International Journal of Advance Research in Science and Engineering
Vol. No.6, Issue No. 08, August 2017
www.ijarse.com
IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

3D Convolutional Neural Network for Object Recognition

Rahul Dev Singh[1], Ajay Mittal[2], Rajesh K. Bhatia[1]

[1] PEC University of Technology, Chandigarh-160 012,(India)

[2] UIET, Panjab University, Chandigarh-160 036, (India)

ABSTRACT

3D Object recognition is an important task in computer vision applications. After the success of convolutional neural networks for object recognition in 2D images, many researchers have designed convolution neural network (CNN) for 3D object recognition. The state of art methods provide favourable results. However, the availability of large/dynamic 3D dataset and computational complexity of CNN are the biggest challenge in 3D CNN. In this paper, a model for object recognition problem using volumetric data representation has been proposed. The aim of this paper is to improve CNN architecture for volume based 3D objects. We implemented two separate CNN architectures and tested them on ModelNet datasets, which represent data in the form of CAD models. We compare our results with VoxNet, which is a state-of-art recognition method.

Keywords: 3D Image, Convolutional neural network, Deep learning, Voxel Grid

I. INTRODUCTION

Object Recognition is an important part of modern intelligent machines and systems. It is used in many applications related to multiple fields such as character recognition for mail sorting service [1], traffic monitoring [2], surveillance for security purposes [3], self-driving vehicle [4], human behaviour analysis [5], and medical imaging [6]. Object recognition has been performed generally using volumetric parts (i.e. generalized cylinders, geons and super-quadrics) [7], appearance based (i.e. edges, lines, corners) [8], pattern recognition (i.e. SIFT, HOG) [9], graph-based [10], and learning based methods [11]. Although the techniques mentioned above are still used in many applications, after the success of the award winning deep learning architecture alexnet [12], Convolutional Neural Networks (CNN) become ubiquitous. In recent years, deep learning has achieved outstanding results in 2D object recognition [13]. These results motivate researchers to apply deep learning methods for 3D object recognition, and some of them provide better results in comparison to state-of-art (SoA) approaches.

After the easy availability of a stereo camera and range-based sensors, the research in 3D object recognition is growing. The success of CNNs on object recognition using 2D images motivated researchers to uplift the CNN architecture for 3D objects. In the context of 3D data, CNN has been used on motion-based data taking time as the third dimension. However, research in 3D object recognition is restricted to certain applications. Many researchers have extended CNNs to process 3D object using RGBD data [14], [15]. These approaches do not contain the full geometric information of an object and make it difficult to combine information between different viewpoints. Volumetric and multi-view based CNNs are two methods for 3D data that gained popularity in recent years. In contrast, this paper proposes a CNN-based 3D object recognition approach that can recognise 3D objects from their 3D volumetric representations, and compare their accuracy on different voxel size. In literature, existing approaches of 3D CNNs use 3D point cloud data as training data [16] or RGBD image to build 3D

ISBN: 978-93-86171-60-3

CNNs, but CNNs can also be applied directly to recognise the voxel of 3D volumetric representation of objects. In this paper, we present object recognition on 3D CAD based model.

The rest of paper is outlined as follows: Section II explains related work of voxel-based 3D object recognition. Dataset, method, and experiment results are presented respectively in Section III, IV, V. Finally, conclusions and future work are discussed in Section VI.

II. RELATED WORK

3D Shape Descriptors: Modern 3D object recognition models have its origin in the 1960s. The earliest recognition frameworks were created on geometry-based models [7]. However, most of the recognition works rely on other handcrafted feature descriptors, such as Point Feature Histograms, 3D Shape Context, or Spin Images. CNN for 3D recognition has first used for RGB-D images, where depth is treated as an additional input channel [17]. The depth based approaches are conceptually very similar to 2D based recognition methods. The depth is just used as the fourth channel in CNN, or 3D synthetic models are used as input to CNN. However, using depth channel along with colour channels produce 2.5D, and it does not provide full geometric information of objects. In recent years, the researcher tried better shape descriptors based on multi-view [18], and volumetric information [19].

