UNIVERSITY OF ALICANTE

PHD THESIS

TBD

Author
Alberto GARCIA-GARCIA

Advisors
Jose GARCIA-RODRIGUEZ
Sergio ORTS-ESCOLANO

A thesis submitted in fulfilment of the requirements for the degree of Doctor of Philosophy

in the

3D Perception Lab
Department of Computer Technology

February 21, 2019

This document was proudly made with LaTeX and Ti*k*Z.

"Will robots inherit the earth? Yes, but they will be our children."

Marvin Minsky

Abstract

Aliquam lectus. Vivamus leo. Quisque ornare tellus ullamcorper nulla. Mauris porttitor pharetra tortor. Sed fringilla justo sed mauris. Mauris tellus. Sed non leo. Nullam elementum, magna in cursus sodales, augue est scelerisque sapien, venenatis congue nulla arcu et pede. Ut suscipit enim vel sapien. Donec congue. Maecenas urna mi, suscipit in, placerat ut, vestibulum ut, massa. Fusce ultrices nulla et nisl.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Nulla ac nisl. Nullam urna nulla, ullamcorper in, interdum sit amet, gravida ut, risus. Aenean ac enim. In luctus. Phasellus eu quam vitae turpis viverra pellentesque. Duis feugiat felis ut enim. Phasellus pharetra, sem id porttitor sodales, magna nunc aliquet nibh, nec blandit nisl mauris at pede. Suspendisse risus risus, lobortis eget, semper at, imperdiet sit amet, quam. Quisque scelerisque dapibus nibh. Nam enim. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Nunc ut metus. Ut metus justo, auctor at, ultrices eu, sagittis ut, purus. Aliquam aliquam.

Resumen

Sed commodo posuere pede. Mauris ut est. Ut quis purus. Sed ac odio. Sed vehicula hendrerit sem. Duis non odio. Morbi ut dui. Sed accumsan risus eget odio. In hac habitasse platea dictumst. Pellentesque non elit. Fusce sed justo eu urna porta tincidunt. Mauris felis odio, sollicitudin sed, volutpat a, ornare ac, erat. Morbi quis dolor. Donec pellentesque, erat ac sagittis semper, nunc dui lobortis purus, quis congue purus metus ultricies tellus. Proin et quam. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos hymenaeos. Praesent sapien turpis, fermentum vel, eleifend faucibus, vehicula eu, lacus.

Sed feugiat. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Ut pellentesque augue sed urna. Vestibulum diam eros, fringilla et, consectetuer eu, nonummy id, sapien. Nullam at lectus. In sagittis ultrices mauris. Curabitur malesuada erat sit amet massa. Fusce blandit. Aliquam erat volutpat. Aliquam euismod. Aenean vel lectus. Nunc imperdiet justo nec dolor.

Aliquam lectus. Vivamus leo. Quisque ornare tellus ullamcorper nulla. Mauris porttitor pharetra tortor. Sed fringilla justo sed mauris. Mauris tellus. Sed non leo. Nullam elementum, magna in cursus sodales, augue est scelerisque sapien, venenatis congue nulla arcu et pede. Ut suscipit enim vel sapien. Donec congue. Maecenas urna mi, suscipit in, placerat ut, vestibulum ut, massa. Fusce ultrices nulla et nisl.

Acknowledgements

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor. Proin tempus nibh sit amet nisl. Vivamus quis tortor vitae risus porta vehicula.

Fusce mauris. Vestibulum luctus nibh at lectus. Sed bibendum, nulla a faucibus semper, leo velit ultricies tellus, ac venenatis arcu wisi vel nisl. Vestibulum diam. Aliquam pellentesque, augue quis sagittis posuere, turpis lacus congue quam, in hendrerit risus eros eget felis. Maecenas eget erat in sapien mattis porttitor. Vestibulum porttitor. Nulla facilisi. Sed a turpis eu lacus commodo facilisis. Morbi fringilla, wisi in dignissim interdum, justo lectus sagittis dui, et vehicula libero dui cursus dui. Mauris tempor ligula sed lacus. Duis cursus enim ut augue. Cras ac magna. Cras nulla. Nulla egestas. Curabitur a leo. Quisque egestas wisi eget nunc. Nam feugiat lacus vel est. Curabitur consectetuer.

Suspendisse vel felis. Ut lorem lorem, interdum eu, tincidunt sit amet, laoreet vitae, arcu. Aenean faucibus pede eu ante. Praesent enim elit, rutrum at, molestie non, nonummy vel, nisl. Ut lectus eros, malesuada sit amet, fermentum eu, sodales cursus, magna. Donec eu purus. Quisque vehicula, urna sed ultricies auctor, pede lorem egestas dui, et convallis elit erat sed nulla. Donec luctus. Curabitur et nunc. Aliquam dolor odio, commodo pretium, ultricies non, pharetra in, velit. Integer arcu est, nonummy in, fermentum faucibus, egestas vel, odio.

Contents

List of Figures

List of Tables

List of Acronyms

3D three-dimensional

BVLC Berkeley Vision and Learning Center

CAD Computer Aided Design

CDBN Convolutional Deep Belief Network

CNN Convolutional Neural Network

GPU Graphics Processing Unit

OFF Object File Format

PCD Point Cloud Data

PCL Point Cloud Library

PLY Polygon File Format

Chapter 1

Introduction

Suspendisse vitae elit. Aliquam arcu neque, ornare in, ullamcorper quis, commodo eu, libero. Fusce sagittis erat at erat tristique mollis. Maecenas sapien libero, molestie et, lobortis in, sodales eget, dui. Morbi ultrices rutrum lorem. Nam elementum ullamcorper leo. Morbi dui. Aliquam sagittis. Nunc placerat. Pellentesque tristique sodales est. Maecenas imperdiet lacinia velit. Cras non urna. Morbi eros pede, suscipit ac, varius vel, egestas non, eros. Praesent malesuada, diam id pretium elementum, eros sem dictum tortor, vel consectetuer odio sem sed wisi.

Suspendisse vitae elit. Aliquam arcu neque, ornare in, ullamcorper quis, commodo eu, libero. Fusce sagittis erat at erat tristique mollis. Maecenas sapien libero, molestie et, lobortis in, sodales eget, dui. Morbi ultrices rutrum lorem. Nam elementum ullamcorper leo. Morbi dui. Aliquam sagittis. Nunc placerat. Pellentesque tristique sodales est. Maecenas imperdiet lacinia velit. Cras non urna. Morbi eros pede, suscipit ac, varius vel, egestas non, eros. Praesent malesuada, diam id pretium elementum, eros sem dictum tortor, vel consectetuer odio sem sed wisi.

