ATLANTIS PRESS

*Advances in Computer Science Research, volume 79*

2nd International Conference on Sensor Network and Computer Engineering (ICSNCE 2018)

# 3D Target Recognition Based on Decision Layer Fusion

Ma xing[a*], Yu Fan[b], Yu Haige, Wei Yanxi, Yang Wenhui

School of computer science and engineering

Xi'an Techonolgical University

Xi'an, 710021, Shaanxi

e-mail: [a*]512066020@qq.com; [b]yffshun@163.com

***Abstract*—Target recognition has always been a hot research topic in computer image and pattern recognition. This paper proposes a target recognition method based on decision layer fusion. ModelNet[1]—The 3D CAD model library，which is used to be identified. Features are extracted from the model's point cloud data and multi-view images. The image is identified using the AlexNet[2] network, the point cloud is identified by the VoxNet[3] network. The fusion algorithm is used in the decision layer to complete the fusion of features. The results show that the proposed method improves the accuracy of object recognition.**

***Keywords-Target Recognition; Convolutional Neural Network; Decision Fusion***

## I. INTRODUCTION

At present, the methods for identifying objects are mainly divided into two categories. The first is to identify the images generated by the objects, and the second is to identify the point clouds generated by the objects.

In terms of image recognition, the current deep learning method has a high recognition rate. For instance, Xie etal. [4] adopt the multi-view depth image representation and propose multi-view deepextreme learning machine (MVD-ELM) to achieve fast and quality projective feature learning for 3D shapes. Zhu et al.[5] alsoproject 3D shapes into 2D space and use autoencoder for featurelearning on 2D images. In 2012, the ImageNet contest champion model--Alexnet became a classical convolutional neural network image classification model.

For the identification of point clouds, domestic and foreign scholars have done a lot of research. Some scholars use the method of manually extracting features to classify and identify. R.B.Rusu[6]and others used the relationship between the normal vectors of a region as a feature to classify objects for recognition. Yasir Salihet al. [7] used VFH as a feature and used a support vector machine as a classifier to classify and recognize point clouds. Manually extracting features requires very professional knowledge and rich experience. Convolutional neural networks can automatically extract features and classify them, and they are invariant to displacements, scaling, and other forms of rigid body changes. Some experts and scholars have used convolutional neural networks to classify and recognize point cloud images, of which the VoxNet network has the highest recognition rate.

The above method achieves higher accuracy in target classification recognition. I have seen that image recognition is affected by factors such as lighting and viewing angles. The accuracy of point cloud recognition is lower than that of image recognition. Therefore, this paper integrates the image and the point cloud at the decision-making level to improve the accuracy of object recognition.

## II. NETWORK STRUCTURE

### *A. Convolutional neural network*

Traditional shallow learning methods such as support vector machines require manually extracting image features and then sending the features into the classifier for training. This leads to a problem that the manually extracted feature is not necessarily the best description of the current image. Even if the selected feature is very suitable for the current image, when the external conditions of the object such as the angle, size, and illumination of the image change, the manually selected feature cannot adapt well to this change, and it is necessary to artificially adjust the selected feature according to the situation. Different from the traditional shallow learning method, the input of the convolutional neural network is the entire image. It continuously adjusts the parameters of the network through a learning algorithm and adaptively extracts the most significant features of the current image, which avoids manual intervention and saves a large amount of manpower, with the continuous updating of the input pictures, the essential features of the current picture are extracted with the times, ensuring the accuracy and efficiency of the recognition. As a special architecture that is particularly suitable for classifying images, compared with conventional shallow machine learning methods such as support vector machines, convolutional neural networks can be much smaller than normal when faced with large-scale high-resolution image classification problems. The method learns more picture information in the training time, and the classification accuracy is higher than the conventional method, which is due to its unique network architecture.

The convolutional neural network consists of three parts: the convolutional layer, the down-sampling layer and the fully connected layer. The down-sampling is usually after the convolutional layer, alternating with the convolutional layer, and finally connected to the fully connected layer. Convolutional neural networks use local connections, weight sharing, and spatial or temporal correlation down-sampling to obtain good translation, scaling, and distortion invariance, making the extracted features more distinguishable. CNNs

training includes forward propagation. The two processes of forward propagation and reverse propagation are the process of the input signal output from the input layer, through several hidden layers, and the output layer. The reverse propagation is the process of back propagation of the error signal from the output layer to the input layer. It mainly uses errors. The back propagation (Error Back Propagation, EBP) algorithm and gradient descent adjust the weights at each level of the network and is similar to the training process of an ordinary neural network.