Convolutional Neural Networks: The CNNs have been designed for 2D data such as images and audio signals. It is widely used in various computer vision tasks and data science. The reason behind the acceptance of CNN in 2D image based tasks is due to the availability of large benchmark datasets, and these large datasets help to generate better image descriptor in comparison to handcrafted features, which provides better results. The similar approach for feature learning is used in this paper for 3D data instead of 2D data.

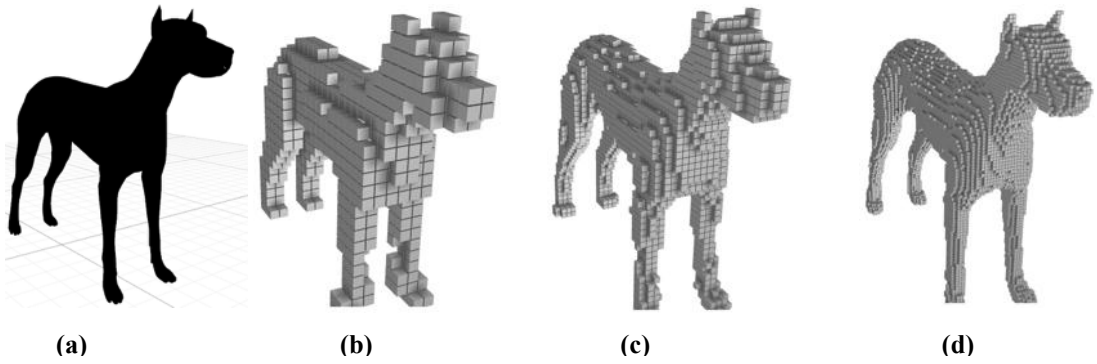

Fig. 1. Voxel representation of a CAD model. (a) CAD model, (b) 32^3 voxel, (c) 64^3 voxel, (d) 128^3 voxel.

Convolutional Neural Networks on 3D data: In recent years, with the availability of range based and 3D CAD datasets, researchers have additional information of objects in the image in comparison to the 2D image. 3D data is represented by motion [20], multi-view [18] and volume based methods in deep networks. Motion based CNN architectures are used successfully in action recognition task. Wu et al. [19] developed first CNN architecture for volume based data, namely 3DShapeNets.

ISBN: 978-93-86171-60-3

3DShapeNets used a Deep Belief Network to represent geometric 3D shapes as a probability distribution of binary variables on a 3D voxel grid. They use their method for shape completion from depth maps, too. The ModelNet dataset was introduced along with their work. However, our work is inspired by VoxNet, which is designed by Maturana & Sherer [21]. The VoxNet is composed of a simple but effective CNN, accepting as input voxel grids similar to Wu et al. 3D ShapeNets and VoxelNet provide state-of-art results. The other remarkable works are the panoramic-view proposed by Shi et al. [22], and the multi-view CNN proposed by Su et al. [18] that use multiple views of the same object from different angles in CNN.

III. DATBASE

Several 3D datasets have become available in recent years. However, these datasets are not large as the ImageNet[1] dataset that contains 2D images. There are many 3D datasets with reasonable size. Mostly 3D datasets are point-cloud based that obtained from range-scanners, such as the ModelNet[2], Sydney Urban Objects[3], SUN-3D[4]. The ModelNet dataset is used for training and testing of proposed CNN models. ModelNet dataset contains 127,915 CAD 3D images from 662 different object classes. ModelNet40, a subset of ModelNet, is a benchmark for 3D object recognition and used which further splits in 9,843 training and 2,468 test images. We also test our network on ModelNet10 in our experiments. ModelNet10 has ten object categories, and it is a subset of ModelNet40.