Suspendisse vitae elit. Aliquam arcu neque, ornare in, ullamcorper quis, commodo eu, libero. Fusce sagittis erat at erat tristique mollis. Maecenas sapien libero, molestie et, lobortis in, sodales eget, dui. Morbi ultrices rutrum lorem. Nam elementum ullamcorper leo. Morbi dui. Aliquam sagittis. Nunc placerat. Pellentesque tristique sodales est. Maecenas imperdiet lacinia velit. Cras non urna. Morbi eros pede, suscipit ac, varius vel, egestas non, eros. Praesent malesuada, diam id pretium elementum, eros sem dictum tortor, vel consectetuer odio sem sed wisi.

Suspendisse vitae elit. Aliquam arcu neque, ornare in, ullamcorper quis, commodo eu, libero. Fusce sagittis erat at erat tristique mollis. Maecenas sapien libero, molestie et, lobortis in, sodales eget, dui. Morbi ultrices rutrum lorem. Nam elementum ullamcorper leo. Morbi dui. Aliquam sagittis. Nunc placerat. Pellentesque tristique sodales est. Maecenas imperdiet lacinia velit. Cras non urna. Morbi eros pede, suscipit ac, varius vel, egestas non, eros. Praesent malesuada, diam id pretium elementum, eros sem dictum tortor, vel consectetuer odio sem sed wisi.

Suspendisse vitae elit. Aliquam arcu neque, ornare in, ullamcorper quis, commodo eu, libero. Fusce sagittis erat at erat tristique mollis. Maecenas sapien libero, molestie et, lobortis in, sodales eget, dui. Morbi ultrices rutrum lorem. Nam elementum ullamcorper leo. Morbi dui. Aliquam sagittis. Nunc placerat. Pellentesque tristique sodales est. Maecenas imperdiet lacinia velit. Cras non urna. Morbi eros pede, suscipit ac, varius vel, egestas non, eros. Praesent malesuada, diam id pretium elementum, eros sem dictum tortor, vel consectetuer odio sem sed wisi.

Suspendisse vitae elit. Aliquam arcu neque, ornare in, ullamcorper quis, commodo eu, libero. Fusce sagittis erat at erat tristique mollis. Maecenas sapien libero, molestie et, lobortis in, sodales eget, dui. Morbi ultrices rutrum lorem. Nam elementum ullamcorper leo. Morbi dui. Aliquam sagittis. Nunc placerat. Pellentesque tristique sodales est. Maecenas imperdiet lacinia velit. Cras non urna. Morbi eros pede, suscipit ac, varius vel, egestas non, eros. Praesent malesuada, diam id pretium elementum, eros sem dictum tortor, vel consectetuer odio sem sed wisi.

Suspendisse vitae elit. Aliquam arcu neque, ornare in, ullamcorper quis, commodo eu, libero. Fusce sagittis erat at erat tristique mollis. Maecenas sapien libero, molestie et, lobortis in, sodales eget, dui. Morbi ultrices rutrum lorem. Nam elementum ullamcorper leo. Morbi dui. Aliquam sagittis. Nunc placerat. Pellentesque tristique sodales est. Maecenas imperdiet lacinia velit. Cras non urna. Morbi eros pede, suscipit ac, varius vel, egestas non, eros. Praesent malesuada, diam id pretium elementum, eros sem dictum tortor, vel consectetuer odio sem sed wisi.

Suspendisse vitae elit. Aliquam arcu neque, ornare in, ullamcorper quis, commodo eu, libero. Fusce sagittis erat at erat tristique mollis. Maecenas sapien libero, molestie et, lobortis in, sodales eget, dui. Morbi ultrices rutrum lorem. Nam elementum ullamcorper leo. Morbi dui. Aliquam sagittis. Nunc placerat. Pellentesque tristique sodales est. Maecenas imperdiet lacinia velit. Cras non urna. Morbi eros pede, suscipit ac, varius vel, egestas non, eros. Praesent malesuada, diam id pretium elementum, eros sem dictum tortor, vel consectetuer odio sem sed wisi.

Suspendisse vitae elit. Aliquam arcu neque, ornare in, ullamcorper quis, commodo eu, libero. Fusce sagittis erat at erat tristique mollis. Maecenas sapien libero, molestie et, lobortis in, sodales eget, dui. Morbi ultrices rutrum lorem. Nam elementum ullamcorper leo. Morbi dui. Aliquam sagittis. Nunc placerat. Pellentesque tristique sodales est. Maecenas imperdiet lacinia velit. Cras non urna. Morbi eros pede, suscipit ac, varius vel, egestas non, eros. Praesent malesuada, diam id pretium elementum, eros sem dictum tortor, vel consectetuer odio sem sed wisi.

Suspendisse vitae elit. Aliquam arcu neque, ornare in, ullamcorper quis, commodo eu, libero. Fusce sagittis erat at erat tristique mollis. Maecenas sapien libero, molestie et, lobortis in, sodales eget, dui. Morbi ultrices rutrum lorem. Nam elementum ullamcorper leo. Morbi dui. Aliquam sagittis. Nunc placerat. Pellentesque tristique sodales est. Maecenas imperdiet lacinia velit. Cras non urna. Morbi eros pede, suscipit ac, varius vel, egestas non, eros. Praesent malesuada, diam id pretium elementum, eros sem dictum tortor, vel consectetuer odio sem sed wisi.

1.1 Motivation

1.2 Approach

1.3 Contributions

1.4 Co-Authored Papers

This thesis is the result of continuous effort throughout the last years. Such efforts have sometimes crystallized in form of co-authored publications and conference talks.