### *B. AlexNet network structure*

The standard convolutional neural network(CNN) is a special multilayer feed forward neural network. It has a deep network structure and is generally composed of an input layer, a convolutional layer, a down-sampling layer, a fully connected layer, and an output layer. The convolutional layer, the down-sampling layer, and the fully connected layer are hidden layers. AlexNet uses two GPU services. The model is divided into eight layers, five convolutional layers and three fully connected layers. Each convolutional layer contains the excitation function RELU and local response normalization (LRN) processing, and then after down-sampling (POOL). The network structure designed in this paper is shown in Figure 1.

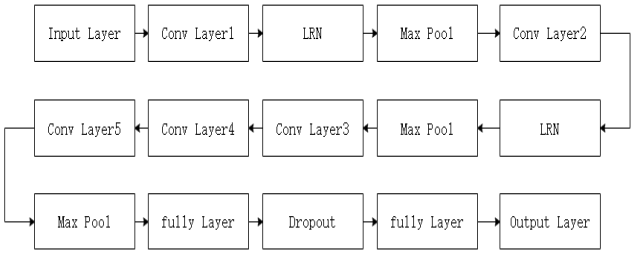


Figure 1. Multi-view image network structure diagram

The input layer are images, there are 5 convolutional layers. The number of feature maps are 55, 27, 31, 13,and 6 features respectively. The convolution kernel sizes respectively are 11, 5, 3, 3, and 3,Below the first two convolutional layers there is a largest pooled layer with an LRN layer for localized normalization. The largest pooled layer is to take the maximum value of the feature points in the domain. After processing through the convolution layer, many features are obtained. The amount of direct calculation is very large, and the increase of features is particularly prone to overfitting. Therefore, the network constructed in this paper is Each time a convolution process is performed, a Max-Pooling layer is added. The Dropout layer has a discard rate of 0.5. Dropout temporarily discards some networks in the training process with a certain probability, and each mini-batch discards different networks. It can reduce the amount of calculation, and more importantly it can prevent over-fitting. The number of neurons in two fully connected layers is 256 and 10, respectively. The last output layer is the Softmax[6]layer, which does not directly output the classification of the identified image, but rather the probability that the output image belongs to each classification.

### *C. VoxNet Network Structure*

Figure 2 shows the network structure of VoxNet. The input layer accepts data as a form. There are a total of two convolutional layers, and the number of feature maps is 32, using the sum of the convolution kernels. The Dropout layer discard rates are 0.2 and 0.3, respectively, to prevent overfitting while reducing the amount of computation. The largest pooled layer, the filter used. Finally, there is a fully connected layer with 128 neurons and a Dropout layer with a discard rate of 0.4. The seventh layer is the output layer and the number of neurons is 10.

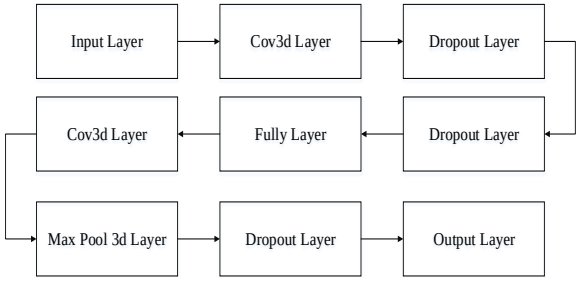


Figure 2. VoxNet network structure diagram

### *D. Network convergence structure*

Multi sensor information fusion is to extract and integrate the same target image with a multi-source information channel for further processing. Information fusion can be divided into three layers: the fusion based on data layer, the fusion based on the feature layer, and the fusion based on the decision layer. The level of fusion was from low to high. So this paper uses the method of decision layer fusion.

The feature fusion of the decision layer is usually the fusion of the prediction results of multiple classifiers. We extract various features from feature extraction algorithm. We assume that all kinds of features are independent of each other, and these features can separately predict the result of recognition separately. On this basis, we send data into their respective classifiers, get the prediction results of each classifier, then combine all the classifier's prediction to get the final recognition results, and complete the fusion of multiple features in the decision level.

