IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received July 28, 2019, accepted August 7, 2019, date of publication August 12, 2019, date of current version August 23, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2934650

Deeply Exploiting Long-Term View Dependency for 3D Shape Recognition

YONG XU[1,2,3], CHAODA ZHENG[1], RUOTAO XU[1], AND YUHUI QUAN[1,4]

[1]School of Computer Science and Engineering, South China University of Technology, Guangzhou 510000, China
[2]Peng Cheng Laboratory, Shenzhen 518000, China
[3]Communication and Computer Network Laboratory of Guangdong, Guangzhou 510000, China
[4]Guangdong Provincial Key Laboratory of Computational Intelligence and Cyberspace Information, Guangzhou 510000, China

Corresponding author: Yuhui Quan (csyhquan@scut.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61672241, Grant 61602184, Grant 61872151, and Grant U1611461, in part by the Natural Science Foundation of Guangdong Province under Grant 2016A030308013 and Grant 2017A030313376, in part by the Science and Technology Program of Guangzhou under Grant 201707010147 and Grant 201802010055, in part by the Guangdong Provincial Engineering and Technology Research Center of Big Data Analysis and Processing under Grant 20140904-160, and in part by the Fundamental Research Funds for the Central Universities under Grant x2js-D2181690.

ABSTRACT Recognition of 3D shapes is a fundamental task in computer vision. In recent years, view-based deep learning has emerged as an effective approach for 3D shape recognition. Most existing view-based methods treat the views of an object as an unordered set, which ignores the dynamic relations among the views, *e.g.* sequential semantic dependencies. In this paper, modeling the views of an object by a sequence, we aim at exploiting the long-term dependencies among different views for shape recognition, which is done by constructing a sequence-aware view aggregation module based on the bi-directional Long Short-Term Memory network. It is shown that our view aggregation module not only captures the bi-directional dependencies in view sequences, but also enjoys the robustness to circular shifts of input sequences. Incorporating the aggregation module into a standard convolutional network architecture, we develop an effective method for 3D shape classification and retrieval. Our method was evaluated on the ModelNet40/10 and ShapeNetCore55 datasets. The results show the encouraging performance gain from exploiting long-term dependencies in view sequences, as well as the superior performance of our method compared to the existing ones.

INDEX TERMS 3D shape recognition, long-term dependency, multi-view deep learning, view aggregation.

I. INTRODUCTION

Understanding 3D objects has been a fundamental problem since the establishment of computer vision, with a broad spectrum of applications including multimedia [1], augmented reality [2], [3], entertainment [4], robotics [5], [6], autonomous driving [7]–[10], 3D reverse engineering [11], [12], medical imaging [13], [14], and monitoring [15]. Due to the limited availability of 3D data, the early works focus on either theories of 3D representation or methods of image-based object recognition. Until recent years, 3D data has become ubiquitous with the rapid development of 3D object acquisition hardware (*e.g.* 3D scanners) as well as 3D modeling software (*e.g.* computer graphics), opening up opportunities to practical 3D object recognition. Compared to the 2D cases in images, 3D objects contain additional cues of 3D shapes, which are very essential and crucial for visual understanding. Therefore, a majority of existing works on 3D object recognition are devoted to building 3D shape features for recognition (*i.e.* 3D shape recognition).

Inspired by the great success of deep learning in image classification [16], [17], many approaches (*e.g.* [18]–[22]) to 3D shape recognition have been proposed based on neural networks (NNs). These approaches consume different formats of 3D data in NNs, such as point clouds, voxelized data, and multiple views of objects. Among these approaches, the ones that accept multiple views of objects as NN's input, which are called view-based approaches, have shown much more encouraging results than other types of approaches; see *e.g.* [21]–[25]. Besides, view-based approaches have

The associate editor coordinating the review of this article and approving it for publication was Aysegul Ucar.