1.4.1 Chapter 2

- Alberto Garcia-Garcia, Francisco Gomez-Donoso, Jose Garcia-Rodriguez, et al. “PointNet: A 3D Convolutional Neural Network for real-time object class recognition”. In: *2016 International Joint Conference on Neural Networks, IJCNN 2016, Vancouver, BC, Canada, July 24-29, 2016*. 2016, pp. 1578–1584. DOI: `10.1109/IJCNN.2016.7727386`. URL: `https://doi.org/10.1109/IJCNN.2016.7727386`
- Alberto Garcia-Garcia, Jose Garcia-Rodriguez, Sergio Orts-Escolano, et al. “A study of the effect of noise and occlusion on the accuracy of convolutional neural networks applied to 3D object recognition”. In: *Computer Vision and Image Understanding* 164 (2017), pp. 124–134. DOI: `10.1016/j.cviu.2017.06.006`. URL: `https://doi.org/10.1016/j.cviu.2017.06.006`
- Francisco Gomez-Donoso, Alberto Garcia-Garcia, Jose Garcia-Rodriguez, et al. “LonchaNet: A Sliced-based CNN Architecture for Real-time 3D Object Recognition”. In: *2017 International Joint Conference on Neural Networks, IJCNN 2017, Anchorage, Alaska, May 14-19, 2017*. 2017. URL: `https://ieeexplore.ieee.org/document/7965883/`

1.4.2 Chapter 3

- Alberto Garcia-Garcia, Jose Garcia-Rodriguez, Sergio Orts-Escolano, et al. “A study of the effect of noise and occlusion on the accuracy of convolutional neural networks applied to 3D object recognition”. In: *Computer Vision and Image Understanding* 164 (2017), pp. 124–134. DOI: `10.1016/j.cviu.2017.06.006`. URL: `https://doi.org/10.1016/j.cviu.2017.06.006`
- Alberto Garcia-Garcia, Pablo Martinez-Gonzalez, Sergiu Oprea, et al. “The RobotriX: An eXtremely Photorealistic and Very-Large-Scale Indoor Dataset of Sequences with Robot Trajectories and Interactions”. In: *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2018, pp. 6790–6797. URL: `https://ieeexplore.ieee.org/abstract/document/8594495`
- TODO: UnrealROX

1.4.3 Chapter 4

- TODO: TactileGCN

1.4.4 Other

During the years spent working on the main topics of this thesis, several collaborations and side works were carried out that also were published either as journal papers, conference proceedings, or preprints:

- Sergiu Oprea, Alberto Garcia-Garcia, Jose Garcia-Rodriguez, et al. “A Recurrent Neural Network based Schaeffer Gesture Recognition System”. In: *2017 International Joint Conference on Neural Networks, IJCNN 2017, Anchorage, Alaska, May 14-19, 2017*. 2017. URL: `https://ieeexplore.ieee.org/document/7965885/`

- Francisco Gomez-Donoso, Sergio Orts-Escolano, Alberto Garcia-Garcia, et al. "A robotic platform for customized and interactive rehabilitation of persons with disabilities". In: *Pattern Recognition Letters* 99 (2017), pp. 105–113. DOI: 10.1016/j.patrec.2017.05.027. URL: https://doi.org/10.1016/j.patrec.2017.05.027

- Sergiu Oprea, Alberto GarciaGarcia, Sergio OrtsEscolano, et al. "A long short-term memory based Schaeffer gesture recognition system". In: *Expert Systems* 0.0 (2017), e12247. DOI: 10.1111/exsy.12247. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/exsy.12247

- Alberto Garcia Garcia, Andreas Beckmann, and Ivo Kabadshow. "Accelerating an FMM-Based Coulomb Solver with GPUs". In: *Software for Exascale Computing-SPPEXA 2013-2015*. Springer, 2016, pp. 485–504. URL: https://link.springer.com/chapter/10.1007/978-3-319-40528-5_22

- Alberto Garcia-Garcia, Sergio Orts-Escolano, Sergiu Oprea, et al. "Multi-sensor 3D object dataset for object recognition with full pose estimation". In: *Neural Computing and Applications* 28 (2016), pp. 941–952. ISSN: 1433-3058. DOI: 10.1007/s00521-016-2224-9. URL: http://dx.doi.org/10.1007/s00521-016-2224-9

- Marcelo Saval-Calvo, Jorge Azorin-Lopez, Andres Fuster-Guillo, et al. "Evaluation of sampling method effects in 3D non-rigid registration". In: *Neural Computing and Applications* 28 (2016), pp. 953–967. ISSN: 1433-3058. DOI: 10.1007/s00521-016-2258-z. URL: http://dx.doi.org/10.1007/s00521-016-2258-z

- Sergio Orts-Escolano, Jose Garcia-Rodriguez, Miguel Cazorla, et al. "Bioinspired point cloud representation: 3D object tracking". In: *Neural Computing and Applications* 29 (2016), pp. 663–672. ISSN: 1433-3058. DOI: 10.1007/s00521-016-2585-0. URL: https://doi.org/10.1007/s00521-016-2585-0

- Alberto Garcia-Garcia, Sergio Orts-Escolano, Jose Garcia-Rodriguez, et al. "Interactive 3D object recognition pipeline on mobile GPGPU computing platforms using low-cost RGB-D sensors". In: *Journal of Real-Time Image Processing* 14 (2016), pp. 585–604. ISSN: 1861-8219. DOI: 10.1007/s11554-016-0607-x. URL: https://doi.org/10.1007/s11554-016-0607-x

- Higinio Mora, Jerónimo M Mora-Pascual, Alberto Garcia-Garcia, et al. "Computational analysis of distance operators for the iterative closest point algorithm". In: *PloS one* 11.10 (2016), e0164694. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0164694

- Sergio Orts-Escolano, Jose Garcia-Rodriguez, Vicente Morell, et al. "3D Surface Reconstruction of Noisy Point Clouds Using Growing Neural Gas: 3D Object/Scene Reconstruction". In: *Neural Processing Letters* 43 (2015), pp. 401–423. DOI: 10.1007/s11063-015-9421-x. URL: http://dx.doi.org/10.1007/s11063-015-9421-x

- Sergio Orts-Escolano, Jose Garcia-Rodriguez, Jose Antonio Serra-Perez, et al. "3D model reconstruction using neural gas accelerated on GPU". in: *Applied Soft Computing* 32 (2014), pp. 87–100. DOI: 10.1016/j.asoc.2015.03.042. URL: http://dx.doi.org/10.1016/j.asoc.2015.03.042

- TODO: ICP

1.5 Thesis Structure

Chapter 2

Object Recognition

Abstract

In this chapter, we address the problem of object class recognition. To approach this challenge, we rely on the geometric information provided by 3D object representations such as point clouds. Furthermore, we focus on learning-based methods to distinguish objects from different classes while capturing the variability of shape of different objects which belong to the same class. More specifically, we leverage deep learning for such task. The chapter begins introducing and formulating the object recognition task in Section 2.1 followed by a review of the most relevant literature and datasets in Sections 2.2 and 2.3. After that, we present our first proposal for 3D object recognition, namely PointNet, in Section 2.4. Later, PointNet is improved and thoroughly tested in adverse conditions with noise and occlusion throughout the study in Section 2.5. Next, LonchaNet is introduced in Section 2.6 as the last iteration of our system that incorporates all the lessons learned by the previous work. Finally, Section 2.7 draws conclusions and sets future lines of research.