As shown in Figure 3, point cloud is extracted from VoxNet using point cloud feature, and AlexNet is used to extract image features. Softmax regression model is used to complete recognition and classification respectively. The fusion algorithm is used in the decision layer to complete the fusion of features.

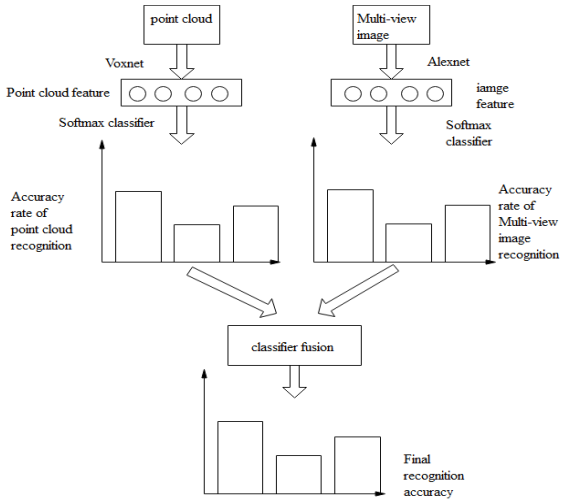


Figure 3. Fusion model

## III. . EXPERIMENTS

### A. Experiment environement

The environment used in this experiment was TensorFlow-GPU 0.12.1 open source software library, Windows 7 operating system, and Nvidia GTX 950 graphics card. The data used in the experiment was ModelNet of Princeton University. ModelNet is a large-scale 3D CAD model database, similar to ImageNet in 2D images.

### B. Experiment Datasets

This article uses the ModelNet40 dataset, where the point cloud dataset is from the dataset in PointNet[12], and the 2D image is from Multi-view. On the topside of Figure 4 is the point cloud image, at the bottom is a two-dimensional image.

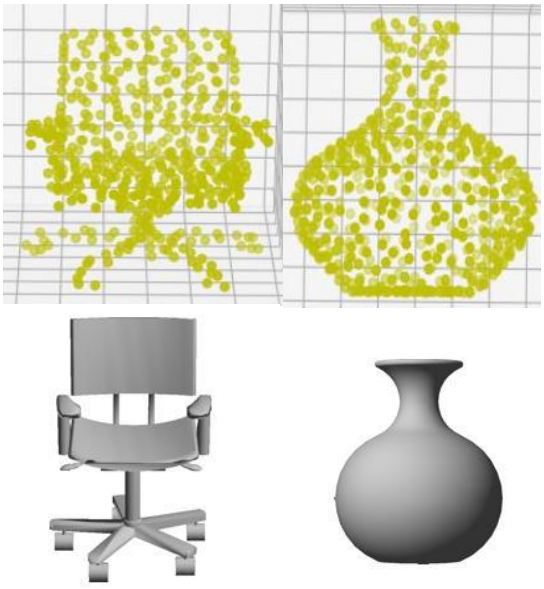

Figure 4. Point cloud image and two-dimensional image

### C. Linear Combination Coefficient Selection

Static linear combination of the results of the AlexNet network and VoxNet network prediction results in the final prediction. Before linear combination, we need to determine the coefficients of each classifier's prediction result, which controls the relative importance of the results predicted by each classifier. It is very important to choose a suitable coefficient. The appropriate coefficient can fully exert the advantages of each classifier and make a joint decision. The final recognition accuracy rate will be better than the recognition accuracy of a single classifier. The use of inappropriate coefficients will result in the classification accuracy of the final joint decision even lower than that of a single classifier.

There will be a Softmax classifier at the last level of the AlexNet and VoxNet network. For each input sample, the output of the Softmax classifier is a probability vector$P = [p_1, p_2, p_3, \ldots p_n]$, that the sample may belong to. Where $p_n$ represents the probability that the sample belongs to class n, and n represents the number of classes of all samples. With $p_1 + p_2 + p_3 + \cdots + p_n = 1$ , the sum of the probabilities that a sample belongs to all classes is 1. In the object recognition of a single classifier, we will choose the probability.

