Robust indoor point cloud classification by fusing LSTM neural networks with supervoxel clustering

M.J. Li[1], L. H. Wang[1*], Z. H. Cai1, M. S. Yang[1], R. J. Wu1,M.M.Yao[2]

[1] Guangdong Power Grid Corporation,Guangzhou, China- limingjun @gd.csg.cn,
379876352@qq.com,zhangsan@mail.com,yangmsir@126.com,wurongji@gd.csg.cn
[2] Shenzhen University, Shenzhen, China - yaomeng 1996@163.com

Commission II, WG II/3

KEY WORDS: Indoor Classification, LSTM, Supervoxel, Point Cloud, Machine Learning.

ABSTRACT:

To address the problems of lack of training data and inaccurate classification of existing 3D point cloud data segmentation and classification methods, this paper proposes a high-precision classification algorithm for indoor point clouds by fusing LSTM neural network and super voxels. The algorithm first performs super voxel segmentation on the original point cloud and uses it as the basic unit for machine learning classification, and then introduces LSTM (Long Short-Term Memory) neural network to model the super voxel domain relationship and optimize the classification results. Finally, the accuracy of the proposed method is evaluated based on open dataset, and the experimental results show that 83.2% classification accuracy can be achieved in the open dataset.

1. INTRODUCTION

With the increasing number of indoor spatial applications, semantic segmentation of indoor 3D data has become a hot topicof research for many researchers and scholars (Kang et al.,2020) It is the key to support various intelligent applications,such as indoor navigation (Choi et al., 2014) indoor navigation, indoor robotics (Taira et al., 2018) and augmented reality, among others. It is key to supporting various intelligent applications such as indoor navigation, indoor robotics and augmented reality. Semantic information extraction from point cloud data is the process of identifying and extracting elements from a cluttered and unorganised point cloud. The core of the process is to use segmentation algorithms to divide the disorganised point cloud data across the scene into a series of point cloud collections, so that each collection contains data with the same semantic and perceptual information, and each collection corresponds to a certain type of entity within the scene, making the point cloud have objectified semantic information (Hu et al., 2020). The point clouds have objectified semantic information. A lot of work has been done to improve the segmentation accuracy and processing speed of indoor point cloud data, but there are still two important challenges. Firstly, the raw point cloud data is heterogeneous. First, the original point cloud data is cluttered, sparse and unstructured, and there are problems such as incomplete data collection, uneven density and noise (Tran et al., 2019).This makes it difficult to generalise point cloud data segmentation algorithms to different scenarios.This makes it difficult to generalise point cloud data segmentation algorithms to different scenarios. Secondly, the current point cloud segmentation algorithms mainly classify point cloud data based on colour and geometric features, which rely on a large amount of training data for model learning, while the complex and diverse structure of objects in indoor space makes the current algorithms prone to low applicability and poor stability (Qi et al., 2017a). These algorithms rely on a large amount of training data for model learning.The current research on semantic segmentation of indoor 3D point clouds consists of three main types: multi-view based point cloud classification, voxel grid based point cloud classification and classification algorithms based on the original 3D point cloud. The multi-view based point cloud classification algorithm is to project the 3D point cloud data into 2D images from different angles according to the 3D imaging principle, and then perform the semantic segmentation of the scene based on the mature 2D image segmentation algorithm (Su et al., 2015). This type of algorithm can be used to initialise the multi-view model parameters using a mature, highly accurate pre-trained 2D convolutional neural network. The difficulty of neural network training is significantly reduced, while the effects of 3D geometric problems such as hollow objects and non-fluid geometry in 3D space can be avoided.

