This is a repository copy of *VoxRec : hybrid convolutional neural network for active 3D object recognition*.

White Rose Research Online URL for this paper:
http://eprints.whiterose.ac.uk/160901/

Version: Published Version

Article:
Karambakhsh, A., Sheng, B., Li, P. et al. (3 more authors) (2020) VoxRec : hybrid convolutional neural network for active 3D object recognition. IEEE Access, 8. pp. 70969-70980.

https://doi.org/10.1109/access.2020.2987177

Reuse
This article is distributed under the terms of the Creative Commons Attribution (CC BY) licence. This licence allows you to distribute, remix, tweak, and build upon the work, even commercially, as long as you credit the authors for the original work. More information and the full terms of the licence here: https://creativecommons.org/licenses/

Takedown
If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk/

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 7, 2020, accepted March 28, 2020, date of publication April 10, 2020, date of current version April 28, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2987177

VoxRec: Hybrid Convolutional Neural Network for Active 3D Object Recognition

AHMAD KARAMBAKHSH[1], BIN SHENG[1], (Member, IEEE), PING LI[2], (Member, IEEE), PO YANG[3], (Senior Member, IEEE), YOUNHYUN JUNG[4,5], AND DAVID DAGAN FENG[6], (Life Fellow, IEEE)

[1]Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240, China
[2]Department of Computing, The Hong Kong Polytechnic University, Hong Kong
[3]Department of Computer Science, The University of Sheffield, Sheffield S1 4DP, U.K.
[4]Department of Software, Gachon University, Seongnam 13120, South Korea
[5]Biomedical and Multimedia Information Technology Research Group, School of Information Technologies, The University of Sydney, Sydney, NSW 2006, Australia
[6]Biomedical and Multimedia Information Technology Research Group, School of Information Technologies, The University of Sydney, Sydney, NSW 2006, Australia

Corresponding author: Bin Sheng (shengbin@sjtu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61872241 and Grant 61572316, in part by the Science and Technology Commission of Shanghai Municipality under Grant 18410750700, Grant 17411952600, and Grant 16DZ0501100, and in part by The Hong Kong Polytechnic University under Grant P0030419 and Grant P0030929.

ABSTRACT Deep Neural Network methods have been used to a variety of challenges in automatic 3D recognition. Although discovered techniques provide many advantages in comparison with conventional methods, they still suffer from different drawbacks, e.g., a large number of pre-processing stages and time-consuming training. In this paper, an innovative approach has been suggested for recognizing 3D models. It contains encoding 3D point clouds, surface normal, and surface curvature, merge them to provide more effective input data, and train it via a deep convolutional neural network on Shapenetcore dataset. We also proposed a similar method for 3D segmentation using Octree coding method. Finally, comparing the accuracy with some of the state-of-the-art demonstrates the effectiveness of our proposed method.

INDEX TERMS Object recognition, recurrent neural networks, multi-layer neural network, octrees.

I. INTRODUCTION

With the fast development of 3D scanning and modelling devices, 3D model's repositories have become huge. These repositories include a mixture of different 3D models which requires to be categorized. Moreover, using VR (Virtual Reality) in some academic environments is obtained much attention, which improves the performance of students and, in some cases, decreases the cost and risk of using other tools for teaching. Due to a large number of models, it is an arduous task to organize them manually. Thus, having 3D recognition methods is necessary for each environment. 3D recognition helps them to organize their dataset and classify any other new 3d models. While, recently, many researchers have concentrated on such the research area, there are still many challenges remained murky. These challenges are not just about the accuracy of their provided methods; there are many other factors to reach an acceptable method which have to be specified according to individual environments. For example, presented researchs on 3D models of human body organs are severely limited. Distinguishing two pieces of human bones with similar shape would need a model to focus on more details rather than just shapes. Therefore, to ease the classification procedure, we require to use an automatic recognition approach, avoiding tedious work and being specialised to the target environment. This triggered us to focus on this research area, and this paper proves that our straightforward method can outperform many other popular methods in terms of 3D recognition.

In this paper, we propose a novel approach that trains a large scale dataset in terms of recognition; it not only focuses on the overall shape of the object but also uses surface features to be more accurate. This method takes advantage of the different presentations of 3D objects, such as surface features and volumetric representation. Each of these representations proved to be helpful in some aspects, and gathering them together would give us an opportunity to benefit from all of them. It means the proposed method has

The associate editor coordinating the review of this manuscript and approving it for publication was Xiaogang Jin.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

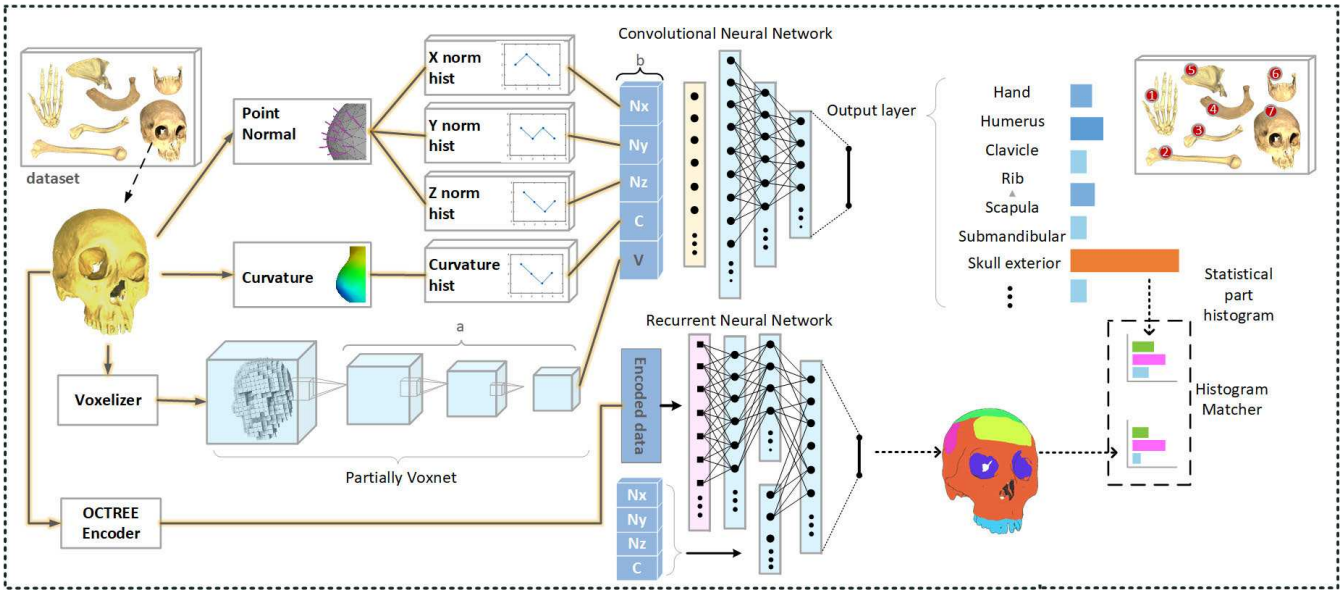


FIGURE 1. The network structure is a combination of five convolutional neural networks to use entire input data for categorization. (a) shows the 3D convolutional layers, and (b) shows the merging layers.

to be a combination of five convolutional neural networks (CNNs). In the first stage, a 3D convolutional method has been prepared to process volumetric data and extract enough features of the input surface shapes. One of the main critiques of volumetric data is the vast memory consumption. In this research, instead of using high-resolution volumetric data, low-resolution has been used. However, to reach the details and train them precisely, the other provided features have been evaluated.