2.1 Introduction

Object recognition is fundamental to computer vision and despite the progress achieved during the last years, it still remains a challenging area of research. Arguably, most of the interest in object recognition is due to its usefulness for robotics.

In that regard, recognizing objects is one of the problems that must be solved to achieve total visual scene understanding. Such deeper and better knowledge of the environment eases and enables the execution of a wide variety of more complex tasks. For instance, accurately recognizing objects in a room can be extremely useful for any robotic system that navigates within indoor environments. Due to the unstructured nature of those environments, autonomous robots need to do reasoning grounded in the dynamic real world. In other words, they need to understand the information captured by their sensors to perform tasks such as grasping, navigation, mapping, or even providing humans with information about their surroundings. Identifying the classes to which objects belong is one key step to enhance the aforementioned capabilities.

Despite the easy intuitive interpretation of the problem, its inherent difficulty can be misleading. We humans recognize numerous objects in difficult settings (e.g., different points of view, occlusion, or clutter) with little to no effort. However, approaching that problem is not that easy for a computer and taking into account all the possible settings and combinations of external factors renders this task a difficult one to solve efficiently and with high precision (which is often required in numerous application scenarios).

From a formal point of view, the object recognition task can be formulated as follows: given an image $\mathcal{I}^{H \times W}$ in which an object $\mathcal{O}$ appears, which can be either a grayscale or RGB array of W pixels in width and H pixels in height, the goal is to predict the class of the object $\mathcal{L}_{\mathcal{O}}$ from a set of N predefined object classes $\mathcal{L} = \{\mathcal{L}_0, \mathcal{L}_1, ..., \mathcal{L}_{N-1}\}$.

Most of the classic literature of this topic tackled such problem by devising handcrafted feature descriptors that are extracted on certain keypoints detected over the bidimensional image and later used either to compare them against pre-existing object descriptors in a database to match them to a certain class or either to feed them as input to a shallow machine learning architecture that learns to classify those descriptors to predict the class of the object that appears in the image. That paradigm shifted recently due to the success of deep learning architectures that are able to exploit their feature learning capabilities to avoid the need of hand engineering descriptors while achieving unprecedented accuracy levels. Furthermore, the adoption and spread of depth sensors has also added a literally new dimension to learn from to boost performance. The approaches introduced in this thesis are part of that cutting-edge trend that takes advantage of the additional geometric information facilitated by commodity range scanners to perform learning over them using deep architectures. A more detailed review of the field, from the very beginning to the current trends using 3D data and deep neural networks, is performed in Section 2.2.

After that literature review, we start describing our first approach to perform object recognition using 3D data, namely PointNet, capable of learning object classes from point clouds discretized as occupancy grids with uniform voxel grids in the tridimensional space. Section 2.4 describes this architecture, its data representation, and also benchmarks it on a standard 3D object classification dataset (ModelNet) to validate it.

Following that, Section 2.5 analyzes how noise and occlusion impact such 3D deep learning architecture and the importance of the data representation when dealing with such adverse conditions that commonly appear in the real world. In that study, we also propose minor changes to the architecture and the representation themselves that significantly boost accuracy with regard to the originally proposed PointNet.

At last, Section 2.6 takes all the lessons learned from the initial PointNet proposal and the extensive study to introduce a novel slice-based architecture to tackle the 3D object class recognition problem, LonchaNet, which achieved state of the art results in the aforementioned benchmark (ModelNet10).

2.2 Related Works

2.2.1 2D Object Recognition

2.2.2 RGB-D Object Recognition

2.2.3 3D Object Recognition

2.3 Datasets

In order to evaluate the performance of our proposal in terms of accuracy we made extensive use of a well-known dataset such as the Princeton ModelNet project [REF]. Its goal, as their authors state, is to provide researchers with a comprehensive clean collection of 3D Computer Aided Design (CAD) models for objects, which were obtained via online search engines. Employees from the Amazon Mechanical Turk service were hired to classify over $150,000$ models into 662 categories.

At the moment, there are two versions of this dataset publicly available for download 2 : ModelNet-10 and ModelNet-40. Those are subsets of the original dataset, only providing the 10 and 40 most popular object categories respectively. They are specially clean since the models that did not belong to the specified categories were manually deleted.

On the one hand, ModelNet-10 is composed of a collection of over 5,000 CAD models classified into 10 categories and divided into training and test sets. In addition, the orientation of all the CAD models was manually aligned. On the other hand ModelNet-40 features over 9,800 models classified into 40 categories and it also includes training and test splits; however, their orientations are not aligned as they are in ModelNet-10.

2.4 PointNet

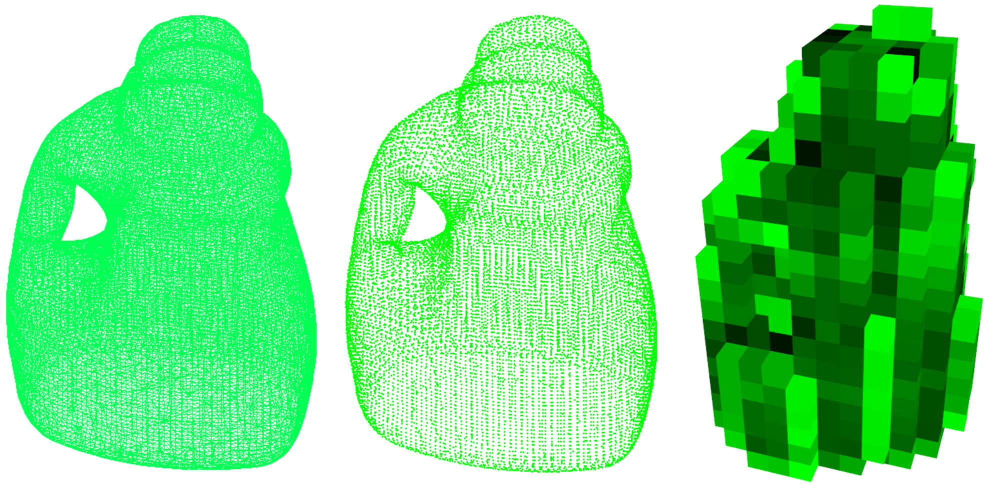
The proposed system takes a point cloud of an object as an input and predicts its class label. In this regard, the proposal is twofold: a volumetric grid based on point density to estimate spatial occupancy inside each voxel, and a pure three-dimensional (3D)-CNN which is trained to predict object classes. The occupancy grid – inspired by VoxNet [16] 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.