GVCNN is a deep neural network for multi-view 3D object recognition, which uses a convolutional neural network with shared view information to extract individual 2D image features for each view, and finally fuses the extracted multi-view feature information into a global 3D object feature information based on the maximum pooling layer of multiple views to achieve global 3D object feature classification.The GVCNN (Feng et al., 2018) and Dominant (Wang et al., 2019) frameworks improve on the MVCNN by using a grouping approach to fuse multi-view features to further exploit the similarity between views to improve recognition accuracy. However, these methods require powerful GPUs for data training and cannot take into account all features in 3D space. Some researchers have therefore investigated the use of 3D point cloud voxel representations to learn features from the scene in 3D space. Unlike point clouds and polygon slices, each voxel has a regular index in the stereo grid. The method extends the 2D convolutional neural network to a 3D convolutional neural network and can be directly applied to 3D voxel convolution. The Rotation Net method (Kanezaki et al., 2018) combines two objective functions, object recognition and view estimation, to build a semantic recognition neural network, and adds the information of each view as an implicit variable in the training of the neural network. 3D-ShapeNet (Wu et al., 2015), the first neural network model to adopt this idea is the 3D-ShapeNet, which expresses a 3D shape in terms of the spatial distribution of binary variables (the presence or absence of objects in voxels) on a grid of stereoscopic voxels. voxNet (Maturana and Scherer, 2015) uses a shallow 3D convolutional neural network to process voxelised 3D point cloud data.The

ORION method (Sedaghat et al., 2016) is an addition to VoxNet that adds a subobjective for estimating the rotational orientation of the object, and the addition of this sub-objective improves the accuracy of semantic recognition. However, the processing time and storage footprint of voxels grows in cubic powers depending on their resolution, and most of the early methods studied were only able to learn with low resolution and shallow neural networks.

Therefore, the OctNet approach (Riegler et al., 2017) In the OctNet method, an unbalanced octree is proposed to divide the 3D stereo grid to solve the sparse problem of effective voxels in the 3D stereo grid,and the algorithm can be used for higher resolution and deeper neural network training. The above methods still suffer from feature loss in the feature computation process, and in recent years, a large number of researchers have investigated how to learn features from raw point clouds for semantic classification.Pointnet (Qi et al., 2017a) is the first neural network model based on 3D point clouds, which first learns the features of each point using a multilayer perceptron (MLP) and then uses a symmetric function to obtain a global object descriptor. pointNet++ (Qi et al., 2017b) adds a hierarchical feature extraction structure to PointNet. It proposes to partition the entire point cloud into several locally grouped ensemble abstraction layers, which act similarly to the convolutional layers in a convolutional neural network, and finally output the perceptual field of features by fusing several ensemble abstraction layers. In contrast to the idea of PointNet++, KCNet (Shen et al., 2018) proposes to use graph pooling layers and kernel correlation to mine the local feature information in the point cloud.

Similar to the aim of KCNet, kd-Net (Klokov and Lempitsky, 2017) is based on the input point cloud and then extracts the feature information hierarchically from the leaf nodes to the root node in a bottom-up manner. However, due to the high complexity of the indoor structure, the data itself is prone to data occlusion, and the training dataset is difficult to obtain, resulting in the current indoor 3D point cloud semantic segmentation methods taking a long time to train and still struggling to achieve the desired classification accuracy.

To address the problem of internal inconsistency of classification targets in existing 3D point cloud data segmentation and classification methods, we propose a high-precision classification algorithm for indoor point clouds jointly optimized by super voxel random forest (Ramiya et al., 2016, Oshiro et al., 2012) and Long Short-Term Memory (LSTM) neural network(Sherstinsky, 2020) . The algorithm is based on the feature that super voxels have internal feature consistency, divides the original point cloud into super voxels, and uses super voxels as the basic unit for multivariate feature calculation, builds the indoor point cloud super voxel random forest classification model, and realizes the coarse classification of point cloud data. On this basis, LSTM is introduced to train and predict the neural network model for the hyper voxel neighborhood connectivity of coarse classification to achieve the optimization of hyper voxel coarse classification results. Finally, the validity and accuracy of the proposed classification method are verified based on the open dataset, and the results show that the classification method of this paper can achieve 83.2% classification accuracy in the public dataset.