IV. METHOD

Object recognition in 3D CNN is done, by selecting most similar features from targeted classes. The object recognition process can be divided into two parts, namely data representation of 3D objects and training of CNN on represented data. We used 3D volumetric data representation in proposed architecture. Most researchers used voxel or point cloud methods for volumetric data representation. We used voxel based data representation in our CNNs. The voxels are generated with the help of binary occupancy grid. The number of models provided by dataset are less in comparison to the 2D dataset. To exploit network, we provide rotation of 30° along the gravity to the model. All the voxels are generated from provided mesh models in ModelNet dataset after rendering them to 12 different orientations. The proposed network is modified version of Voxnet. We used two different networks for training purpose.

Network-1:

Network-1 consists three convolutional layers and two fully connected layers. The network is shown in figure-2. Most researchers used voxel size equal or less than 32×32×32. However, we believe that 32×32×32 pixel per object is very less for accurate prediction of an object. In 2D image tasks, experiments show that less than 227×227 reolution is not good for object recognition task in 2D images. So to exploit the network, we use

1 http://www.image-net.org/
2 http://modelnet.cs.princeton.edu/
3 http://www.acfr.usyd.edu.au/papers/SydneyUrbanObjectsDataset.shtml
4 http://sun3d.cs.princeton.edu/

ISBN: 978-93-86171-60-3

three different sizes of voxels 32×32×32, 64×64×64, and 128×128×128 sizes. Many pixels in voxel represent empty cells that generate a lot of unnecessary matrix multiplication (zero valued matrix) cost. To reduce this cost, we use the kernel of 5×5×5 instead 3×3×3 that is frequently used in volumetric CNNs. The ReLU, and max-pooling (size of 2×2×2) layers are used after convolution (layer) operation. The pooling layer is used to reduce the over-fitting of too many parameters. To avoid data over fitting, due to orientation and similar views, we use dropout layer before first fully connected layer with 0.5 probabilities.

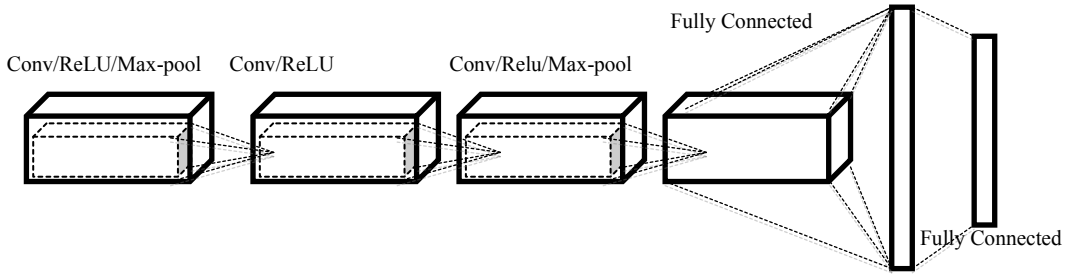

Fig. 2. The architecture of Network-1.

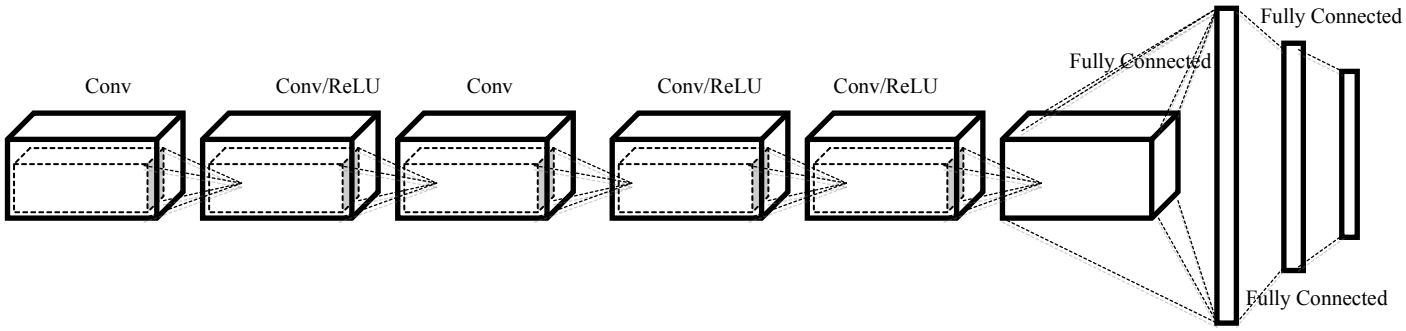

Fig. 3 The architecture of Network-2.