2.4.1 Data Representation

As we mentioned before, our proposed architecture takes a point cloud of an object as input to recognize it. However, point clouds are unstructured representations that cannot be easily handled by common CNN architectures due to the lack of a matrix-like organization. The most straightforward way to apply formal convolutions to that unstructured space is to impose a certain organization into it.

Occupancy grids 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.

As we previously reviewed in Section 2.2, certain 3D deep learning architectures make use of occupancy grids as a representation for the input data to be learned or classified. For instance, 3D ShapeNets [17] is a Convolutional Deep Belief Network (CDBN) which represents a 3D shape as a $30 \times 30 \times 30$ binary tensor in which a one indicates that a voxel intersects the mesh surface, and a zero represents empty space. VoxNet [16] introduces three different occupancy grids ($32 \times 32 \times 32$ voxels) that employ 3D ray tracing to compute the number of beams hitting or passing each voxel and then use that information to compute the value of each voxel depending on the chosen model: a binary occupancy grid using probabilistic estimates, a density grid in which each voxel holds a value corresponding to the probability that it will block a sensor beam, and a hit grid that only considers hits thus ignoring empty or unknown space. The binary and density grids proposed by Maturana *et al.* [16] differentiate unknown and empty space, whilst the hit grid and the binary tensor do not.



(a) Mesh **(b)** Cloud **(c)** Grid

Figure 2.1: Various 3D representations for an object. 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.

VoxNets occupancy grid outperforms 3D ShapeNets in terms of accuracy in the ModelNet challenge for the 3D-centric approaches described above. However, ray tracing grids considerably harmed performance in terms of execution time so that other approaches must be considered for a real-time implementation. In that very same work, the authors show that hit grids performed comparably to other approaches while keeping a low complexity to achieve a reduced runtime.

With PointNet, we propose an occupancy grid inspired by the aforementioned successes but aiming to maintain a reasonable accuracy while allowing a real-time implementation. In our volumetric representation, each point of a cloud is mapped to a voxel of a fixed-size occupancy grid. Before performing that mapping, the object cloud is scaled to fit the grid. Each voxel will hold a value representing the number of points mapped to itself. At last, the values held by each cell are normalized. Figure 2.1 shows the derivation of the proposed occupancy grid representation from other typical tridimensional representations of a sample object.

2.4.2 Network Architecture

As we have previously stated, CNNs have proven to be very useful for recognizing and classifying objects in 2D images. A convolutional layer can recognize basic patterns such as corners or planes and if we stack several of them they can learn a topology of hierarchical filters that highlight regions of the images. What is more, the composition of several of these regions can define a feature of a more complex object. In this regard, a combination of various filters is able to recognize a full object. We apply this approach used in 2D images to 3D recognition. The deep architecture featured by PointNet is represented in Figure 2.3. This setup allows PointNet to be on par with state-of-the-art algorithms while keeping reduced execution times.

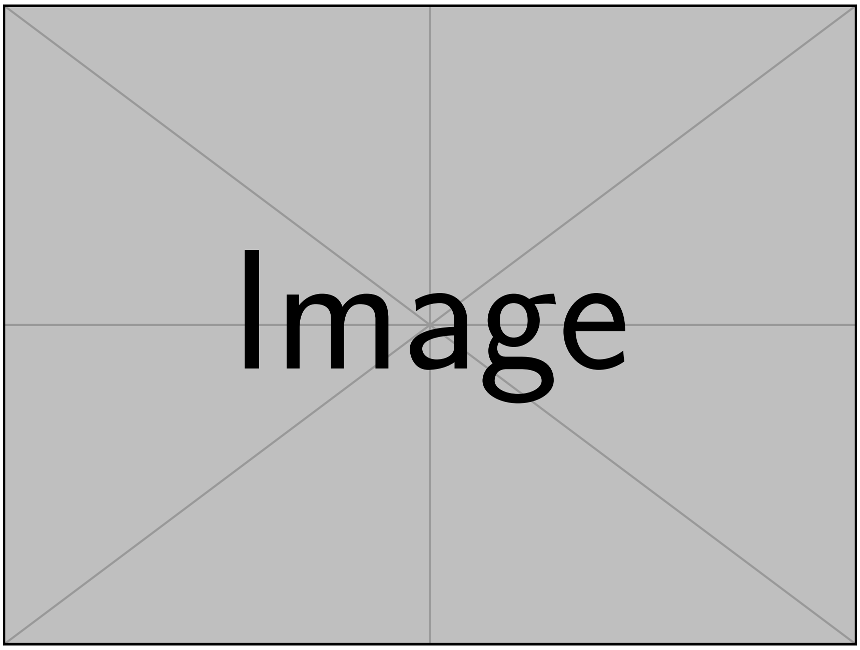


Figure 2.2: PointNet's 3D CNN architecture. [MISSINGDETAILS]

2.4.3 Experiments

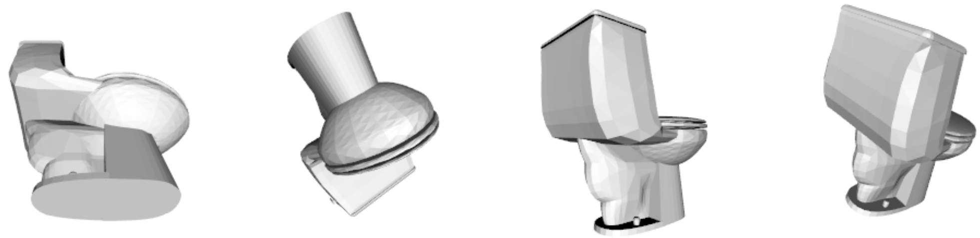
Data Generation

The CAD models are provided in OFF. Firstly, we converted all OFF models into Polygon File Format (PLY) to ease the usage of the dataset with the Point Cloud Library (PCL). As we already mentioned, the input for PointNet are point clouds, but the dataset provides CAD models specifying vertices and faces. In this regard, we converted the PLY models into Point Cloud Data (PCD) clouds by raytracing them. A 3D sphere is tessellated and a virtual camera is placed in each vertex of that truncated icosahedron pointing to the origin of the model then multiple snapshots are rendered using raytracing and the z-buffer data, which contains the depth information, is used to generate point clouds from each point of view. After all points of view have been processed, the point clouds are merged. A voxel grid filter is applied to downsample the clouds after the raytracing operations. Figure 3 illustrates the aforementioned processes. After that, the resulting point clouds are used to train, randomizing the order of the models, and test the system taking into account the corresponding splits.