We divide the suggested neural network model to five networks; one of them is focused on 3D volumes, and the other fours train the surface features. The considered surface features in this method are surface normal and surface curvature. These features that are provided in some of the famous datasets can improve the accuracy prepared by 3D volume to be competitive in 3D recognition purposes. These data have less information in size and more in aspects of identification. It would assist the learning method to achieve a reasonable result in a proper time. It also optimizes the way of mesh recognition by encoding the information, extracting essential features, and learning the encoded data. Besides, it has been beneficial to different environments due to accessing local features as well as the global shape. The surface normal and surface curvature have been considered for using surface features in 3D model recognition. These features generally describe a point or a local area of an object. To take advantage of such features, we design a procedure to convert them into a histogram, which represents the distribution of the features across a whole object. A convolutional neural network (CNN) can be assisted by such the histograms and volumes to distinguish objects in a dataset.

Moreover, we added a part annotation stage to our method for improving the recognition accuracy. In the first step, Octree encoding converts the point cloud to a series of binary data before training. The encoding level contains a modified version of an octree-like method, which was inspired by octnet [1]. The encoded object and surface features would be the input of the training procedure. After the recognition stage, the result can be verified by simple evaluation and the fault classified object. Our main contributions include:

- A new machine learning approach for recognition of a large 3D objects dataset is proposed. The approach preprocesses the 3D models to extract normal and curvature features. Merging these features and voxel data, significantly improves the efficiency of the 3D recognition method.
- Expanding the suggested recognition method by applying a 3D mesh part annotation. The method replaces the voxel data with encoded octree data, which significantly improves the efficiency of our approach. Despite the ability of segmentation, it exploits the segmented result for the improvement of the recognition method. This opinion improves the accuracy to be competitive with other state-of-the-arts.
- The suggested method takes advantage of various features and merges them into one neural network. This process not only improves the recognition accuracy but also provides the method to be more generalizable due to using different objects dataset.

II. RELATED WORK

There are many pieces of research in the computer vision and graphics which have been dedicated to establishing a way of recognizing 3D objects. Several representations are employed to describe 3D models, such as shape descriptors, voxels, and projected view representations. Besides, a variety of methods are used to assess this information and provide

the favour results, such as methods for the informative region selection, feature extraction, and classification [2]. In this section, we describe some of the papers which take advantage of classification methods. In 2017, Czajewski and Kołomyjec [3] published a remarkable paper in 3D mesh recognition based on color and 3D depth (RGB-D) images. The proposed method used Viewpoint Feature Histogram and Camera Roll Histogram as their descriptor. ICP (Iterative closest point) was then employed for the main matcher. From their description, their recognition performance is better than the convolutional neural network (CNN)-recurrent neural network(RNN) method from Socher *et al.* 2012 [4]. The CNN-RNN method proposed a model that merges convolutional and recursive neural networks (RNN) for extracting features and analyzing RGB-D images. Reference [5] by Beserra Gomes *et al.* suggested the moving fovea method to down-sample 3D data and decrease the processing time of the object classifier system from point clouds. They stated their object recognizer could run 7x faster than non-foveated approaches. The central concept shows that the point density should be higher close to the fovea. This density is declined according to the distance from the fovea. It means that they reduce the number of points and calculation at the same time.

VoxNet [6] by Maturana and Scherer concentrated on light detection and ranging(LIDAR) and RGB-D cameras to enhance the robot perception from a real environment. They suggested method, combining a volumetric occupancy grid representation with a supervised 3D CNN. It has to be mentioned; the VoxNet is a groundwork of many other proposed methods afterward. Another related work is [7] by Zhirong *et al.*, which focused on obtaining a volumetric representation of a 3D model from 2.5D range data. This opinion achieved an exciting result on depth sensors such as the Kinect. Meanwhile, Su *et al.* by [8] proposed to render 12-views from 3D meshes and categorize the rendered images rather than working on 3D meshes. To do this, they applied VGG [9] that is already trained on ImageNet data. MVCNN-MultiRes [10] enhance MVCNN by using rendered images from different resolutions. Moreover, the 3D object recognition of the ModelNet dataset is addressed by FusionNet [11] using two data representations: pixel representation and volumetric representation. They merged two different voxel CNN networks with a single multi-view network and the result performed 92% accuracy on modelnet10 and 90% accuracy in modelnet40, which was the most recognition accuracy in 2016.

CNN have performed the best performance in various computer vision tasks, including action recognition [12] and object recognition such as large-scale classification [13]. Through jointly encoding convoluted information in the training process, 2D convolutional networks have achieved the best performance in object detection and classification. Other investigations used 3D CNN structures to deliver recognition and detection tasks in a video by tuning the networks using video frames [14]. Gkioxari and Malik [15] proposed an action detection system that aimed to detect bounding boxes

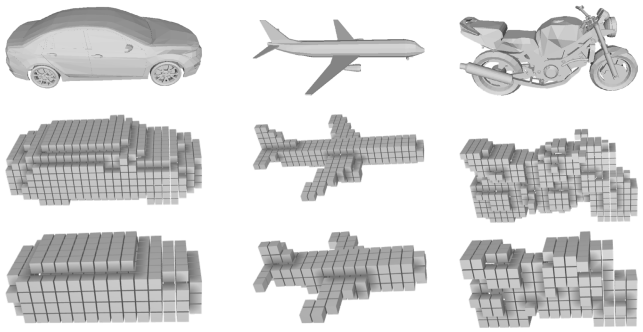


FIGURE 2. Three different 3D meshes from the Shapenetcore-part dataset with a varying resolution of their voxel data. The second raw shows that the smaller voxel size can keep the details more than the larger one.

of actions frame-by-frame from a video. On the other hand, a series of supervised methods achieve great performances on mesh part annotation and part labelling. But, for example, Yi *et al.* [16] rely on databases of segmented objects, which is a highly labor-intensive process. Moreover, catching the right scale part is a very intensive task for a non-expert manual annotator. To handle these challenges, SyncSpecCNN [17] is researched on a spectral CNN method on a graph of triangulated vertices by Yi *et al.* Or VoxelNet [18] by Zhou and Tuzel investigated on using convolutional method and volumetric data to deliver an accurate 3D recognizer in 2017. The results achieved by them not only work for classification purposes but also to localize the objects in Kitti dataset.

Zhi *et al.* by Lightnet [19] propose a real-time method of anticipating class label and orientation information, without extra annotation data. They introduced a shallow network for 3D object recognition accuracy that overcomes some of the state-of-the-art methods in the number of training parameters. The other research paper is [20] that leverages GAN (Adversarial Generative Network) to generate object structure implicitly. This network is also able to establish a 3D model from a low-dimensional probabilistic model. In addition to these, it provides robust shape descriptors enabling us for recognition purposes. In 2014, Liang *et al.* published [21] that focused on 3D object recognition and position estimation from multiple projected views of a 3D mesh. Their model adopted two DBFs (deep belief network) to obtain the image features and use escape connection to the last layer to match the features and analyze the input data. Besides, they apply a new DBF that merged two traditional DBFs and estimated the camera position, the same as the classification method. In this method, they also applied the K-mean clustering to overcome the weakness in object detection,which provides accurate results.

PointNet [22] cannot recognize local features via the metric space points, restricting its ability to achieve fine-grained patterns. It has some difficulties in complicated scenes. An improved version of PointNet, so-called PointNet++ by [23] presented a hierarchical neural network that utilized PointNet recursively on a nested partitioning of a list of 3D points.

By utilizing metric space distances, the network can learn local features with increasing contextual scales. Liu *et al.* by [24] suggested using volumetric representation and unsupervised deep network to obtain the features of point cloud data. They likewise applied the Hough Forest method on the gathered features and achieve object detection and position estimation concurrently. They compared the results on the 2.5D dataset of Tejani *et al.* [25], and performed an almost acceptable score.