Network-2:

Network-2 inspired by the network architecture of AlexNet. It consists five convolutional layers and three fully connected layers. The network is shown in figure-3. Network-2 is also used multiple views of the voxel. We train network with 30-degree rotation. This network has all the settings as used in network-1 except pooling. To train network, we did not use pooling layers in the network-2. In our view, pooling can create ambiguity in shape of an object. Network-2 helps us to understand that how the depth of CNN affects the performance of recognition model.

ISBN: 978-93-86171-60-3

Type	Filter size/Dropout rate	Stride	Output size	Number of parameters
Convolution	5*5*5	1	32*28*28*28	4032
Max Pooling	2*2*2	-	32*14*14*14	-
Convolution	3*3*3	1	32*12*12*12	27,680
ReLU	-	-	32*12*12*12	-
Convolution	3*3*3	1	32*10*10*10	27,680
ReLU	-	-	32*10*10*10	-
Max Pooling	2*2*2	-	32*5*5*5	-
Dropout	0.5	-	32*5*5*5	-
Fully Con-nected	-	-	128	5,12,128
Fully Con-nected	-	-	40	5,160

Table 1: Details of Network-1

Type	Filter size/Dropout rate	Stride	Output size	Number of parameters
Convolution	5*5*5	1	32*28*28*28	4032
Convolution	3*3*3	1	32*26*26*26	27,680
ReLU	-	-	32*26*26*26	-
Convolution	3*3*3	1	32*24*24*24	27,680
Convolution	3*3*3	1	32*22*24*24	27,680
ReLU	-	-	32*22*24*24	-
Dropout	0.3	-	32*22*24*24	-
Convolution	3*3*3	1	32*22*22*22	27,680
ReLU	-	-	32*22*22*22	-
Dropout	0.4	-	32*22*22*22	-
Fully Con-	-	-	256	27,26,144

ISBN: 978-93-86171-60-3

nected				
Fully Connected	-	-	128	32,896
Fully Connected	-	-	40	5,160

Table 2: Details of Network-2

V. EXPERIMENTS

The ModelNet dataset is used for training and testing purpose in this network. All the 3D CAD images are converted in voxels using provided script by [19]. We test our networks for different size of voxels 32×32×32, 64×64×64, and 128×128×128. We implement both networks separately and compare their results. However, results of both networks have very less difference. We compare results for the same size of voxels for ModelNet10 and ModelNet40 datasets. Results show that higher resolution voxels improved the accuracy of recognition task. In comparison to VoxNet, our model gives better accuracy. However, VoxNet has less than 1 million parameters in its architecture while our network-2 has more than 2 million parameters. We train network-2 without pooling to test the effect of pooling on volumetric CNN. But results show that there is no significant effect of pooling layer on results. One of the reasons for no effect of pooling layer on performance is that the all the voxels used in training and testing has solid shape. There is not a single object model in the ModelNet dataset, which has the hollow object. The results are shown in figure 4, and figure 5 shows the loss and accuracy

