Information Sciences 366 (2016) 188–201

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

3D object understanding with 3D Convolutional Neural Networks

Biao Leng*, Yu Liu, Kai Yu, Xiangyang Zhang, Zhang Xiong

School of Computer Science & Engineering, Beihang University, Beijing, 100191, PR China

ARTICLE INFO

Article history:
Received 15 June 2014
Revised 26 June 2015
Accepted 9 August 2015
Available online 12 August 2015

Keywords:
Deep representation
3D Convolutional Neural Networks
3D object understanding

ABSTRACT

Feature engineering plays an important role in object understanding. Expressive discriminative features can guarantee the success of object understanding tasks. With remarkable ability of data abstraction, deep hierarchy architecture has the potential to represent objects. For 3D objects with multiple views, the existing deep learning methods can not handle all the views with high quality. In this paper, we propose a 3D convolutional neural network, a deep hierarchy model which has a similar structure with convolutional neural network. We employ stochastic gradient descent (SGD) method to pretrain the convolutional layer, and then a back-propagation method is proposed to fine-tune the whole network. Finally, we use the result of the two phases for 3D object retrieval. The proposed method is shown to out-perform the state-of-the-art approaches by experiments conducted on publicly available 3D object datasets.

1. Introduction

The rapid development in computation ability, graphics devices and mobile-networks has made it more practicable to make use of the 3D object information. The key of 3D object utilization is 3D object retrieval and recognition, and effective algorithms for them are increasingly demanded.

For the recognition task in machine learning, feature engineering is a crucial step. The ability of discovering and expressing the underlying concept of the raw data is one of the important factors that determine overall performance of the algorithm. Therefore, our goal is to find features that represent the 3D objects discriminatively [4]. In recent decades, plenty of 3D model descriptors have been proposed. These descriptors can be generally divided into model-based and view-based methods [19]. Model-based methods, such as geometric moment [8], surface distribution [56] and volumetric descriptors, capture the features of objects from the object-level perspective. View-based methods, on the other hand, consider each 3D object as a collection of 2D projections from several fixed viewpoints [2]. Since 2D views contain rich information, these methods are flexible and effective for 3D object retrieval. Extensive researches [44,48,51,55,66,73] have dedicated their efforts to the view-based methods for 3D object representation, and many reports [9,67] have illustrated the advantages of view-based methods over the model-based ones. Furthermore, view-based methods can be improved by adopting hybrid descriptors, which combine several feature descriptors together, such as DESIRE [71], MADE [42] and CMVD [14].

However, these methods have not achieved satisfactory performance yet due to the immature feature detection techniques. On the other hand, some recent works [63] have illustrated that deep architectures, such as restricted Boltzmann machines

* Corresponding author. Tel.: +86 1082338177.
E-mail address: lengbiao@buaa.edu.cn, lengbiao@gmail.com (B. Leng).

http://dx.doi.org/10.1016/j.ins.2015.08.007
0020-0255/© 2015 Elsevier Inc. All rights reserved.

(RBM), deep belief networks (DBN) [3], deep Boltzmann machines (DBM) [64,65], and stacked autoencoder [69,70] have powerful representational capacities to approximate the distributions of the input data. Further more, models with the ability to detect local features (deep convolutional neural network) are proved to be better than fully connected models (DBN and DBM) [35]. With the powerful representation ability, deep learning is widely studied and successfully applied in many fields, such as 2D image processing [35], speech recognition [13] and natural language processing [12].

For view-based 3D object retrieval methods, each object contains multiple views. The general idea is to handle each view separately and then deal with the results together. During this procedure, the information with regard to some correlations among different views may be lost. Therefore, the method with the ability to handle all the views of the 3D object simultaneously is urgently needed.