Several 2D recognition methods have been studied in terms of investigation on recent research in recognition topics. For example, if we focus on image recognition, RVM (Representative vector machine) [26] is highlighted. It concentrated on character recognition with PC-2DLSTM (Principal Component 2-D Long Short-Term Memory) [27] and metric learning-based recognition [28] and achieved the accurate result using deep neural network and in of face recognition and facial expression. Also, SSP (superimposed sparse parameter) classifier [29] and AFERS (facial expression recognition system) [30] have been proposed as the top approaches for classification purposes. MIT university in 2018 by [31] described the EdgeConv layer in deep networks to obtain local geometric features of point clouds. The entire architecture followed the Pointnet architecture except applying EdgeConv Blocks; it has produced an inevitable outcome on the experimental results. They had got an accuracy of 92% for the classification of ModelNet40. It was better than the state of the arts such as Pointnet++, VoxNet, and KD-Net. Also, [32] by Jin Xie *et al.* examined using nonlinear distance metric in 3D shape descriptors for retrieval. They assessed their results on SHREC'10, ShapeGoogle, McGill, and SHREC'14 datasets. The progress of methods using CNN not only can reach more accuracy, but also, some researches are designed to be simple for operating such as Sun *et al.* [33]. Such the method does not need to process the data before or after CNN, and thus it is easily manageable.

III. 3D OBJECT RECOGNITION

We propose a combination of 5 different CNNs that work using voxel, surface normal, and surface curvature (see FIGURE 1). The main difference from our recognizer and other proposed approaches is using some parameters which explain the adjacency of a point, such as surface normal, which contains the direction information and curvature data, which gives us information of edges on the surface. The proposed method can train the entire categories in a short time, besides having a top and reliable accuracy.

A. PREPROCESSING

The first step of the proposed method is to convert the data from a list of the points to a series of information that can quickly be learned by our light and deep neural network. This information-processing includes: converting a point cloud to volumetric data, and converting normal and curvature information to a series of one-dimensional histograms. Firstly, the volumetric data for our primary target dataset, Shapenetcore-part [16], has already been provided (see FIGURE 2). However, in the case that volumetric data is necessary to be acquired, the 3D occupancy grid [34] is an accurate method. This method allows us efficiently estimate occupied space between two 3D points. Also, it can be stored and be operated with efficient and straight forward data structures. Secondly, as the Algorithm 1 is illustrated, we take advantage of normal and curvature data by converting them to a series of one-dimensional histograms at the preprocessing step. In this step, surface normal, which contains nx, ny, and nz parameters, provide three histograms that represent the distribution of different point direction in our target point cloud. Thus, at the end of the preprocessing step, provided data, containing 3D volumes, histograms of the surface normal in three directions, and histogram of surface curvature are ready for the learning process.

Algorithm 1 Histogram Generator

Require: Input data I_d,Histogram size h_s
Ensure: Histogram array H

```
procedure Normalization(I)
    find min_I
    Decrease min_I from I
    find max_I
    Out = I / max_I
    return Out
end procedure
[N_d] = Normalization([I_d])
Ind_d = N_d × h_s
Ind_d = Interger(Ind_d)
L_h = Initialize an Array with Size(I_d) Width and h_s Height
Num = Make an array with values from 0 to Size(I_d)
Set Ind_list[Num, Ind_d] to 1
H = Summarize the Ind_list based on the value of Num in each row
```

An important preprocessing step is to convert the estimated data into a format which shows the overall changes of computed data and modify it to have less redundancy. Therefore, a histogram of one-dimensional data, an accurate representation of the distribution of the geometric features, could be applicable in this field. In other words, it can abstract the estimated data to comprehensive and trainable information. To process a dataset, we design an appropriate method to be agile and efficient. Algorithm 1 is designed to extract a histogram in a loopless and matrix calculation based method, which is operational in TensorFlow as well.

B. DEEP NEURAL NETWORK STRUCTURE

Most of the deep neural network is designed for processing images, in particular in learning methods such as detection and classification. However, there are many different methods to train and classification on any sorts of challenges, such as gesture recognition, 3D reconstruction, and others. In this

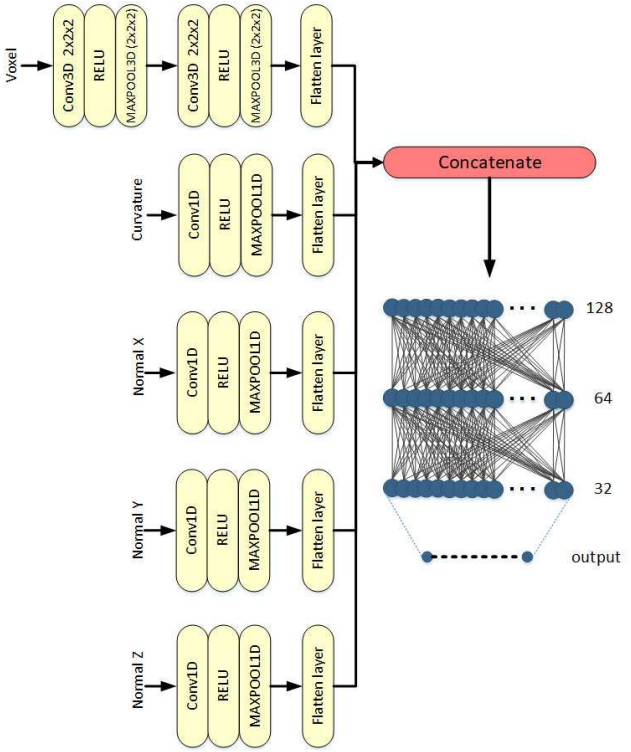


FIGURE 3. The network structure of the proposed approach containing five different convolutional networks that are concatenated and classified through a fully connected layer.

work, we explored many methods of 3D object recognition using a deep neural network to train a vast dataset and find an efficient and accurate approach. The designed CNN of the voxel data is contained two convoluted layers which are adapted to abstract the input data to more critical information. There are two different convolutional layers in the proposed network, which 1D and 3D Layers; 1D layers are applied to histogram data of the surface normal and curvature input arrays, and the 3D layer is used to refer on volumetric data. Besides, the results of these layers should be merged in 1 layer before the fully connected layers. Thus convolutional layers not only assisted us in decreasing the input size but also prepare the data to be ready for fully connected layers. FIGURE 3 represent each step of the suggested neural network in detail. It contains five CNNs which are concatenated and then connected to the fully connected layers before the output layer.

C. PARAMETERS ADJUSTMENT

To explore the parameter of voxelization, we conduct a series of experiments where use different grid size in each. The size of the histogram of normal and curvature information is changed relatively, and the processed data is recorded for each 3D model separately. For this investigation, the Shapenetcore-part dataset is used that contains 16 different categories divided to train, test, and validation, according to [16]. The evaluation demonstrates that increasing the grid resolution improves accuracy until the grid length of 8. If the grid size is increased, the more convolutional layers are required to keep the efficiency of the network. Additionally, the training level would need a high-performance processor and more memory when the grid size was increased. The comparison of FIGURE 4 shows us that the grid size with the value of 6 would be the best choice. It needs low memory, the training process is much faster, and the accuracy can be counted as one of the highest. On the other hand, increasing the grid size could seriously decrease our performance by importing many details of the 3D model, which is not necessary to be learned.

D. 3D MESH PART ANNOTATION

To verify the result of the recognition method, we apply an almost similar network to the challenge of 3D mesh part annotation, with some modification of the input data. The overall architecture of the part annotation approach is explained in FIGURE 5, which contains encoding 3D points to octree encoded binary arrays, the histogram of adjacency points' normal (in x, y, and z-axis), and the histogram of curvature as well. The octree encoding is a procedure of converting a point location to a series of binary data; This method by divide-and-concur procedure, divide the data space to some smaller area and search to find the area which includes our intended point. Having a point in one area leads the algorithm to move to that area and continue the procedure there. This procedure continues until the method reaches the sequence of the binary array that leads us to find the point's area. For example, if the process continues until six times, we have six binary arrays with a length of 3. With following these 6 arrays data, we get closer to the point location step by step that is helpful to learn a location by intended tolerance. In this approach, each stage provides a number, which is the selected area index in the model space. The algorithm saves the number in each step and converts them to binary data to be ready for the learning method, as shown in FIGURE 6.