2. COARSE CLASSIFICATION OF INDOOR POINT CLOUDS IN SUPER VOXEL RANDOM FORESTS

As shown in Figure 1, the process of indoor point cloud segmentation method is optimized jointly by super voxel random forest and LSTM network. In the coarse classification stage, the original point cloud is clustered by super voxels to obtain the super voxel centroids, and the multi-dimensional features of super voxels are calculated and used for the training of the random forest model, which mainly consists of randomization, decision tree generation and voting classification steps. The process of coarse classification mainly consists of randomization, decision
tree generation, and voting classification. The hyper voxel features involved in this paper contain four main types, which are local density features, Point Feature Histogram (PFH) features (Rusu et al., 2009), normal vector features, color information, relative elevation features, and shape features. After the supervoxel RF classification, the keras deep learning framework is used as the basis for the construction of super voxel LSTM neural network, which is used for the optimization for point cloud classification.

1) After indoor scene hyper voxels are clustered, the scene is partitioned into several blocks and the hyper voxels are linked to each other.

2) Iterative search is performed on the neighborhood information of each super voxel divided in the scene, and the surrounding voxels of the current super voxel are searched by KDtree, combined with their feature information, and combined into a spatial sequence set by distance size.

3) For the super voxel LSTM network training, a model with three LSTM layers plus one fully connected layer is designed, and the LSTM layers all use The LSTM layer is used as the activation function, and finally enters the fully connected layer, and uses The LSTM layer is used as the activation function and finally enters the fully connected layer, and the multi-objective classification of the scene is achieved using the activation function, in which the parameters of the neural network are initialized in a random way and rmsprop is used as the optimizer.

In the training process, the model training batch size is set to 128 and the number of iterations epoch is 80. Considering the problem of unevenness of different types of super voxels in the training data, the category weights are calculated and added to the training process according to the number of categories in the training set.

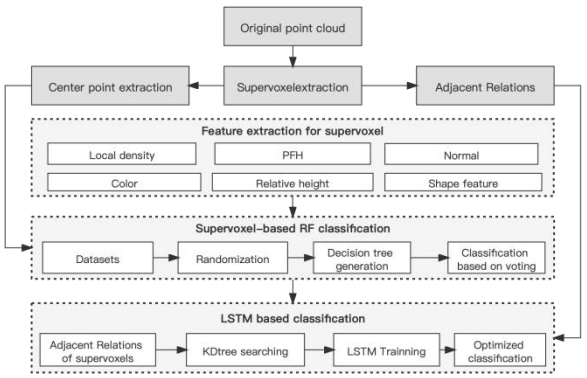

Figure 1 The framework of the proposed method

2.1 Super voxel characterisation

There are four main types of hyper voxel features involved in this paper, namely local density features, Point Feature Histogram (PFH) features, normal vector features, colour information, relative elevation features and shape features. The method proposed in this paper uses the super voxels as the basic classification unit for classification, therefore the features extracted below are the feature information of the centroid of each super voxel.

Local density feature: The local density feature is the average distance from a point to the k nearest neighbour points. Therefore, for each centroid in the super voxel, fast retrieval of neighbourhood points is achieved by constructing a KdTree and fast library for approximate nearest neighbors (FLANN) algorithm, which in turn obtains the local density feature of a point by calculating the average Euclidean distance between two pairs of neighbouring points.

PFH feature: The PFH feature is a description of the geometric properties of a point's k-neighbourhood by parameterising the spatial differences between the query point and its neighbourhood points and forming a multi-dimensional histogram. Specifically it is based on the relationship between points and their k-neighbourhoods and their normal vectors to describe the geometric features of the sample. In this paper, the PFH features of each centroid are obtained by creating a kdtree of the original point cloud, which is computed by k-neighbourhood search.

Normal vector features: The normal vector of each point in the point cloud represents the direction of the surface on which the point is located, and can accurately describe both planar and surface information. In this paper, we calculate the normal vector information of the super voxels by plane fitting and calculate their feet to the vertical direction as random forest features.

Colour features: Most of the classification targets in indoor environments have colour consistency, and therefore RGB colour plays an important role in the indoor point cloud segmentation process. Considering the super voxels as the basic classification unit in this paper, the colour information of each super voxel is determined by the average RGB value of the points within the super voxel.

Relative elevation features: The relative elevation features of the super voxels are obtained from the difference between the height of the centre point of the super voxels and the elevation of the ground plane.