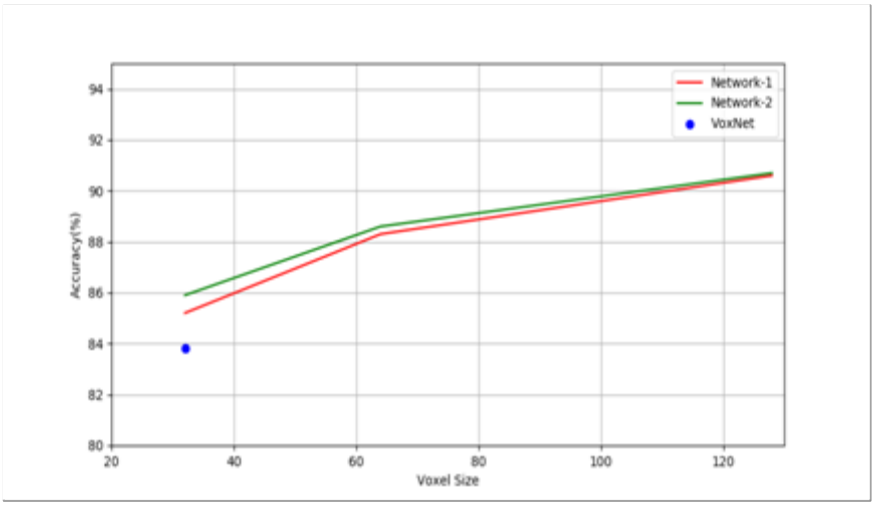

chart of network-1. All the experiments are done using NVidia Titan X Pascal GPU.

Fig. 4 Performance comparison of tested CNN networks on ModelNet40 dataset with different voxel resolution

ISBN: 978-93-86171-60-3

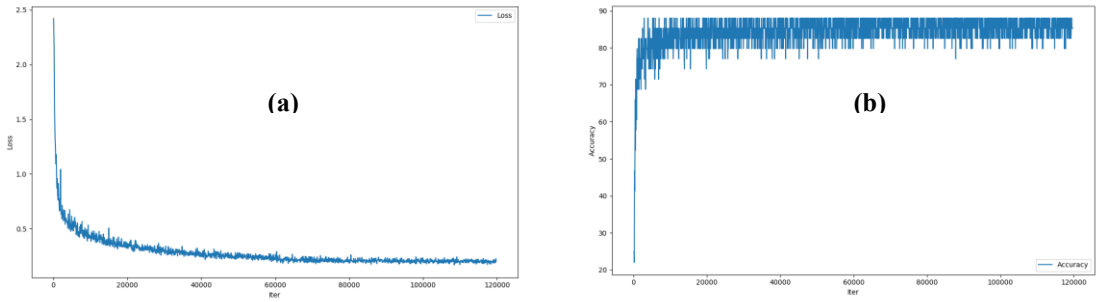

Fig. 5: Plot of Network-1 model: (a) loss, and (b) accuracy on training modelnet40 dataset.

VI. CONCLUSION

The paper describes the recognition task on 3D data using voxel based 3D data representation. The performance of CNN on different size of voxels has been analyzed. The analysis further motivated to design new networks and test different sizes of the voxel to find a suitable size for CNN operations. It can be concluded from the experimental results that size of a voxel has a direct impact on object recognition task. But on the other hand increasing the size of voxel creates a bottleneck in performance in volumetric CNN architecture. To overcome this problem an optimized data structure regarding time, as well as space, should be explored to handle the large size of voxels.

REFERENCES

[1] S. Basu, N. Das, R. Sarkar, M. Kundu, M. Nasipuri, and D. Kumar Basu, "A novel framework for automatic sorting of postal documents with multi-script address blocks," *Pattern Recognition*, vol. 43, no. 10, pp. 3507–3521, Oct. 2010.

[2] S. R. E. Datondji, Y. Dupuis, P. Subirats, and P. Vasseur, "A Survey of Vision-Based Traffic Monitoring of Road Intersections," *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 10, pp. 2681–2698, Oct. 2016.

[3] M. J. Gómez, F. García, D. Martín, A. de la Escalera, and J. M. Armingol, "Intelligent surveillance of indoor environments based on computer vision and 3D point cloud fusion," *Expert Systems with Applications*, vol. 42, no. 21, pp. 8156–8171, Nov. 2015.