This method is using almost the same procedure as the proposed recognizer does. However, the network uses a RNN instead of a convolutional network, as it is evident in FIGURE 7. Also, the encoding step has to concentrate on just a point of the mesh, not the entire mesh. This suggested method showed its efficiency, but to improve its power, we can add normal and curvature to increase the accuracy. This kind of supervised 3D part annotation has a variety of applications. One example could be [35] by Karambakhsh *et al.*, which was the first step of our collaboration with the medical school of Shanghai Jiao Tong University. Organizing a medical dataset and segmentation would be a significant application for our proposed method.

In the first step of the proposed approach, the position of the intended point should be encoded by the Octree method, which helps us to recognize the area of the point with our suggested resolution. To convert the point information to encoded binary arrays of the Octree, more than position data, we need to have a minimum and maximum

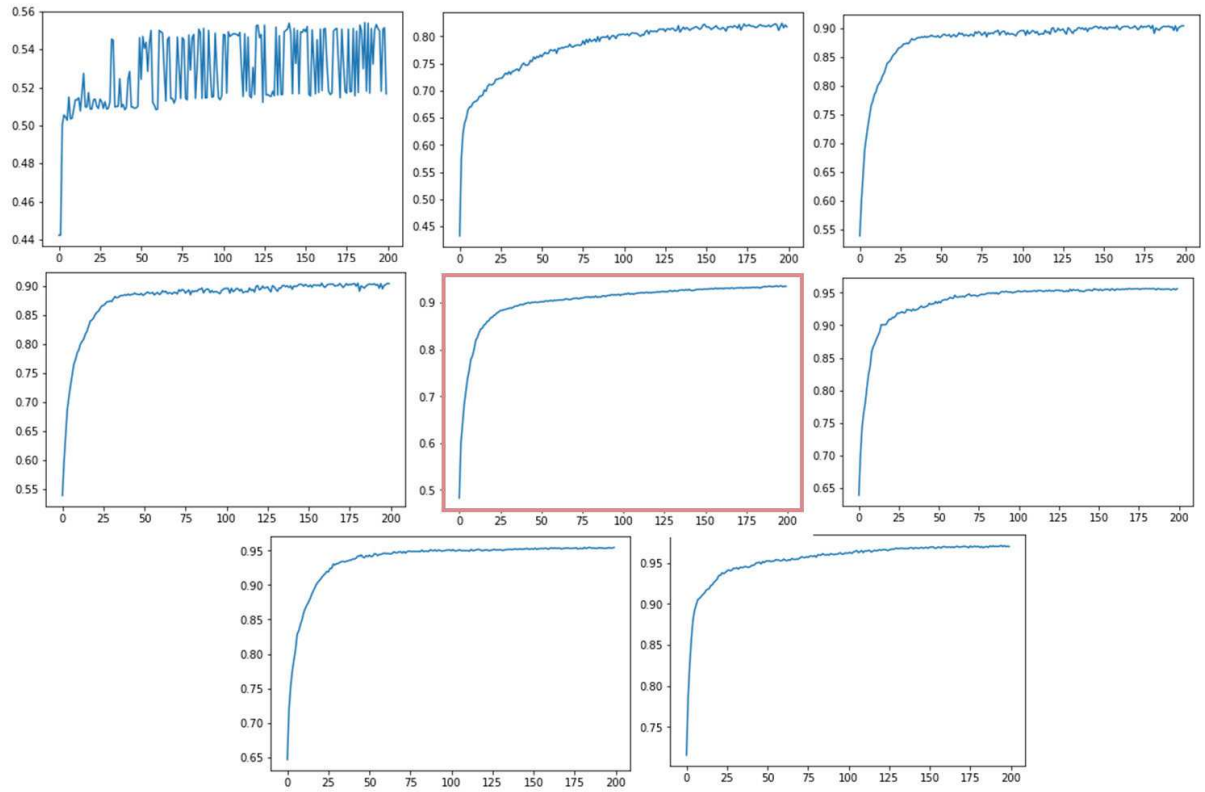


FIGURE 4. The accuracy results are shown by sending voxels with different grid sizes, from 2 to 9. It shows that the accuracy increases by a larger grid size, but it also affects the training speed. We find out that grid size with a length of 6 is more efficient than the other sizes by having less training time and high accuracy.

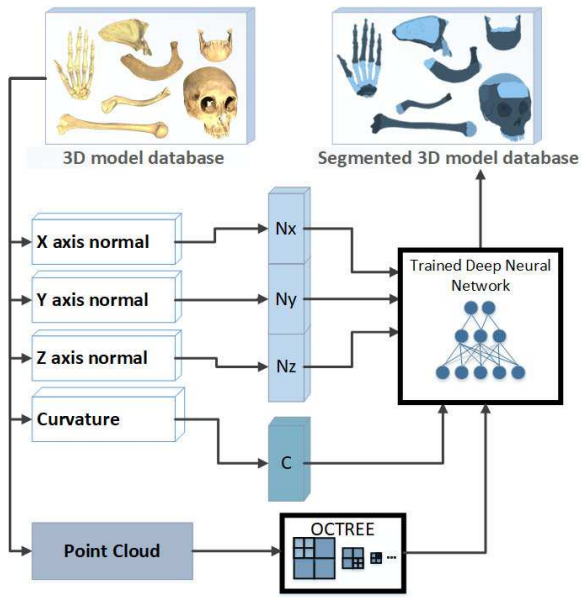


FIGURE 5. The overall procedure of 3D object part-recognition via the proposed approach. In the train level, the network should learn each 3D model parts separately.

of the 3D model in x, y, and z-axis. To be scale-invariant, we require to normalize the position of vertices as well. Since the encoded data is sequential and distinguishable, the method can use an RNN for learning the point position. In this paper, we use the same encoding method to recognize the location of the point. However, the result showed that having less encoding steps ends to having less location accuracy, and more encoding steps lead to be time-consuming. Finally, in our approach, it is decided to use less encoding step but some extra local features to recognize the shapes accurately.

As mentioned in the first section, using point normal and curvature provides us with an ability to recognize 3D models in a variety of datasets. Since each point of the 3D models has surface normal and curvature, they can be used in the 3D part annotation as well. Finally, by matching the acquired part's histogram with the selected category's histogram, the result of recognition is evaluated. FIGURE 8 showed the procedure of verifying the recognition result by part annotation of the selected object. FIGURE 8(a) shows the sample Shapenetcore-part objects in four categories, and FIGURE 8(b) is the categorized result after applying the classification method. The histogram of parts can be almost distinguishable for each category; thus, the suggested part segmentation can help us to check whether parts are similar to the selected category or not. Similarly, FIGURE 8(c) demonstrates a fault in classification, where part-histogram is not matching to the expected histogram of the category.

TABLE 1. The accuracy of the proposed network by a different combination of data. Employing only voxel data to the suggested network achieved adequate performance for the suggested environments. But to increase the accuracy, the combination of the surface normal and curvature is proposed.

Input			Accuracy
Volumetric data	Point's Normal	Curvature	
✓			90.1%
	✓		69.0%
		✓	44.3%
✓	✓		91.8%
✓		✓	90.3%
	✓	✓	70.1%
✓	✓	✓	92.1%

TABLE 2. Classification accuracy of our suggested approach in comparison with the state of the arts. The results are divided into 3 different types of Image-based, feature-based, and volumetric approaches.