Shape features: The shape feature parameters are mainly calculated by the Eigen feature values obtained from the local PCA decomposition of the point cloud and combined to obtain them. The traditional Eigen eigenvalue is calculated based on the local point cloud obtained by K-neighborhood search. In order to obtain a more accurate domain point cloud, this paper takes the super voxel itself as the domain information of the current super voxel centroid and uses the point cloud inside the super voxel for the calculation of the Eigen eigenvalue,after the feature decomposition, three eigenvalues are obtained, which areλ_1、λ_2、λ_3, where the three eigenvalues are listed in order from largest to smallest, i.e. ($\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq 0$). Based on this, the curvature, linearity, planarity, scattering and anisotropy of the super voxels are calculated according to the shape feature calculation method. The calculations are shown in the

Table 1 Calculation methods for shape features

Shape features	Calculation method
Curvature (Curvature)	$C_e = \frac{\lambda_3}{\lambda_1 + \lambda_2 + \lambda_3}$
Linearity (Linearity)	$L_e = \frac{\lambda_1 - \lambda_2}{\lambda_1}$
Flatness (Planarity)	$P_e = \frac{\lambda_2 - \lambda_3}{\lambda_1}$
Dispersion (Scattering)	$S_e = \frac{\lambda_1}{\lambda_3}$
Anisotropy (Anisotropy)	$A_e = \frac{\lambda_3 - \lambda_1}{\lambda_3}$

2.2 Super voxel random forest model construction

The random forest model construction in this paper uses super voxels as the basic unit for training and prediction, similar to the traditional random forest construction method, the super voxel random forest consists of N decision trees $\{h(X, \theta_n), n = 1,2,3, \cdots, N\}$ as the initial classifier, and the final combined classifier is obtained by integrated learning. The random forest counts the results of each decision tree classification and votes on the output classification. Of these $\{\theta_n, n = 1,2,3, \cdots, N\}$ are sequences of random variables, determined by the Bagging strategy and the feature subspace strategy in the random forest. Specifically: 1) The Bagging strategy is to randomly sample N training samples of the same size as the original dataset from the original dataset $\{T_n, n = 1,2,3, \cdots, N\}$ (about 63% of the samples are sampled each time), and a decision tree is trained for each training sample set.T_n 2) The feature subspace strategy is to split and refine each node in the decision tree by selecting a subset of features from the data features and choosing the best feature segmentation node.

Finally, Random Forest is a combined classifier that integrates multiple decision tree classifiers and ultimately decides the classification result by classifier voting. The basic process of classification is as follows:

1) A bootstrap sampling method was used to randomly select K training sample sets from the original sample set.

2) A decision tree model is constructed for each of the K training sample sets to obtain K classification results. Specifically, each decision tree will select N features from the M features of the input variables. Generally the value of N is taken according to the formula $N = \sqrt{M}$ The value of N is determined according to the formula. In turn, information entropy (entropy) and Gini index (Gini) are used as node splitting criteria, as shown in Eqs. 1 and 2, where$n$ denotes the number of categories contained in the training data set D, and p_i denotes the probability of the training data belonging to a certain category.

$$\text{entropy}(D) = -\sum_{i=1}^{n} p_i \log_2 p_i \quad (1)$$

$$\text{Gini}(D) = 1 - \sum_{i=1}^{n} p_i^2 \quad (2)$$

(3) The final classification is determined by voting based on the K classifications.

2.3 LSTM neural network optimization for fine classification of indoor point clouds

Unlike the original point cloud data, the point cloud can obtain the connection relationship between super voxels and super voxels after super voxel clustering, and the connection relationship contains the association characteristics between different types of elements, for example, the super voxels of the desktop have a certain correlation with the desktop clutter super voxels, and the judgment of the correlation can avoid the clutter from being incorrectly segmented into objects such as chairs.

Therefore, this paper proposes a Long short-term memory (LSTM) neural network optimisation method for modelling hyper voxel association sequences based on the results of random forest coarse classification of hyper voxels, and

optimises the classification results of indoor 3D point clouds. The core reason for modelling the spatial connectivity of super voxels based on LSTM networks is that the LSTM models the sequence data in such a way that the horizontal neurons in its internal structure run through the series of data, and the state information of the neurons can be transmitted sequentially throughout the chain with only linear interactions, so that the information on the neurons in the chain can remain approximately constant, thus preserving long-term information. This gives it a significant advantage in the extraction of long and short term information, which is why LSTM has been used in the past for data with a clear sequence and correlation, such as long text classification and time series data prediction, where it is often possible to obtain better results than traditional time series prediction methods.