[4] G. H. Lee, F. Faundorfer, and M. Pollefeys, "Motion Estimation for Self-Driving Cars with a Generalized Camera," in *2013 IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 2746–2753.

[5] T. Ko, "A survey on behavior analysis in video surveillance for homeland security applications," *Applied Imagery Pattern Recognition Workshop, 2008. AIPR '08. 37th IEEE*, pp. 1–8, 2008.

[6] A. P. James and B. V Dasarathy, "Medical image fusion: A survey of the state of the art," *Information Fusion*, vol. 19, pp. 4–19, Sep. 2014.

[7] A. Andreopoulos and J. K. Tsotsos, "50 Years of object recognition: Directions forward," *Computer Vision and Image Understanding*, vol. 117, no. 8, pp. 827–891, 2013.

[8] X. Gao, S. Member, Y. Su, X. Li, S. Member, and D. Tao, "A Review of Active Appearance Models," vol. 40, no. 2, pp. 145–158, 2010.

ISBN: 978-93-86171-60-3

[9] N. Dalal and W. Triggs, "Histograms of Oriented Gradients for Human Detection," *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition CVPR05*, vol. 1, no. 3, pp. 886–893, 2004.

[10] A. S. Mian, M. Bennamoun, and R. Owens, "Three-dimensional model-based object recognition and segmentation in cluttered scenes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 10, pp. 1584–1601, 2006.

[11] R. Poppe, "Vision-based human motion analysis: An overview," *Computer Vision and Image Understanding*, vol. 108, pp. 4–18, 2007.

[12] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," *Advances In Neural Information Processing Systems*, pp. 1097--1105, Feb. 2012.

[13] J. W. H. Tangelder and R. C. Veltkamp, "A survey of content based 3D shape retrieval methods," *Multimedia Tools and Applications*, vol. 39, no. 3, pp. 441–471, Sep. 2008.

[14] S. Gupta, P. Arbeláez, R. Girshick, and J. Malik, "Aligning 3D models to RGB-D images of cluttered scenes," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 07–12–June, pp. 4731–4740, 2015.

[15] M. Schwarz, H. Schulz, and S. Behnke, "RGB-D Object Recognition and Pose Estimation based on Pre-trained Convolutional Neural Network Features," *IEEE International Conference on Robotics and Automation (ICRA'15)*, no. May, pp. 1329–1335, 2015.

[16] A. Aldoma, Z. C. Marton, F. Tombari, W. Wohlkinger, C. Potthast, B. Zeisl, R. Rusu, S. Gedikli, and M. Vincze, "Tutorial: Point cloud library: Three-dimensional object recognition and 6 DOF pose estimation," *IEEE Robotics and Automation Magazine*, vol. 19, no. 3, pp. 80–91, 2012.

[17] S. Gupta, P. Arbeláez, R. Girshick, and J. Malik, "Indoor Scene Understanding with RGB-D Images: Bottom-up Segmentation, Object Detection and Semantic Segmentation," *International Journal of Computer Vision*, vol. 112, no. 2, pp. 133–149, Apr. 2015.

[18] H. Su, S. Maji, E. Kalogerakis, and E. Learned-Miller, "Multi-view Convolutional Neural Networks for 3D Shape Recognition," in *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 945–953.

[19] Zhirong Wu, S. Song, A. Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and J. Xiao, "3D ShapeNets: A deep representation for volumetric shapes," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1912–1920.

[20] S. Ji, M. Yang, K. Yu, and W. Xu, "3D convacolutional neural networks for human action recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 1, pp. 221–31, 2013.

[21] D. Maturana and S. Scherer, "VoxNet: A 3D Convolutional Neural Network for Real-Time Object Recognition," *Iros*, pp. 922–928, 2015.

[22] B. Shi, S. Bai, Z. Zhou, and X. Bai, "DeepPano: Deep Panoramic Representation for 3-D Shape Recognition," *IEEE Signal Processing Letters*, vol. 22, no. 12, pp. 2339–2343, 2015.