Types	Methods	Accuracy
Image based approach	Cao et al. [36]	91
	MVCNN-MultiRes [10]	90.01
	MVCNN [8]	88.93
	ShapePFCN [37]	88.4
Feature based approach	ShapeBoost [38]	83
	Guo et al. [39]	82.2
Volumetric approach	OctNet [1]	88.03
Mixed approach	Ours	92.1

E. INTERACTIVE TRAINING

One of the crucial challenges in different environments is updating the database without consuming too much time and effort. In most of the approaches, a network can learn a dataset with a static number of inputs, but there is a possibility in the suggested approach that assists us to cover this request. The proposed method can receive the data part by part, due to our input generator settings. There are two types of training in the suggested approach; (i)offline training that receives all the dataset at the first step and (ii) active training that can apply a new entry to the pre-trained network. The active training makes an opportunity to add a new scanned 3D object and recognize it, the same as other existing objects in the dataset. For sure, there is a risk of decreasing the accuracy of the network if the operator added a wrong entry. The structure of our interactive training is shown in FIGURE 9. The method loads the input from a dataset directory gradually, which almost affects training speed, but it makes the technique able to receive data and continue training even after convergence. We should make sure that the new input information has to be in the right structure and similar to our origin input data. This ability assists researchers in pedagogical environments to update their databases frequently.

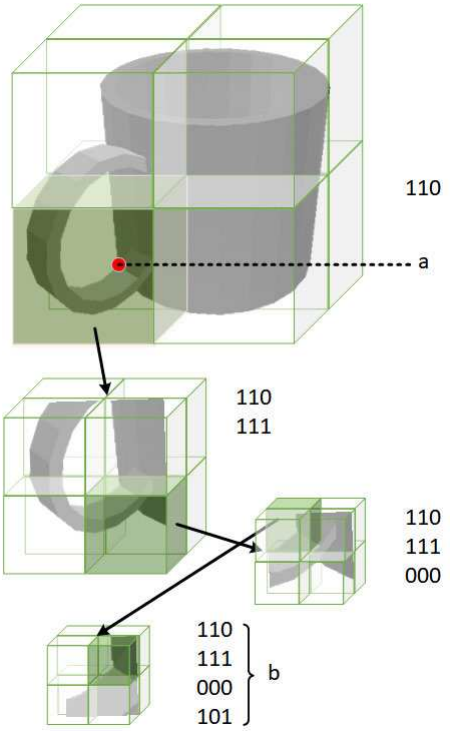


FIGURE 6. Applying the octree method on a 3D cup model from Shapenetcore-part. (a) shows the selected point to be a series of binary coded array. (b) the coded array, which is useful to localize the selected point's position.

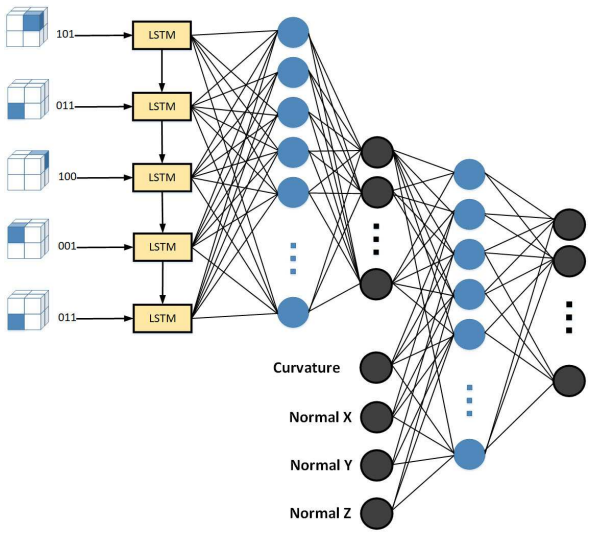


FIGURE 7. The overview of the proposed approach to 3D mesh part annotation. The network structure is a combination of a recurrent neural network with some extra nodes, such as surface normal and surface curvature. The input data are the sample encoded 3D models in 5 levels.

IV. EXPERIMENTAL RESULTS

The comparison of the proposed approach with the other state of the arts shows undeniable progress in the 3D mesh recognition issue. As we already discussed, the main suggestion in this paper is an acceptable and accurate method for classifying 3D models. However, we have achieved an almost precise method for part annotation by modifying the process and applying a nearly similar procedure and network. For comparison, we used the proposed method on the Shapenetcore-part dataset, which is an almost large dataset with 16 different categories. TABLE 1 demonstrates that the

TABLE 3. The results that the proposed network has shown. These numbers are the mean of the output layer on each category.

	airpla	bag	cap	car	chair	earpho	guitar	knife	lamp	laptop	bike	mug	pistol	rocket	skatebo	table
airpla	**0.94**	0	0	0	0	0	0	0	0	0	0	0	0.01	0.01	0.04	0.02
bag	0.01	**0.87**	0.01	0.01	0.01	0	0.01	0.01	0.02	0.01	0.04	0.04	0	0	0.07	0
cap	0	0	**0.73**	0	0	0.01	0.01	0.03	0	0.01	0.13	0.01	0.06	0.02	0.03	0
car	0.03	0.01	0.01	**0.49**	0	0.05	0.01	0.25	0.09	0.04	0.01	0.01	0.01	0.02	0.07	0.02
chair	0.03	0.01	0.01	0	**0.7**	0	0.01	0.09	0.02	0.06	0	0	0.01	0	0.1	0.03
earpho	0	0	0.01	0	0	**0.96**	0	0.02	0	0.01	0	0.01	0.02	0.01	0.01	0
guitar	0.01	0	0	0	0	0	**0.96**	0.04	0.01	0	0	0	0	0.01	0.03	0
knife	0	0	0	0	0.01	0	0	**0.97**	0	0.01	0	0	0	0	0.02	0.01
lamp	0	0	0	0.01	0	0.01	0	0.01	**0.94**	0.02	0.01	0.01	0	0.03	0.03	0.01
laptop	0	0	0.02	0	0.04	0.05	0.01	0.05	0.04	**0.55**	0.02	0	0.01	0.06	0.16	0.04
bike	0	0.01	0	0.01	0	0	0	0.01	0.01	0.01	**0.99**	0	0	0.01	0.03	0
mug	0	0.01	0	0	0	0	0	0.01	0	0	0.01	**0.89**	0.03	0.01	0.08	0.02
pistol	0	0	0.03	0	0.01	0.1	0	0.04	0	0.02	0.02	0.02	**0.67**	0.03	0.07	0.02
rocket	0.01	0	0	0	0	0.01	0.01	0.02	0	0	0	0	0	**0.96**	0.01	0
skatebo	0	0	0	0	0	0	0	0	0	0.01	0.01	0	0	0.02	**0.93**	0.04
table	0.02	0	0.01	0.01	0	0.01	0	0.01	0	0	0	0.01	0.01	0.01	0.29	**0.72**

TABLE 4. The accuracy result of the proposed part annotator network that is the highest accuracy in comparison with the state of the arts. The table shows the accuracy in different object categories separately.