This paper uses the keras deep learning framework as the basis for the construction of the super voxel LSTM neural network.

1) After clustering of indoor scene hyper voxels, the scene is partitioned into blocks and the hyper voxels are linked to each other.

2) The neighbourhood information of each super voxel divided in the scene is searched iteratively, and the surrounding voxels of the current super voxel are searched through the KDtree, combined with their feature information, and combined into a spatial sequence set by distance size.

3) A model with three LSTM layers plus one fully connected layer was designed for training the super voxel LSTM network, with the LSTM layers all using The LSTM layer is used as the activation function, and finally enters the fully connected layer, and uses The LSTM layer is used as the activation function and finally enters the fully-connected layer, and the multi-objective classification of scenes is achieved using the activation function, where the neural network parameters are initialized in a random way and rmsprop is used as the optimizer. In the training process, the model training batch_size was set to 128 and the number of iterations epoch was 80. Considering the problem of unevenness of different types of super voxels in the training data, category weights were calculated and added to the training process based on the number of categories in the training set.

3. EXPERIMENTS

3.1 Datasets

The point cloud dataset used in this experiment is a publicly available dataset from Stanford University, referred to Figure Figure 1. The S3DIS dataset is a semantic dataset with pixel-level semantic annotations developed by Stanford University, and is divided into 6 regions containing 272 scenes, which can be classified into 11 categories of scenes, door, wall, floor, beam, window, chair, cluster, column, etc. In this paper, regions 1-5 are selected as training data and region 6 is used as test region for accuracy evaluation.

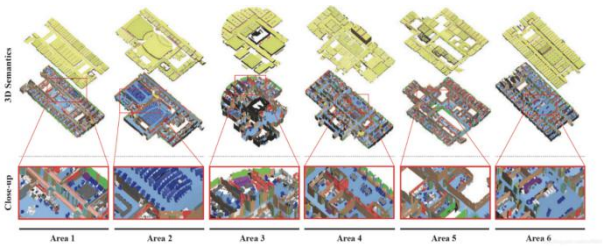

Figure 2 S3DIS datasets

3.2 Experimental results and analysis

Analysis of super voxel segmentation results: reasonable super voxel parameter settings can avoid the occurrence of incorrect segmentation. As shown in Figure 3The overall segmentation accuracy of super voxels is about 94.5%, and the segmentation accuracy of chair, floor, bookcase and sofa can reach over 97%, shown in Figure 4. This is closely related to the distribution of indoor point clouds. For the point cloud data close to walls, ceilings and tables, there are sparse quantities and incomplete structures, while the super voxel segmentation algorithm only considers their vector, colour and shape information. It is difficult to avoid mis-segmentation of point clouds by considering only vector, colour and shape information. Based on the segmentation results of the super voxels, a high-precision indoor point cloud classification method based on the joint optimization of super voxel random forest and LSTM network is proposed in this paper to perform semantic classification experiments on the modified point clouds.