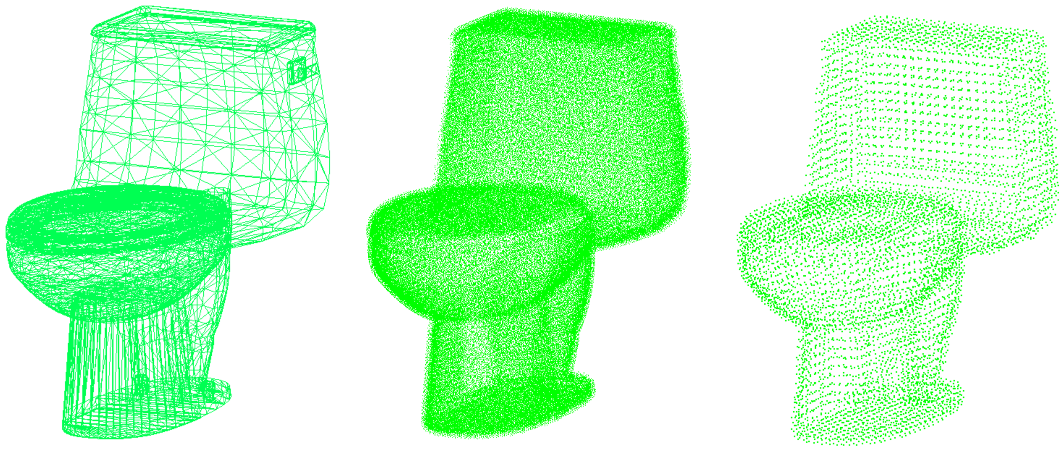
Implementation

This architecture was implemented using the PCL [18][19] which provides state-of-the-art algorithm implementations for 3D point cloud processing and Caffe [20], 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 enables researchers to design, train, and deploy CNN architectures efficiently, mainly thanks to its drop-in integration of NVIDIA cuDNN [21] to take advantage of Graphics Processing Unit (GPU) acceleration.

[1] `https://github.com/BVLC/caffe`



(a) Rendered views.



(b) Mesh. **(c)** Cloud. **(d)** Downsampled.

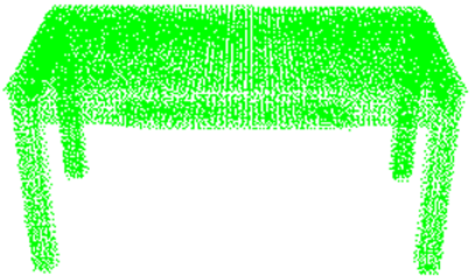
Figure 2.3: Dataset model processing example to generate the point clouds for PointNet. Some rendered views of a toilet model are shown in (a). The original OFF mesh is shown in (b). The generated point cloud after merging all points of view is shown in (c), and (d) shows the downsampled cloud using a voxel grid filter with a leaf size of $0.7 \times 0.7 \times 0.7$.

Setup

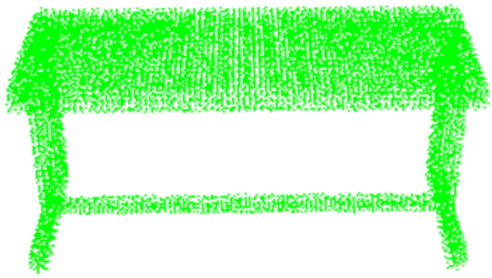
All the timings and results were obtained by performing the experiments in the following test setup: Intel Core i5-3570 with 8 GB of 1600 MHz DD3 RAM on an ASUS P8H77-M PRO motherboard (Intel H77 chipset). Additionally, the system includes an NVIDIA Tesla K20 GPU, and a Seagate Barracuda 7200.14 secondary storage. Caffe RC2 was run over ElementaryOS Freya 0.3.1, an Ubuntu-based Linux distribution. It was compiled using CMake 2.8.7, g++ 4.8.2, CUDA 7.0, and cuDNN v3.

Results and Discussion

As a result of training PointNet with a learning rate of 0.0001 and a momentum of 0.9 during 200 iterations using the ModelNet-10 dataset, it obtained a success rate of 77.6%. As shown in Figure [MISSINGREF], the confusion matrix reveals the stability of the system, mainly confusing items that look alike such as desk and table. Because of the nature of CNNs, which heavily rely on detecting combinations of features, these kind of errors are common. As we can observe in Figure 2.4, the visual features that define a desk and a table are almost the same, making it hard to distinguish between



(a)



(b)

Figure 2.4: Similarity between two objects of different classes: Table and Desk. The point cloud shown in (a) represents an object of the Table class, whilst the point cloud in (b) represents an object whose class is Desk but it is misclassified as a Table due to their resemblance.

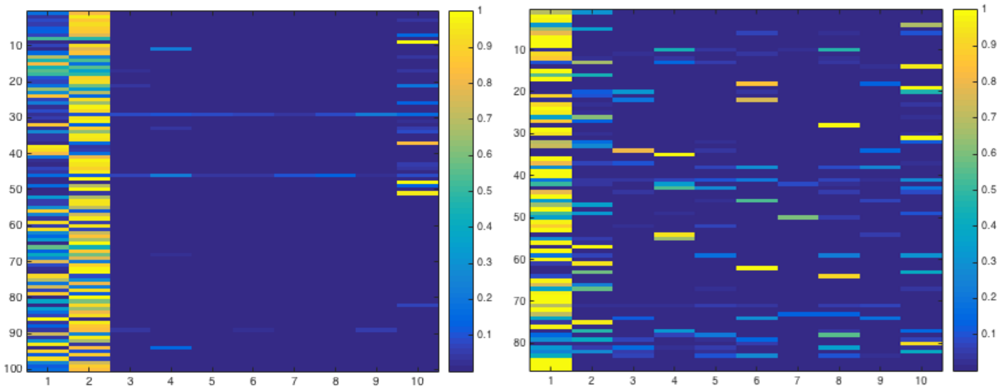
both classes. Figure 2.5 shows the neuron activations for the output layer of the architecture, proving that *Desk* and *Table* are consistently confused during the tests. In light of these experiments, and taking into account the knowledge of the CNNs principles, it is conceivable to think that a deeper network would provide better results so more experiments were carried out.