	Voxel CNN	ACNN	Yi et al. 2016	Qi et al. 2016	Yi et al. 2016	Pointnet++ 2017	ours
knife	79.58	81.98	86.1	85.9	85.4	85.9	**88.43**
aero	75.14	76.35	81.6	**83.4**	81	82.4	75.6
cap	73.28	70.8	81.9	82.5	77.7	**87.7**	83.4
lamp	74.43	77.43	**84.7**	80.8	82.5	83.7	73.31
guitar	88.35	87.84	**93**	91.5	92	91	88.63
mug	91.79	89.49	92.7	93	91.9	94.1	**96.51**
skate	65.25	82.05	82.9	72.8	69.8	76.4	**83.82**
rocket	51.16	49.23	60.6	57.9	53.1	58.7	**69.17**
pistol	76.41	77.41	81.6	81.2	**85.9**	81.3	80.35
laptop	93.92	95.49	95.6	95.3	**95.7**	95.3	94.5
earphone	63.5	71.14	74.9	73	61.9	71.8	**75.38**
motor	58.67	45.68	66.7	65.2	70.6	71.6	**72.69**
bag	72.8	72.89	81.7	78.7	78.4	79	**83.08**
car	70	72.72	75.2	74.9	75.7	**77.3**	73.9
chair	87.17	86.12	90.2	89.6	87.6	**90.8**	83.95
mean	74.76	75.77	**81.96**	80.38	79.28	81.8	81.51

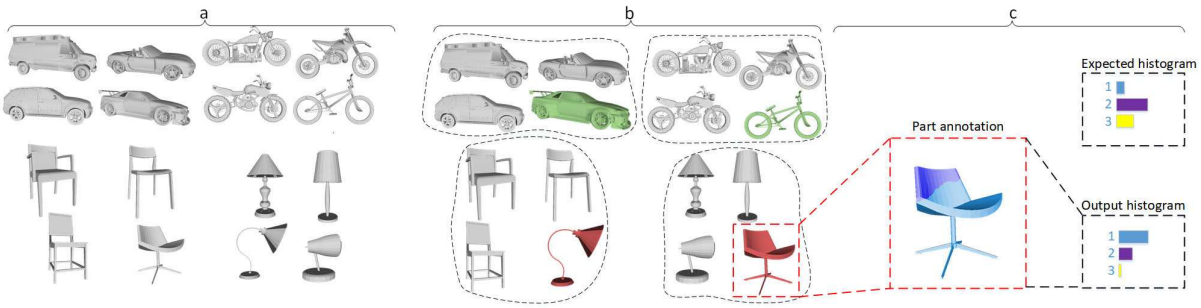


FIGURE 8. Overview of the proposed approach to verify the recognition result by using part annotation method. (a) shows four categories of the Shapenetcore-part dataset include car, bike, chair, and lamp, which used as the ground truth of our approach. (b) shows 4 samples of categorization results that two of them are selected in the right category (green objects), and two are wrong (red objects). (c) shows the histogram matching between the expected lamp histogram and the output of the recognition method.

investigation of the recognition with different options and inputs, which is done by proposed neural networks. TABLE 1 clearly shows the progress of recognition accuracy after concatenating networks. The precise result of the network just by using volumetric data verifies the power of the proposed system. However, by adding two more parameters, which are point normal and curvature, the more exciting results are achieved. The advantages of the method are not limited to its accuracy but also training speed faster than some of the popular methods in the same issue. Using matrix-based calculation made preprocessing part enormously faster than ordinary loop-based implementation. Also, the network structure is simple, which means the number of nodes and layers is less than many other methods as well. TABLE 2 shows the comparison table in terms of recognition accuracy and based on categorization on Shapenetcore-part.

Dataset: The evaluation of the proposed approach is done on the accessible dataset of Shapenetcore-part contains

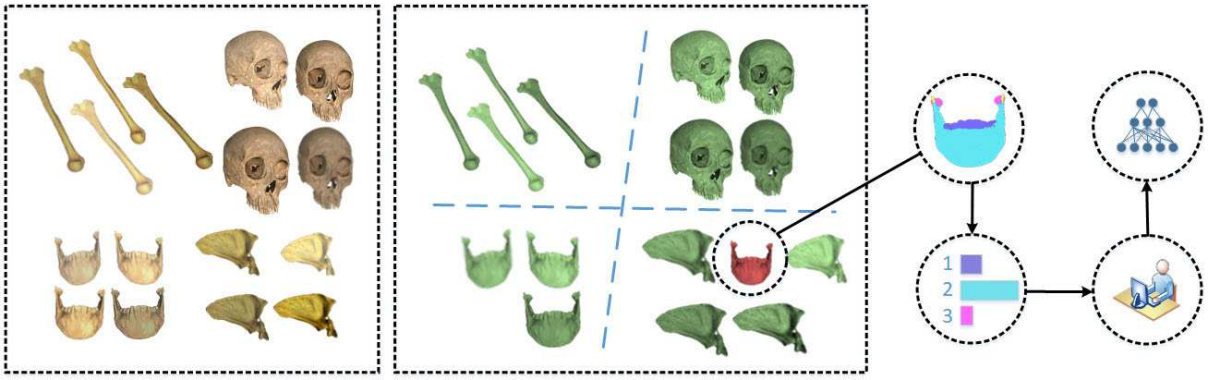


FIGURE 9. Structure of proposed interactive learning which allows the user to add a new entry, test the result on part histogram matching, and let the operator check the selected category to make sure of correctness.

18045 shapes across 16 categories with a different range of objects. We use the default train-test split [16], which allow us to compare our approach with another method in the same situation.

Device: the processing is done in a desktop-PC with Intel Core i7 4 GHz processor, 16 GB RAM, with 64-bit Windows 10. The ordinary PC is selected to show the lightweight deep network structure, which is optimized and efficient.

TABLE 2 collects the overall classification results of different methods on the Shapenetcore-part dataset. As the result shows, the proposed merged network result is superior in all the mentioned results. The comparison demonstrates the power of combined five networks into one. Besides, in comparison of image-based technics, such as MVCNN [8], the proposed approach does not require to render the model with different views. The only time-consuming task is preprocessing the data to extract voxel, surface normal, and surface curvature. However, our implementation using matrix calculation enabled this step to be as efficient as possible.

TABLE 3 shows the estimation of the proposed deep neural network in each category separately. Indeed, there are many objects in each category, and we show the mean of the results in this table. On the other hand, FIGURE 10 visualizes the list of recognition accuracy in both stages of the proposed method. The results of the first stage demonstrate that the method has done the expected classification, especially on objects that have enough details to distinguish them, and the recognizer supporting by part annotation shows more improvement in some of the categories. Although the part annotation method shows precise results, using them in recognition of Shapenetcore-part has not got a perfect result; this problem is according to the variety of shapes on the Shapenetcore-part dataset. Thus, when the part annotator stage is applied on body organs 3D models, they do not have many differences in one category, the results would be more enhanced.

The proposed idea, which is combining the volumetric data with the point normal and curvature information, works well on the recognition issue. But to improve the abilities of the proposed approach, we decided to apply a similar network that received octree coded information of a 3D point instead

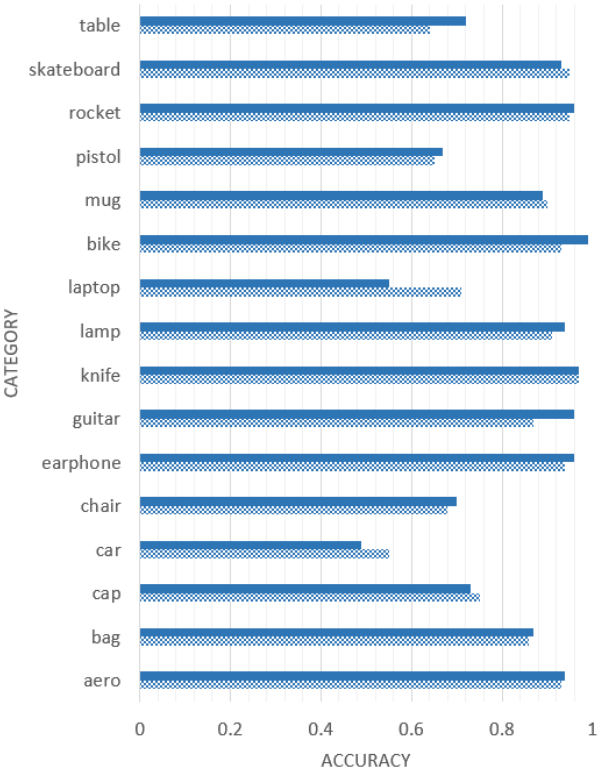


FIGURE 10. The recognition's accuracy of the suggested network on the Shappnetcore-part dataset. It shows the accuracy in different categories by horizontal bars. The top horizontal bar shows the recognition accuracy using just voxel, point normal, and point curvature. The bottom bar shows the results after applying the part annotator.