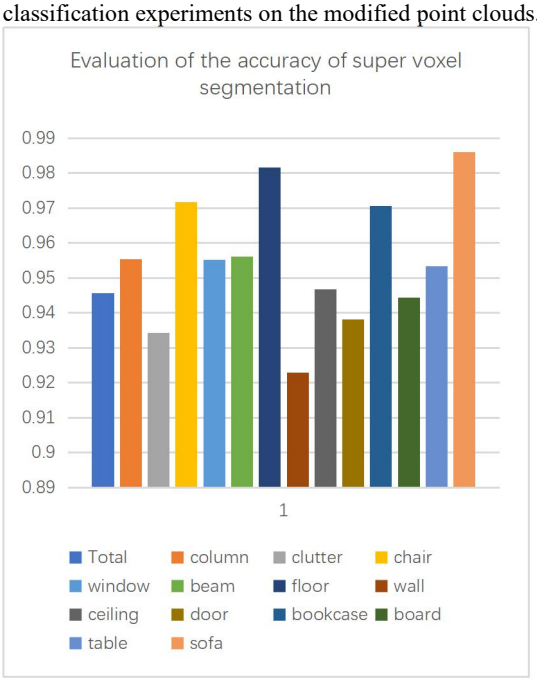

Figure 3 Super voxel segmentation accuracy

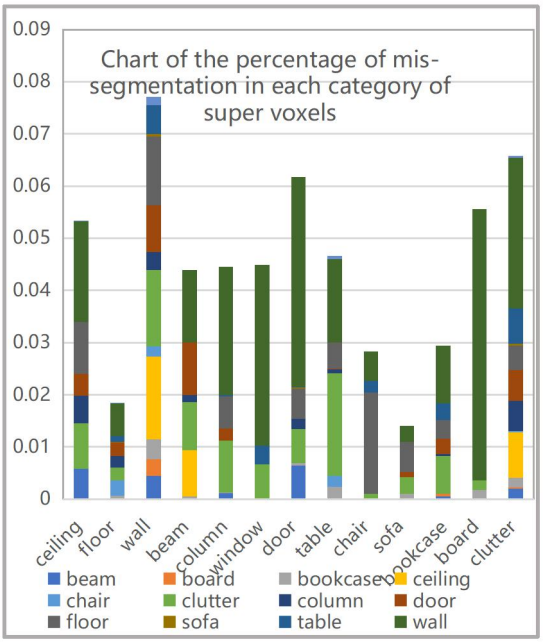

Figure 4 Chart of the percentage of mis-segmentation in each category of super voxels

LSTM neural network training: In order to obtain an optimal training result, the convergence of the model under different epoch values was investigated, and iterative tests showed that the epoch parameter of the LSTM network proposed in this paper was set at around 80, which can effectively avoid the problem of model overfitting.

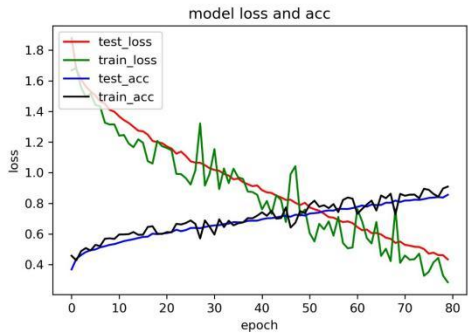

Figure 5 Loss vs epoch relationship

Analysis of point cloud classification results: The average intersection and concurrence ratio mIoU and mAcc are used to evaluate the accuracy of point cloud classification. mIoU represents the ratio of the intersection and concurrence of the two sets of true and predicted values of data classification, and mAcc represents the ratio of the intersection and true values of the true and predicted values of classification. The specific calculation method is shown in Equation (1). Suppose there are $k+1$ There are two categories (including the background category), denote p_{ij} is the ratio of the i is the number of predicted categories asj the number of points in the class, andp_{ii} denotes the true value ofi and the predicted value isi the number of points for which the predicted value isp_{ji} denotes the number of points for which the true value isj and the predicted value isi the number of points with a true value of , and the number of points with a predicted value of .

$$\mathrm{mIoU} = \frac{1}{k+1}\sum_{i=0}^{k}\frac{p_{ii}}{\sum_{j=0}^{k}p_{ij}+\sum_{j=0}^{k}p_{ji}-p_{ii}} \qquad (3)$$

$$\mathrm{mAcc} = \frac{1}{k+1}\sum_{i=0}^{k}\frac{p_{ii}}{\sum_{j=0}^{k}p_{ij}+\sum_{i=0}^{k}p_{ii}} \qquad (4)$$

In this paper, four commonly used point cloud classification frameworks, including RF classification based on the original point cloud, PointCNN, PVCNN++ and PointNet++, are selected and their classification results are compared and analysed. The classification accuracies of different classification methods are listed in Table 2.