The class label corresponding to the largest element in the vector P is the class corresponding to the sample. In this chapter, for each test sample, the obtained recognition rate results are $P_{AlexNet}$、$P_{VoxNet}$, and the coefficients of each classifier are α and β respectively. Then we can complete the fusion of all base classifiers according to Eq. 1.

$$P = \alpha \times P_{AlexNet} + \beta \times P_{VoxNet} \quad (1)$$

After the fusion is complete, we get a k-dimensional vector, $P = [p_1, p_2, p_3, \ldots p_n]$ where n is the number of categories. We take the largest element among them as the final sample tag.

### D. Recognition results

In order to test the accuracy of this experimental method, the method of this paper is compared with the recognition accuracy of VoxNet and AlexNet. The experimental results are shown in Table 1.

TABLE I. ACCURACY OF DIFFERENT METHODS

| recognition methods | accuracy rate/% |
|---|---|
| AlexNet | 85 |
| VoxNet | 83 |

In this paper, the coefficients α and β before AlexNet and VoxNet are set to different values, the method with higher accuracy is used to set larger coefficients, the method with lower accuracy is set with smaller coefficients, and the comparison between different combinations of coefficients is

accurate for the network. influences. The experimental results are shown in Table 2.

TABLE II. EFFECTS OF DIFFERENT COMBINATIONS OF COEFFICIENTS ON NETWORK ACCURACY

| weight（α, β） | accuracy/% |
|---|---|
| （0.5，0.5） | 79.8 |
| （0.6，0.4） | 81.2 |
| （0.7，0.3） | 91 |
| （0.8，0.2） | 87.5 |
| （0.9，0.1） | 86.3 |

From Table 2, the network has the highest recognition rate when α and β are set to 0.7 and 0.3. Compare this method with the recognition accuracy of VoxNet and AlexNet. Experimental results show that this method has the highest recognition rate.

## IV. CONCLUSION

This paper presents a decision-level three-dimensional target fusion algorithm, we using different convolutional neural network frameworks to extract point cloud features and visual features of three-dimensional objects, respectively, and finally achieved effective fusion. Experimental results show that feature fusion in the decision-making layer is also an effective feature fusion method. This method improves the accuracy of object recognition. In the process of integration, a method with a higher recognition accuracy rate sets a larger coefficient, and a method with a low recognition accuracy rate sets a smaller coefficient, and the final accuracy rate is the highest.

## REFERENCES

[1] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang and J. Xiao. 3D ShapeNets: A Deep Representation for Volumetric Shapes. CVPR2015.

[2] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks[C]// International Conference on Neural Information Processing Systems. Curran Associates Inc. 2012:1097-1105.

[3] D. Maturana and S. Scherer. VoxNet: A 3D Convolutional Neural Network for Real-Time Object Recognition. IROS2015.

[4] H. Su, S. Maji, E. Kalogerakis, E. Learned-Miller. Multi-view Convolutional Neural Networks for 3D Shape Recognition. ICCV2015.

[5] Charles R. Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation. CVPR 2017.

[6] Z. Xie, K. Xu, W. Shan, L. Liu, Y. Xiong, H. Huang, Projective feature learning for 3D shapes with multiview depth images, in: Computer Graphics Forum, vol.34, Wiley Online Library, 2015, pp.1–11.

[7] Z. Zhu, X. Wang, S. Bai, C. Yao, X. Bai, Deep learning representation using autoencoder for 3D shape retrieval, in: Proceedings of the International Conference on Security, Pattern Analysis, and Cybernetics(SPAC), IEEE, 2014, pp.279–284.

[8] Holz D, Holzer S, Rusu R B, et al. Real-Time Plane Segmentation Using RGB-D Cameras[C]// Robot Soccer World Cup XV. Springer-Verlag, 2012:306-317.

[9] Salih Y, Malik A S. Comparison of stochastic filtering methods for 3D tracking[J]. Pattern Recognition, 2011, 44(10–11):2711-2737.

[10] C. R. Qi, H. Su, M. Nießner, A. Dai, M. Yan, and L. Guibas.Volumetric and multi-view cnns for object classification on 3d data. In Proc. Computer Vision and Pattern Recognition (CVPR), IEEE, 2016.

[11] http://www.cnblogs.com/graphics/archive/2010/08/05/1793393.html The Princeton ModelNet. http://modelnet.cs.

[12] The Princeton ModelNet. http://modelnet.cs.

[13] R. B. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proc. CVPR, 2014.