of 3D mesh voxel data to the part annotation task and use this second network to verify the recognition results. TABLE 4 shows a forward step in terms of part annotation accuracy of Shapenetcore-part. The proposed approach contains a combination of an RNN on the position data and simple MLP layers on normal and curvature information that end to reach more accurate results on part annotation issue on the Shapenetcore-part dataset. A comparison of the results confirms that the number of objects in each category is essential. If there were many different shapes in a class and it has variety in shapes,

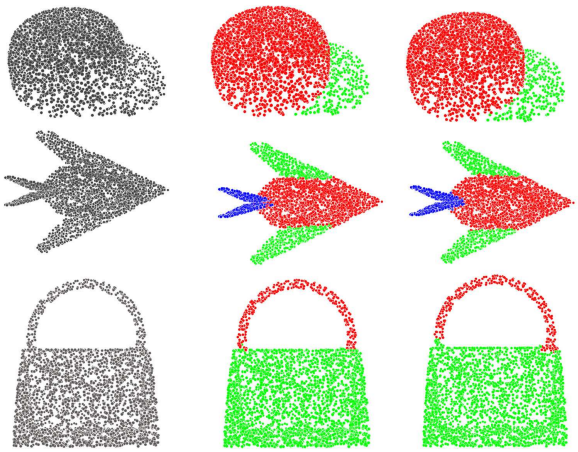


FIGURE 11. The part-annotation results in 3 categories of Shapenetcore-part dataset. The left column in each category demonstrates the original point cloud, the middle one represents the ground truth, and the right column is our part annotation result.

the accuracy would be moderate. Conversely, if the number were lower than an acceptable amount for training, the network would not be able to distinguish the category correctly. Therefore, the best network results are in classes that have an adequate number of objects. As it is shown in this section, the result of the proposed recognition network shows precise results, which are better than some of the top methods in this field. But to improve the recognition accuracy, we decided to use segmentation results to improve the classification result. Thus, the Shapenetcore-part has been used in this research that had already part annotated. Therefore, in this research, we also proposed to use an almost similar neural network model in terms of part annotation.

The result of our part annotation approach is shown in TABLE 4, which is a comparison with some other methods that used the same dataset, such as Voxnet [6], Pointnet++ [23], and others. We also demonstrate the part annotation results on the Shpaenetcore-part dataset in FIGURE 11, segmenting point cloud data using their trained category's network. The proposed network stage can segment a 3d model according to the selected category in the previous step (recognition stage). If the recognition stage classified the model in the correct category, the part annotation stage would be statistically matched to the selected category. Accordingly, in this paper, we show that these results can improve the recognition as well as the other 3D global features, and FIGURE 10 shows the results. On the other hand, many body-organ parts are scanning every day in terms of educating and investigating. Through an offline learning method, we would not be able to recognize them, except that the learning procedure train the whole dataset one more time. For this purpose, the interactive stage is added to this method to be able to continue training by new entries without starting from the first. This method requires an interactive interface and an open-loop learning procedure, which is correctly done in the suggested approach.

V. CONCLUSION AND FUTURE WORK

In this paper, we have suggested a combination of deep neural networks that can categorize the 3D object dataset and verify the results by statistical histogram from the part annotation. Our approach relies on two essential features of the 3D data, surface normal and curvature; the first one leverages direction variation while the second one concentrates on changes through the point cloud's surface. On the other hand, as the main feature of our model, we have used voxel information that naturally can remove noisy points. The experimental result shows that the proposed method is competitive with the most well-known approaches on the Shapenetcore-part dataset. One of our plans is to find a learning method that can achieve 3D features automatically and replace them with point normal and curvature parameters. Also, AutoEncoder networks have shown exciting results in terms of extracting 3D features from a 3D object, which encourages us to continue this research on methods that can recognize features using such the network in an optimized way. On the other hand, one of the preprocessing steps which can be attractive to 3D vision researchers is estimating the transformation of 3D objects that is one of our targets to be investigated as our next step.

REFERENCES

[1] G. Riegler, A. O. Ulusoy, and A. Geiger, "OctNet: Learning deep 3D representations at high resolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6620–6629.
[2] Z.-Q. Zhao, P. Zheng, S.-T. Xu, and X. Wu, "Object detection with deep learning: A review," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 11, pp. 3212–3232, Nov. 2019.
[3] W. Czajewski and K. Kołomyjec, "3D object detection and recognition for robotic grasping based on RGB-D images and global features," *Found. Comput. Decis. Sci.*, vol. 42, no. 3, pp. 219–237, Sep. 2017.
[4] R. Socher, B. Huval, B. Bhat, C. D. Manning, and A. Y. Ng, "Convolutional-recursive deep learning for 3D object classification," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 656–664.
[5] R. Beserra Gomes, B. M. Ferreira da Silva, L. K. D. M. Rocha, R. V. Aroca, L. C. P. R. Velho, and L. M. G. Gonçalves, "Efficient 3D object recognition using foveated point clouds," *Comput. Graph.*, vol. 37, no. 5, pp. 496–508, Aug. 2013.
[6] D. Maturana and S. Scherer, "VoxNet: A 3D convolutional neural network for real-time object recognition," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Sep. 2015, pp. 922–928.
[7] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, and J. Xiao, "3D ShapeNets: A deep representation for volumetric shapes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1912–1920.
[8] H. Su, S. Maji, E. Kalogerakis, and E. Learned-Miller, "Multi-view convolutional neural networks for 3D shape recognition," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 945–953.
[9] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, pp. 1–14, Sep. 2014.
[10] C. R. Qi, H. Su, M. NieBner, A. Dai, M. Yan, and L. J. Guibas, "Volumetric and multi-view CNNs for object classification on 3D data," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 5648–5656.
[11] V. Hegde and R. Zadeh, "FusionNet: 3D object classification using multiple data representations," *CoRR*, vol. abs/1607.05695, pp. 1–9, Nov. 2016.
[12] Z. Yang, Y. Li, J. Yang, and J. Luo, "Action recognition with Spatio–Temporal visual attention on skeleton image sequences," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 8, pp. 2405–2415, Aug. 2019.
[13] Y. Lin, F. Lv, S. Zhu, M. Yang, T. Cour, K. Yu, L. Cao, and T. Huang, "Large-scale image classification: Fast feature extraction and SVM training," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2011, pp. 1689–1696.

[14] C. Gan, N. Wang, Y. Yang, D.-Y. Yeung, and A. G. Hauptmann, "DevNet: A deep event network for multimedia event detection and evidence recounting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 2568–2577.

[15] G. Gkioxari and J. Malik, "Finding action tubes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 759–768.

[16] L. Yi, V. G. Kim, D. Ceylan, I.-C. Shen, M. Yan, H. Su, C. Lu, Q. Huang, A. Sheffer, and L. Guibas, "A scalable active framework for region annotation in 3D shape collections," *ACM Trans. Graph.*, vol. 35, no. 6, pp. 1–12, Nov. 2016.

[17] L. Yi, H. Su, X. Guo, and L. Guibas, "SyncSpecCNN: Synchronized spectral CNN for 3D shape segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6584–6592.

[18] Y. Zhou and O. Tuzel, "VoxelNet: End-to-end learning for point cloud based 3D object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4490–4499.

[19] S. Zhi, Y. Liu, X. Li, and Y. Guo, "Toward real-time 3D object recognition: A lightweight volumetric CNN framework using multitask learning," *Comput. Graph.*, vol. 71, pp. 199–207, Apr. 2018.

[20] J. Wu, C. Zhang, T. Xue, W. T. Freeman, and J. B. Tenenbaum, "Learning a probabilistic latent space of object shapes via 3D generative-adversarial modeling," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 82–90.