Table 2 Comparison of the classification accuracy of different methods

Methods	mIoU	mAcc
Super voxel random forest coarse classification	39.20%	72.40%
Super voxel LSTM optimization	46.30%	83.20%
RF classification based on raw point cloud	8.70%	24.30%
PointCNN	65.40%	75.60%
PVCNN++	59.00%	87.10%
PointNet++	54.50%	-----

BRISK has the best performance of recall within the 0.25 meters and 2 degrees threshold, comparing with ORB and SIFT. And the average time cost of one image is 1.46 second, so we use BRISK in most visual localization experiment we carry out. Image retrieval based on multi-features provides more reliable results, so there would be more inliers of 2D-3D correspondences, which improve the localization performance. However, the time cost of this method is rather high, because image retrieval of different features is carried out instead of the classic single one. Though it reaches high precision, which feature strategy to choose is still based on the requirement of scene and experiment.

It can be seen that the original point cloud-based spontaneous classification method has the lowest accuracy, with its mIoU reaching only 8.7% accuracy and its mAcc only 24.3% accuracy, the lowest accuracy among all classification algorithms. In contrast, the original super voxel random forest coarse classification method can achieve 72.4% accuracy of mAcc, which indicates that the pre-processing of super voxels can effectively improve the classification accuracy of point cloud data. On the basis of coarse classification, the LSTM optimized mIoU can reach 46% and the mAcc is 83.2%, which is similar to the accuracy obtained by the deep learning frameworks PointCnn and PVCNN++. It is worth mentioning that the training data of the LSTM optimised network proposed in this paper only used the label information of region 1 for model training, while other deep learning frameworks used regions 1-5 for model training, therefore, from the perspective of training data requirements, the point cloud data classification framework proposed in this paper can achieve a relatively better prediction result with a small training data set.

4. CONCLUSION

In this paper, we propose a high-precision indoor point cloud classification method jointly optimised by super voxel random forest and LSTM network, which makes full use of the internal feature consistency of super voxels, divides the original point cloud into super voxels, calculates the geometric, colour and shape features of super voxels as the basic unit, and builds a super voxel random forest classification model for indoor point clouds to achieve coarse The classification is based on the idea of coarse to fine classification. Based on the idea of coarse to fine classification, this paper introduces LSTM to train and predict the coarse classification of super voxel neighbourhood connections, taking into account the correlation characteristics between different types of elements contained in the connection relations between super voxels, and achieves the optimization of the coarse classification results of super voxels. Finally, the validity and accuracy of the proposed classification method were verified based on open datasets, and the results showed that the classification method in this paper could achieve 83.2% classification accuracy in open datasets.

In future work, current LSTM network structures are more biased towards sequential inputs with logical order, whereas there is a two-by-two connection between spatial sequence data extracted by super voxels. How to design a sequential neural network structure with non-sequential super voxel arrangement will be an important research direction.

ACKNOWLEDGMENTS

This work was supported in part by the National Key Research and Development Program of China (Projects Nos. 2019YFB210310, 2019YFB2103104) and in part by a Research Program of Shenzhen S and T Innovation Committee grant (Projects Nos. JCYJ20210324093012033, JCYJ20210324093600002), the Natural Science Foundation of Guangdong Province grant (Projects No. 2121A1515012574), the Open Fund of Key Laboratory of Urban Land Resources Monitoring and Simulation, MNR(Nos. KF-2021-06-125,KF-2019-04-014), the National Natural Science Foundation of China grant (Projects Nos. 71901147, 41901329, 41971354, 41971341) and the Foshan City to promote scientific and technological achievements of universities to serve industrial development support projects(Projects No. 2020DZXX04).

REFERENCES

Choi, J., Choi, J., Kim, I., 2014. Development of BIM-based evacuation regulation checking system for high-rise and complex buildings. Automation in construction, 46, 38–49.

Feng, Y., Zhang, Z., Zhao, X., Ji, R., Gao, Y., 2018. Gvcnn: Group-view convolutional neural networks for 3d shape recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 264–272.