In the deeper network experiment we added several layers to the PointNet architecture. One more convolutional layer was added since these layers are coupled to the detection of the features of the objects, so the more layers there are, a better or more expressive model is produced. An Inner Product layer was also added. Since these layers make the classification possible, adding more of them would theoretically provide better classification results.

This architecture was trained during $1,000$ iterations and tested every 200 iterations. The best result was provided by the 800 iterations test with an accuracy of 76.7%, while the $1,000$ iterations test dropped the performance to a 75.9% due to overfitting.

It is well known that training using an unbalanced dataset tends to harm those classes with the least number of examples and to benefit those with the most, as stated by [MISSINGREF]. Having this in mind, and knowing that ModelNet-10 is highly unbalanced as shown in Table [MISSINGTABLE], the dataset was balanced by limiting the number of examples of each class to 400 using random undersampling. This does not fully solve the problem but improves the difference between the classes with the least number of examples and those with the most. The network was trained and tested with this more balanced dataset and it achieved an accuracy of 72.9%. The fact is that balancing the training set makes the accuracy of the classes with less examples higher, but it harms the success rate on classes with more, as seen in Figure 2.6.

After analyzing the results, it can be stated that neither a deeper network nor balancing the dataset increase accuracy. In fact, the experiments of the original architecture with the unbalanced ModelNet-10 offered the best recognition results with a 77.6% success rate. In addition, PointNet featuring the architecture exposed in Figure [MISSINGREF] takes an average time of 24.6 miliseconds to classify an example (in comparison with Voxnet, which can take up to half a second for its raytracing-based implementation). These results prove the system as a fast and accurate 3D object class recognition tool.



(a) (b)

Figure 2.5: Neuron activations for the output layer of the architecture when classifying all the test samples for both *Desk*(b) and *Table*(a) classes. Each row represents an activation vector for a specific sample, so each column is a position of the vector: the activation to that particular class. The first column corresponds to the *Desk* class, while the second one is the *Table*. The activation shows the clear confusion between *Desk* and *Table*. Although the latter one is much less confused with other classes, many *Tables* are misclassified as *Desks* thus lowering the accuracy for this class.

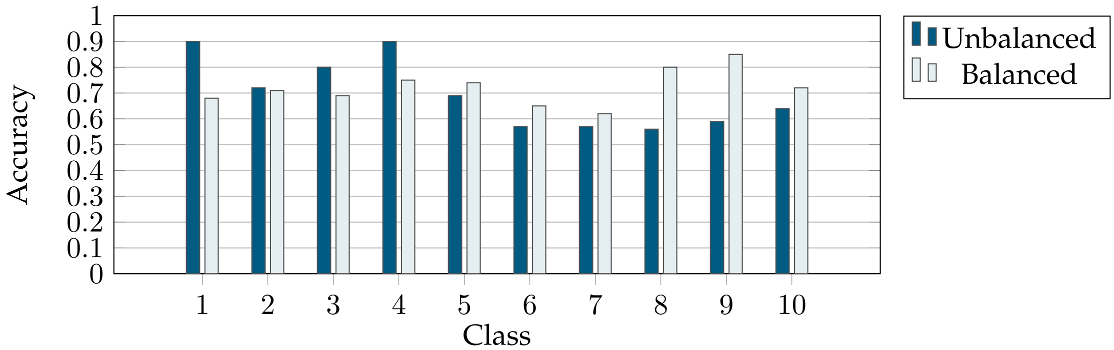


Figure 2.6: Comparison of accuracy per class using an unbalanced dataset and a balanced one with a maximum of 400 models per class via random undersampling. Accuracy is harmed in the classes in which models are removed but gained otherwise.

2.4.4 Conclusion

PointNet is a brand new kind of CNN for object class recognition that handles tridimensional data, inspired by VoxNet and 3D ShapeNets but using density occupancy grids as inner representation for input data. It was implemented in Caffe and provides a faster method than the state of art ones yet obtaining a high success rate as the experiments over the ModelNet10 dataset. This fact enlightens a promising future in real-time 3D recognition tasks.

Following on this work, we plan to improve the inner representation by using adaptable occupancy grids instead of fixed-size ones. In addition, we will integrate the system in an object recognition pipeline for 3D scenes. Our network will receive a point cloud segment of the scene where the object lies, produced by a preprocessing method, and that segments will be used to generate the occupancy grids that will be learned by the system. This implies adapting the system for learning partial views of the objects and dealing with occlusions and scale changes. As an additional feature, we will include pose estimation in that pipeline, all of this with goal of developing an end-to-end 3D object recognition system.

2.5 Noise and Occlusion

2.6 LonchaNet

2.7 Conclusion

Chapter 3

Semantic Segmentation

3.1 Introduction

3.2 Related Works

3.3 The RobotriX

3.4 UnrealROX

3.5 2D-3D-SeGCN

Chapter 4

Tactile Sensing

4.1 Introduction

4.2 Related Works

4.3 TactileGCN

4.4 Conclusion

Chapter 5

Conclusion

5.1 Findings and Conclusions

5.2 Limitations

5.3 Future Work

Bibliography

[1] Alberto Garcia-Garcia, Francisco Gomez-Donoso, Jose Garcia-Rodriguez, et al. "PointNet: A 3D Convolutional Neural Network for real-time object class recognition". In: *2016 International Joint Conference on Neural Networks, IJCNN 2016, Vancouver, BC, Canada, July 24-29, 2016*. 2016, pp. 1578–1584. DOI: 10.1109/IJCNN.2016.7727386. URL: https://doi.org/10.1109/IJCNN.2016.7727386.

[2] Alberto Garcia-Garcia, Jose Garcia-Rodriguez, Sergio Orts-Escolano, et al. "A study of the effect of noise and occlusion on the accuracy of convolutional neural networks applied to 3D object recognition". In: *Computer Vision and Image Understanding* 164 (2017), pp. 124–134. DOI: 10.1016/j.cviu.2017.06.006. URL: https://doi.org/10.1016/j.cviu.2017.06.006.