[21] D. Liang, K. Weng, C. Wang, G. Liang, H. Chen, and X. Wu, "A 3D object recognition and pose estimation system using deep learning method," in *Proc. 4th IEEE Int. Conf. Inf. Sci. Technol.*, Apr. 2014, pp. 401–404.

[22] R. Q. Charles, H. Su, M. Kaichun, and L. J. Guibas, "PointNet: Deep learning on point sets for 3D classification and segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 77–85.

[23] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "PointNet++: Deep hierarchical feature learning on point sets in a metric space," *CoRR*, vol. abs/1706.02413, pp. 1–14, Jun. 2017.

[24] H. Liu, Y. Cong, and Y. Tang, "Deep learning of volumetric representation for 3D object recognition," in *Proc. 32nd Youth Academic Annu. Conf. Chin. Assoc. Autom. (YAC)*, May 2017, pp. 663–668.

[25] A. Tejani, D. Tang, R. Kouskouridas, and T.-K. Kim, "Latent-class Hough forests for 3D object detection and pose estimation," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 462–477.

[26] J. Gui, T. Liu, D. Tao, Z. Sun, and T. Tan, "Representative vector machines: A unified framework for classical classifiers," *IEEE Trans. Cybern.*, vol. 46, no. 8, pp. 1877–1888, Aug. 2016.

[27] D. Tao, X. Lin, L. Jin, and X. Li, "Principal component 2-D long short-term memory for font recognition on single chinese characters," *IEEE Trans. Cybern.*, vol. 46, no. 3, pp. 756–765, Mar. 2016.

[28] M. Sepahvand, F. Abdali-Mohammadi, and F. Mardukhi, "Evolutionary metric-learning-based recognition algorithm for online isolated Persian/Arabic characters, reconstructed using inertial pen signals," *IEEE Trans. Cybern.*, vol. 47, no. 9, pp. 2872–2884, Sep. 2017.

[29] Q. Feng, C. Yuan, J.-S. Pan, J.-F. Yang, Y.-T. Chou, Y. Zhou, and W. Li, "Superimposed sparse parameter classifiers for face recognition," *IEEE Trans. Cybern.*, vol. 47, no. 2, pp. 378–390, Feb. 2017.

[30] A. Majumder, L. Behera, and V. K. Subramanian, "Automatic facial expression recognition system using deep network-based data fusion," *IEEE Trans. Cybern.*, vol. 48, no. 1, pp. 103–114, Jan. 2018.

[31] Y. Wang, Y. Sun, Z. Liu, S. E. Sarma, M. M. Bronstein, and J. M. Solomon, "Dynamic graph CNN for learning on point clouds," *CoRR*, vol. abs/1801.07829, pp. 1–14, Jan. 2018.

[32] J. Xie, G. Dai, F. Zhu, L. Shao, and Y. Fang, "Deep nonlinear metric learning for 3-D shape retrieval," *IEEE Trans. Cybern.*, vol. 48, no. 1, pp. 412–422, Jan. 2018.

[33] Y. Sun, B. Xue, M. Zhang, and G. G. Yen, "Completely automated CNN architecture design based on blocks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 4, pp. 1242–1254, Apr. 2020.

[34] S. Thrun, "Learning occupancy grid maps with forward sensor models," *Auto. Robots*, vol. 15, no. 2, pp. 111–127, Sep. 2003.

[35] A. Karambakhsh, A. Kamel, B. Sheng, P. Li, P. Yang, and D. D. Feng, "Deep gesture interaction for augmented anatomy learning," *Int. J. Inf. Manage.*, vol. 45, pp. 328–336, Apr. 2019.

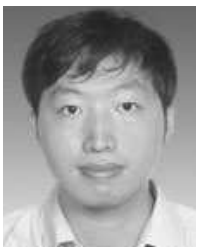
[36] Z. Cao, Q. Huang, and R. Karthik, "3D object classification via spherical projections," in *Proc. Int. Conf. 3D Vis. (3DV)*, Oct. 2017, pp. 566–574.

[37] E. Kalogerakis, M. Averkiou, S. Maji, and S. Chaudhuri, "3D shape segmentation with projective convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6630–6639.

[38] E. Kalogerakis, A. Hertzmann, and K. Singh, "Learning 3D mesh segmentation and labeling," *ACM Trans. Graph.*, vol. 29, no. 4, pp. 1–12, Jul. 2010.

[39] K. Guo, D. Zou, and X. Chen, "3D mesh labeling via deep convolutional neural networks," *ACM Trans. Graph.*, vol. 35, no. 1, pp. 1–12, Dec. 2015.

AHMAD KARAMBAKHSH is currently pursuing the Ph.D. degree with the Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China. He is also with the Visual Media and Data Management Laboratory, Department of Computer Science and Engineering, Shanghai Jiao Tong University. His current research interests include image/video processing, mixed reality, object recognition, depth estimation, and deep neural networks.



BIN SHENG (Member, IEEE) received the B.A. degree in English and the B.Eng. degree in computer science from the Huazhong University of Science and Technology, Wuhan, China, the M.Sc. degree in software engineering from the University of Macau, Taipa, Macau, and the Ph.D. degree in computer science and engineering from The Chinese University of Hong Kong, Hong Kong. He is currently an Associate Professor with the Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China. His current research interests include virtual reality and computer graphics. He is also an Associate Editor of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY.

PING LI (Member, IEEE) received the Ph.D. degree in computer science and engineering from The Chinese University of Hong Kong, Hong Kong. He is currently a Research Assistant Professor with The Hong Kong Polytechnic University, Hong Kong. He has excellent research project reported worldwide by ACM TechNews. He has one image/video processing national invention patent. His current research interests include image/video stylization, artistic rendering and synthesis, and creative media.

PO YANG (Senior Member, IEEE) received the B.Sc. degree in computer science from Wuhan University, Wuhan, China, the M.Sc. degree in computer science from the University of Bristol, Bristol, U.K., and the Ph.D. degree in electronic engineering from the Staffordshire University, Stoke-on-Trent, U.K. He is currently a Senior Lecturer of large scale data fusion with the Department of Computer Science, The University of Sheffield, Sheffield, U.K. He holds a strong tracking of high-quality publications and research experiences. He has published over 40 articles. His current research interests include the Internet of Things, RFID and indoor localization, pervasive health, image processing, GPU, and parallel computing.

YOUNHYUN JUNG received the B.Sc. degree in computer science from Inha University, Incheon, South Korea, in 2008, and the Ph.D. degree in computer science from The University of Sydney, Sydney, Australia, in 2016. From 2007 to 2010, he worked as a Software Engineer with Samsung Electronics. He is currently a Postdoctoral Research Fellow in computer science with The University of Sydney. His current research interests are in volume rendering and multimodal medical image visualization.

DAVID DAGAN FENG (Life Fellow, IEEE) received the M.Eng. degree in electrical engineering and computer science (EECS) from Shanghai Jiao Tong University, Shanghai, China, in 1982, and the M.Sc. degree in biocybernetics and the Ph.D. degree in computer science from the University of California at Los Angeles (UCLA), Los Angeles, CA, USA, in 1985 and 1988, respectively. He is currently the Head of the School of Information Technologies, the Director of the Biomedical and Multimedia Information Technology Research Group, and the Research Director of the Institute of Biomedical Engineering and Technology, The University of Sydney, Sydney, Australia. More importantly, however, is that many of his research results have been translated into solutions to real-life problems and have made tremendous improvements to the quality of life for those concerned. He has been invited to give over 100 keynote presentations in 23 countries and regions. He has published over 700 scholarly research articles, pioneered several new research directions, and made a number of landmark contributions in his field. He is a Fellow of the Australian Academy of Technological Sciences and Engineering. He received the Crump Prize for Excellence in Medical Engineering from UCLA. He has served as the Chair for the International Federation of Automatic Control (IFAC) Technical Committee on Biological and Medical Systems. He has organized/chaired over 100 major international conferences/symposia/workshops.