Hu, Q., Yang, B., Xie, L., Rosa, S., Guo, Y., Wang, Z., Trigoni, N., Markham, A., 2020. Randla-net: Efficient semantic segmentation of large-scale point clouds. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 11108–11117

Kanezaki, A., Matsushita, Y., Nishida, Y., 2018. Rotation net: Joint object categorization and pose estimation using multiviews from unsupervised viewpoints. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 5010–5019.

Kang, Z., Yang, J., Yang, Z., Cheng, S., 2020. A review of techniques for 3d reconstruction of indoor environments. ISPRS International Journal of Geo-Information, 9(5), 330.

Klokov, R., Lempitsky, V., 2017. Escape from cells: Deep kd networks for the recognition of 3d point cloud models. Proceedings of the IEEE international conference on computer vision, 863–872.

Li, Y., Bu, R., Sun, M., Wu, W., Di, X., Chen, B., 2018. Pointcnn: Convolution on x-transformed points. Advances in neural information processing systems, 31, 820–830.

Liu, Z., Tang, H., Lin, Y., Han, S., 2019. Point-voxel cnn for efficient 3d deep learning. arXiv preprint arXiv:1907.03739.

Maturana, D., Scherer, S., 2015. Voxnet: A 3d convolutional neural network for real-time object recognition. 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, 922–928.

Oshiro, T. M., Perez, P. S., Baranauskas, J. A., 2012. How any trees in a random forest? International workshop on machine learning and data mining in pattern recognition, Springer, 154–168.

Qi, C. R., Su, H., Mo, K., Guibas, L. J., 2017a. Pointnet: Deep learning on point sets for 3d classification and segmentation. Proceedings of the IEEE conference on computer vision and pattern recognition, 652–660.

Qi, C. R., Yi, L., Su, H., Guibas, L. J., 2017b. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. arXiv preprint arXiv:1706.02413.

Ramiya, A. M., Nidamanuri, R. R., Ramakrishnan, K., 2016. A supervoxel-based spectro-spatial approach for 3D urban point cloud labelling. International Journal of Remote Sensing, 37(17), 4172–4200.

Riegler, G., Osman Ulusoy, A., Geiger, A., 2017. Octnet: Learning deep 3d representations at high resolutions. Proceedings of the IEEE conference on computer vision and pattern recognition, 3577–3586.

Rusu, R. B., Blodow, N., Beetz, M., 2009. Fast point feature histograms (fpfh) for 3d registration. 2009 IEEE international conference on robotics and automation, IEEE, 3212–3217.

Sedaghat, N., Zolfaghari, M., Amiri, E., Brox, T., 2016. Orientation-boosted voxel nets for 3D object recognition. arXiv preprint arXiv:1604.03351.

Shen, Y., Feng, C., Yang, Y., Tian, D., 2018. Mining point cloud local structures by kernel correlation and graph pooling. Proceedings of the IEEE conference on computer vision and pattern recognition, 4548–4557.

Sherstinsky, A., 2020. Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. Physica D: Nonlinear Phenomena, 404, 132306.

Su, H., Maji, S., Kalogerakis, E., Learned-Miller, E., 2015. Multi-view convolutional neural networks for 3d shape recognition. Proceedings of the IEEE international conference on computer vision, 945–953.

Taira, H., Okutomi, M., Sattler, T., Cimpoi, M., Pollefeys, M., Sivic, J., Pajdla, T., Torii, A., 2018. Inloc: Indoor visual local ization with dense matching and view synthesis. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 7199–7209.

Tran, H., Khoshelham, K., Kealy, A., 2019. Geometric comparison and quality evaluation of 3D models of indoor environments. ISPRS journal of photogrammetry and remote sensing,149, 29–39.

Wang, C., Pelillo, M., Siddiqi, K., 2019. Dominant set clustering and pooling for multi-view 3d object recognition. arXiv preprint arXiv:1906.01592.

Wu, Z., Song, S., Khosla, A., Yu, F., Zhang, L., Tang, X., Xiao, J., 2015. 3d shapenets: A deep representation for volumetric shapes. Proceedings of the IEEE conference on computer vision and pattern recognition, 1912–1920.

Xu, Q., Sun, X., Wu, C.-Y., Wang, P., Neumann, U., 2020. Grid gcn for fast and scalable point cloud learning. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 5661–5670.