor to consider jointly, then they have a better chance to be complementary to each other.

Finally, in order to prove that the feature descriptors of the same type of point cloud objects directly extracted by our trained network are consistent, we visualize the feature descriptors of the same category point cloud objects. We use the traffic signs as an example to visualize, the histogram of the feature descriptors was shown in Figure 6. It can be seen that the histogram distribution trends of the traffic signs feature descriptors are broadly consistent, and the changes in the salient features are also consistent. These prove that our network learns the common attributes of the same kind of point cloud objects.

6. CONCLUSION

We use Siamese network to train deep network for the extraction of point cloud feature description. Training such models involve unordered point cloud which constraints the network structure and discriminative power.

In this paper, we introduce a novel training scheme, based on discriminator network, and mining of both positive and negative correspondences enables the network to extract meaningful feature descriptions. Our network generalizes well across different datasets, even in the training sets and testing sets have different data types also get a good performance. They could be used in many fields such as point cloud classification, registration, and matching. In the following work, we will further study how to better match and build a contextual relationship with the same kind of point cloud objects feature descriptors.

ACKNOWLEDGEMENTS

This work described here is supported by grants form Natural Science Foundation of China (Grant No.61371144 and Grant No.U1605254).

REFERENCES

[1] Yang, Bisheng, et al. "Hierarchical extraction of urban objects from mobile laser scanning data." ISPRS Journal of Photogrammetry and Remote Sensing 99 (2015): 45-57.

[2] Rusu, Radu Bogdan, et al. "Persistent point feature histograms for 3D point clouds." Proc 10th Int Conf Intel Autonomous Syst (IAS-10), Baden-Baden, Germany. 2008.

[3] Rusu, Radu Bogdan, Nico Blodow, and Michael Beetz. "Fast point feature histograms (FPFH) for 3D registration." Robotics and Automation, 2009. ICRA'09. IEEE International Conference on. IEEE, 2009.

[4] Rusu, Radu Bogdan, et al. "Fast 3d recognition and pose using the viewpoint feature histogram." Intelligent Robots and Systems (IROS), 2010 IEEE/RSJ International Conference on. IEEE, 2010.

[5] Guo, Yulan, et al. "RoPS: A local feature descriptor for 3D rigid objects based on rotational projection statistics." Communications, Signal Processing, and their Applications (ICCSPA), 2013 1st International Conference on. IEEE, 2013.

[6] Tian, Yurun, Bin Fan, and Fuchao Wu. "L2-Net: Deep learning of discriminative patch descriptor in euclidean space." Conference on Computer Vision and Pattern Recognition (CVPR). Vol. 2. 2017.

[7] Qi C R, Su H, Mo K, et al. "Pointnet: Deep learning on point sets for 3d classification and segmentation." Conference on Computer Vision and Pattern Recognition (CVPR). 2017.

[8] Qi C R, Yi L, Su H, et al. "PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space Supplementary Material." Conference and Workshop on Neural Information Processing Systems (NIPS). 2017.

[9] Zagoruyko, Sergey, and Nikos Komodakis. "Learning to compare image patches via convolutional neural networks." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2015.

[10] Simo-Serra, Edgar, et al. "Discriminative learning of deep convolutional feature point descriptors." Proceedings of the IEEE International Conference on Computer Vision (CVPR). 2015.

[11] Wu, Zhirong, et al. "3D Shapenets: A Deep Representation for Volumetric Shapes." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2015.