In this paper, we propose a deep hierarchy architecture for 3D object understanding. It has a similar structure with convolutional neural networks (CNN). We call it *3D Convolutional Neural Networks (3DCNN)*. Like CNN, it has several convolutional layers and subsampling layers, but the difference lies in the complexity. It is more complex because it handles multiple views at the same time and take the interaction among them into consideration. In order to guarantee the reasonable sequence of the input views, we sort the views before they are fed to the network. We also employ a stacked layer-wise method to pre-train the 3DCNN. When training one convolutional layer, we consider the layer as a convolutional auto-encoder [54], which can be trained with stochastic gradient descent (SGD). A well pre-trained 3DCNN is faster to converge and more likely to extract high-level abstract features,proving that the pre-training step is efficient for 3D object retrieval. After pre-training, we employ the back-propagation method for supervised learning with labeled samples. Some tricks including ‘dropout’ are applied in both pre-training and supervised learning. Experiments are conducted using publicly available 3D datasets. We not only compare this method with those in state-of-the-arts, but also explore the influence of the tricks as well as the effectiveness of the pre-training phase during the supervised learning phase.

The architecture of this paper is organized as follows. In Section 2, we give a brief review of related works on 3D object representation. Section 3 introduces details about 3D object understanding using 3DCNN. The experimental results are shown in Section 4. Finally, we draw the conclusion in Section 5.

2. Related work

In recent decades, the field of 3D model understanding has attracted great attention of a variety of communities, and a lot of related works [2,8,14,15,40,43,45–47] have been proposed.

2.1. 3D object feature detection

In feature detection, the crucial step of 3D object retrieval and recognition, discriminative information of the objects is extracted to represent them. Generally speaking, these methods for 3D object feature extraction are divided into model-based and view-based approaches.

Model-based methods detect the features directly from the original 3D object with the topological and geometric information [15,20,56,61]. Topological methods, statistics-based approaches and geometry-based algorithms are all model-based ones. Topological descriptors [6,8,61,74] are suitable for nonrigid models. Patane et al. [61] presents a minimal contouring algorithm for rapid computation of the Reeb graph. Statistics-based methods [20,53,56,59] sample features by the objects’ characteristics, such as distance between points and surface information, etc. Osada et al. [56] proposes a novel method called shape distribution to describe the signature of 3D models. Park et al. [59] develops a sliced image histogram to represent PCA normalized 3D objects. The geometric-based ones [15,18,34,36,58,72] include instinctive surface characteristics extraction and mathematical representations like the spherical harmonics. Daras et al. [15] applies voxelization on 3D models before extracting radial and spherical information via a radial integration transform (RIT) and a spherical integration transform (SIT). A probability-density based shape descriptor, in which some local features can be extracted via kernel density estimation coupling with Gaussian transform, is proposed by Akgul et al. [1].

View-based methods [2,10,11,60,62,66] first capture some 2D views from the original object with some fixed view points and then regard these view images as the information of the object. We can exploit some mature methods in image processing to process these view images and extract discriminative characteristics from them. View-based methods have attracted great attention due to their flexibility and good performance. The original work is accomplished by Chen et al. [11] who proposes the light field descriptor (LFD). Twenty cameras are placed on the vertices of a regular dodecahedron, taking hundreds of black and white projection images. Papadakis et al. proposes the PANORAMA descriptor [57] which can calculate accurate model attribution for panoramic object representation. Adaptive views clustering (AVC) method [2] selects the best characteristic views from more than 320 views. An algorithm derived from K-means and a Bayesian information criteria is employed to estimate the extent to which the characteristic views fit the data. A generated view method is introduced in [20], where the model information is employed to generate circular images for 3D model representation. In this method, the generated views can ensure the 3D model spatial information perform better in 3D model retrieval tasks. To learn the relevance of 3D objects, Gao et al. proposes the first learning-based method [23], in which 3D objects are formulated within a hypergraph structure, and the learning on multiple hypergraphs is conducted to estimate the object relevance. This method employs the hypergraph structure to avoid the distance measure between two groups of views, which yields a better result than state-of-the-art methods. The CMVD method [14] extracts a set of 2D rotation-invariant shape descriptors from both the binary images and depth images. It is different from