PointNet: A 3D Convolutional Neural Network for Real-Time Object Class Recognition

A. Garcia-Garcia*, F. Gomez-Donoso†, J. Garcia-Rodriguez*, S. Orts-Escolano*, M. Cazorla†, J. Azorin-Lopez*
*Department of Computer Technology, University of Alicante
{agarcia, jgarcia, sorts, jazorin}@dtic.ua.es
†Department of Computer Science and Artificial Intelligence, University of Alicante
fgomez@dccia.ua.es, miguel.cazorla@ua.es

Abstract—**During the last few years, Convolutional Neural Networks are slowly but surely becoming the default method to solve many computer vision related problems. This is mainly due to the continuous success that they have achieved when applied to certain tasks such as image, speech, or object recognition.**

Despite all the efforts, object class recognition methods based on deep learning techniques still have room for improvement. Most of the current approaches do not fully exploit 3D information, which has been proven to effectively improve the performance of other traditional object recognition methods.

In this work, we propose *PointNet*, a new approach inspired by *VoxNet* and *3D ShapeNets*, as an improvement over the existing methods by using density occupancy grids representations for the input data, and integrating them into a supervised Convolutional Neural Network architecture.

An extensive experimentation was carried out, using *ModelNet* – a large-scale 3D CAD models dataset – to train and test the system, to prove that our approach is on par with state-of-the-art methods in terms of accuracy while being able to perform recognition under real-time constraints.

I. Introduction

Object class recognition is one of the main challenges for a computer to be able to understand a scene. This turns out to be a key capability for autonomous robots which operate in real-word environments. Due to the unstructured nature of those environments, mobile robots need to do reasoning grounded in the real world, i.e., they must understand the information provided by their sensors. In this regard, semantic object recognition is a crucial task, among other equally important ones, which robots have to perform in order to achieve a considerable level of scene understanding [1].

The advent of reliable and low-cost range sensors, e.g., RGB-D cameras such as Microsoft Kinect and PrimeSense Carmine, revolutionized the field by providing useful 3D data to feed the prediction systems with a new dimension of useful information. Because of that, traditional 2D approaches were superseded by 3D-based ones. In addition, 3D model and scene repositories are growing on a daily basis, thus providing researchers with enough reliable data sources for training and testing object recognition systems.

The vast majority of 3D object recognition methods [2] are typically based on hand-crafted local feature descriptors [3]. These kinds of approaches rely on specific pipelines [4] consisting of a keypoint detection phase, followed by the computation of descriptors at those characteristic regions, finally they are classified to determine the possible object represented by those descriptors. That classification is performed by using distance metrics or machine learning algorithms, e.g., Support Vector Machines (SVMs) [5], random forests [6], or neural networks, which are trained with object datasets. Despite the popularity and the continuous flow of successful implementations of this kind of pipelines – specially for recognition in cluttered and occluded environments – handcrafting feature descriptors requires domain expertise and remarkable engineering and theoretical skills, and even fulfilling both requirements they are still far from perfection. In this regard, the aim of researchers has been to replace those hand-engineered descriptors with multilayer neural networks able to learn them automatically by using general-purpose training algorithms. This gave birth to the application of Convolutional Neural Networks (CNNs) to image analysis. A brand-new deep learning architecture designed to process data in form of multiple arrays, which quickly surpassed the previous methods with many practical successes [7] – mainly due to the fact that they were easy to train and generalized far better – so that CNNs have become the community standard to tackle the object class recognition problem [8]. However, only a few CNN-based systems have been proposed to use 3D data for this purpose, therefore we strongly believe that there is still room for improvement since most of them do not fully exploit 3D data.

In this work, we propose a new deep learning architecture for object class recognition using pure 3D information, an approach inspired by the success of recently introduced CNN-based systems for 3D object recognition. Its contribution is twofold: a novel way for representing the input data, which is based on point density occupancy grids, and its integration into a CNN specifically tuned for the aforementioned purpose.

This paper is structured as follows. Section II reviews the state of the art of deep learning methods applied to 3D object class recognition. Next, our proposal – namely *PointNet* – is described in Section III. Section IV defines the methodology followed to test our approach, as well as the experiments that were carried out, their results, and the corresponding discussion. At last, Section V draws some conclusions about our work and sets future research lines to improve the proposal.

978-1-5090-0620-5/16/$31.00 ©2016 IEEE

II. Related Work

Deep learning architectures have recently revolutionized 2D object class recognition. The most significant example of such success is the CNN architecture, being *AlexNet* [7] the milestone which started that revolution. Krizhevsky *et al.* developed a deep learning model based on the CNN architecture that outperformed by a large margin (15.315 % error rate against the 26.172 % scored by the runner-up not based on deep learning) state-of-the-art methods on the *ImageNet* [9] LSVRC 2012 challenge.