[3] Francisco Gomez-Donoso, Alberto Garcia-Garcia, Jose Garcia-Rodriguez, et al. "LonchaNet: A Sliced-based CNN Architecture for Real-time 3D Object Recognition". In: *2017 International Joint Conference on Neural Networks, IJCNN 2017, Anchorage, Alaska, May 14-19, 2017*. 2017. URL: https://ieeexplore.ieee.org/document/7965883/.

[4] Alberto Garcia-Garcia, Pablo Martinez-Gonzalez, Sergiu Oprea, et al. "The RobotriX: An eXtremely Photorealistic and Very-Large-Scale Indoor Dataset of Sequences with Robot Trajectories and Interactions". In: *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2018, pp. 6790–6797. URL: https://ieeexplore.ieee.org/abstract/document/8594495.

[5] Sergiu Oprea, Alberto Garcia-Garcia, Jose Garcia-Rodriguez, et al. "A Recurrent Neural Network based Schaeffer Gesture Recognition System". In: *2017 International Joint Conference on Neural Networks, IJCNN 2017, Anchorage, Alaska, May 14-19, 2017*. 2017. URL: https://ieeexplore.ieee.org/document/7965885/.

[6] Francisco Gomez-Donoso, Sergio Orts-Escolano, Alberto Garcia-Garcia, et al. "A robotic platform for customized and interactive rehabilitation of persons with disabilities". In: *Pattern Recognition Letters* 99 (2017), pp. 105–113. DOI: 10.1016/j.patrec.2017.05.027. URL: https://doi.org/10.1016/j.patrec.2017.05.027.

[7] Sergiu Oprea, Alberto GarciaGarcia, Sergio OrtsEscolano, et al. "A long short-term memory based Schaeffer gesture recognition system". In: *Expert Systems* 0.0 (2017), e12247. DOI: 10.1111/exsy.12247. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/exsy.12247.

[8] Alberto Garcia Garcia, Andreas Beckmann, and Ivo Kabadshow. "Accelerating an FMM-Based Coulomb Solver with GPUs". In: *Software for Exascale Computing-SPPEXA 2013-2015*. Springer, 2016, pp. 485–504. URL: https://link.springer.com/chapter/10.1007/978-3-319-40528-5_22.

[9] Alberto Garcia-Garcia, Sergio Orts-Escolano, Sergiu Oprea, et al. "Multi-sensor 3D object dataset for object recognition with full pose estimation". In: *Neural Computing and Applications* 28 (2016), pp. 941–952. ISSN: 1433-3058. DOI: 10.1007/s00521-016-2224-9. URL: http://dx.doi.org/10.1007/s00521-016-2224-9.

[10] Marcelo Saval-Calvo, Jorge Azorin-Lopez, Andres Fuster-Guillo, et al. "Evaluation of sampling method effects in 3D non-rigid registration". In: *Neural Computing and Applications* 28 (2016), pp. 953–967. ISSN: 1433-3058. DOI: 10.1007/s00521-016-2258-z. URL: http://dx.doi.org/10.1007/s00521-016-2258-z.

[11] Sergio Orts-Escolano, Jose Garcia-Rodriguez, Miguel Cazorla, et al. "Bioinspired point cloud representation: 3D object tracking". In: *Neural Computing and Applications* 29 (2016), pp. 663–672. ISSN: 1433-3058. DOI: 10.1007/s00521-016-2585-0. URL: https://doi.org/10.1007/s00521-016-2585-0.

[12] Alberto Garcia-Garcia, Sergio Orts-Escolano, Jose Garcia-Rodriguez, et al. "Interactive 3D object recognition pipeline on mobile GPGPU computing platforms using low-cost RGB-D sensors". In: *Journal of Real-Time Image Processing* 14 (2016), pp. 585–604. ISSN: 1861-8219. DOI: 10.1007/s11554-016-0607-x. URL: https://doi.org/10.1007/s11554-016-0607-x.

[13] Higinio Mora, Jerónimo M Mora-Pascual, Alberto Garcia-Garcia, et al. "Computational analysis of distance operators for the iterative closest point algorithm". In: *PloS one* 11.10 (2016), e0164694. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0164694.

[14] Sergio Orts-Escolano, Jose Garcia-Rodriguez, Vicente Morell, et al. "3D Surface Reconstruction of Noisy Point Clouds Using Growing Neural Gas: 3D Object/Scene Reconstruction". In: *Neural Processing Letters* 43 (2015), pp. 401–423. DOI: 10.1007/s11063-015-9421-x. URL: http://dx.doi.org/10.1007/s11063-015-9421-x.

[15] Sergio Orts-Escolano, Jose Garcia-Rodriguez, Jose Antonio Serra-Perez, et al. "3D model reconstruction using neural gas accelerated on GPU". In: *Applied Soft Computing* 32 (2014), pp. 87–100. DOI: 10.1016/j.asoc.2015.03.042. URL: http://dx.doi.org/10.1016/j.asoc.2015.03.042.

[16] Daniel Maturana and Sebastian Scherer. "Voxnet: A 3d convolutional neural network for real-time object recognition". In: *Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ International Conference on*. IEEE. 2015, pp. 922–928.

[17] Zhirong Wu, Shuran Song, Aditya Khosla, et al. "3D ShapeNets: A Deep Representation for Volumetric Shapes". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015. URL: http://arxiv.org/abs/1406.5670.

[18] Radu B. Rusu. "Point Cloud Library". In: *IEEE International Conference on Robotics and Automation (ICRA)*. 2011, pp. 1–4.

[19] Aitor Aldoma, Zoltan-Csaba Marton, Federico Tombari, et al. "Tutorial: Point cloud library: Three-dimensional object recognition and 6 dof pose estimation". In: *IEEE Robotics & Automation Magazine* 19.3 (2012), pp. 80–91.

[20] Yangqing Jia, Evan Shelhamer, Jeff Donahue, et al. "Caffe: Convolutional Architecture for Fast Feature Embedding". In: *Proceedings of the ACM International Conference on Multimedia*. 2014, pp. 657–678.

[21] Sharan Chetlur, Cliff Woolley, Philippe Vandermersch, et al. "cuDNN: Efficient Primitives for Deep Learning". In: (2014), pp. 1–9. ISSN: 08876266. DOI: `10.1002/polb.23894`. eprint: `1410.0759`.