Due to the successful applications of the CNNs to 2D image analysis, several researchers decided to take the same approach for 2.5D data, treating the depth channel as an additional one together with the RGB ones [10]–[12]. These methods simply extend the architecture to take four channels – matrices – as input instead of the three featured by RGB images. This is still a image-based approach which does not fully exploit the geometric information of 3D shapes despite its straightforward implementation. Apart from 2.5D approaches, specific architectures to learn from volumetric data, which make use of pure 3D convolutions, have been recently developed. Those architectures are commonly referred as 3DCNNs and their foundations are the same as the 2D or 2.5D ones, but the nature of the input data is radically different. Since volumetric data is usually quite dense and hard to process, most of the successful 3DCNNs resort to a more compact representation of the 3D space: the occupancy grids. *VoxNet* [13] and *3D ShapeNets* [14] make extensive use of this representation.

Those 3DCNNs are slowly overtaking other approaches when applying object recognition to complete 3D scenes [15]. This progress has been mainly enabled by two factors: the substantial growth in the number of 3D models available online through repositories, and the reduction of training times thanks to frameworks and libraries which exploit the power of massively parallel architectures for this kind of tasks. On the one hand, there exist many collections of 3D models, but they tend to be small and usually lack annotations and other useful information for training this kind of deep architectures. In contrast, 2D approaches have taken advantage of the numerous and high-quality datasets that already exist such as *ImageNet* [9], *LabelMe* [16], and *SUN* [17]. During the last years, researchers have unified efforts to create large-scale annotated 3D datasets inspired by the success of the 2D counterparts. The most popular 3D datasets which have revamped data-driven solutions – for computer vision in general, and object recognition in particular – are the *Princeton ModelNet* [14], and *ShapeNets* [18] datasets. On the other hand, the creation of deep learning frameworks such as *Caffe* [19], *Theano* [20], *Torch* [21], or *TensorFlow* [22], which allow researchers to easily express and launch their architectures and accelerate the training calculations with Graphics Processing Units (GPUs) by using CUDA or OpenCL, has enabled quick prototyping and testing. Both facts have turned out to be crucial for the development of the field.

III. Approach

The proposed system takes a point cloud of an object as an input and predicts its class label. In this regard, the proposal is twofold: a volumetric grid based on point density to estimate spatial occupancy inside each voxel, and a pure 3DCNN which is trained to predict object classes. The occupancy grid – inspired by *VoxNet* [13] occupancy models based on probabilistic estimates – provides a compact representation of the object's 3D information from the point cloud. That grid is fed to the CNN architecture, which in turn computes a label for that sample, i.e., predicts the class of the object.

This architecture was implemented using the Point Cloud Library (PCL) [23] – which contains state-of-the-art algorithms for 3D point cloud processing – and *Caffe* [19], a deep learning framework developed and maintained by the Berkeley Vision and Learning Center (BVLC) and an active community of contributors on *GitHub*[1]. This BSD-licensed C++ library allows us to design, train, and deploy CNN architectures efficiently, mainly thanks to its drop-in integration of NVIDIA *cuDNN* [24] to take advantage of GPU acceleration.

A. Occupancy Grid

Occupancy grids [25] are data structures which allow us to obtain a compact representation of the volumetric space. They stand between meshes or clouds, which offer rich but large amounts of information, and voxelized representations with packed but poor information. At that midpoint, occupancy grids provide considerable shape cues to perform learning, while enabling an efficient processing of that information thanks to their array-like implementation.

Recent 3D deep learning architectures make use of occupancy grids as a representation for the input data to be learned or classified. *3D ShapeNets* [14] is a Convolutional Deep Belief Network (CDBN) which represents a 3D shape

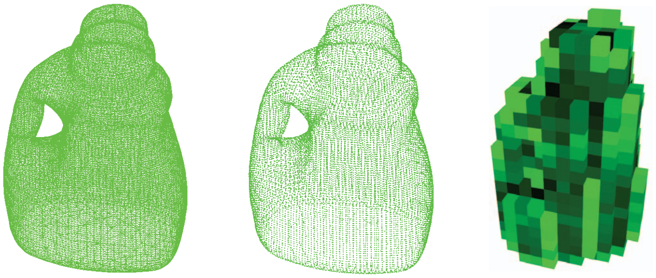
(a) Mesh (b) Cloud (c) Occ. Grid

Fig. 1: A mesh (a) is transformed into a point cloud (b), and that cloud is processed to obtain a voxelized occupancy grid (c). The occupancy grid shown in this figure is a cube of $30 \times 30 \times 30$ voxels. Each voxel of that cube holds the point density inside its volume. In this case, dark voxels indicate high density whilst bright ones are low density volumes. Empty voxels were removed for better visualization.

[1] http://github.com/BVLC/caffe