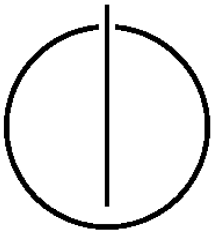
TECHNISCHE UNIVERSITÄT MÜNCHEN

DEPARTMENT OF INFORMATICS

Master's Thesis in Informatics

Single Shot 6D Object Detection in 3D Point Clouds

Mustafa Onur Eken



TECHNISCHE UNIVERSITÄT MÜNCHEN

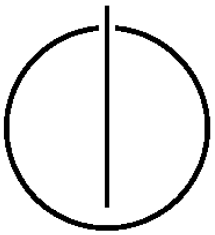
DEPARTMENT OF INFORMATICS

Master's Thesis in Informatics

Single Shot 6D Object Detection in 3D Point Clouds

6D Objekterkennung in 3D Punktwolken in einem Versuch

Author:	Mustafa Onur Eken
Supervisor:	Slobodan Ilic
Advisor:	Tolga Birdal
Submission Date:	14.02.2019



I confirm that this master's thesis in informatics is my own work and I have documented all sources and material used.

Munich, 14.02.2019 Mustafa Onur Eken

Abstract

Object detection is a very hot research topic and the studies have given rise to high impact work such as Faster-RCNN, YOLO, SSD and countless more. A similar task, pose estimation in 3D data is now a major concern thanks to the proliferation of 3D sensors, particularly in domains of autonomous driving and large scale reconstruction.

In this project we develop, on existing frameworks, a deep learning based 3D object detection algorithm, directly consuming 3D data and along with its appearance and differential properties, such as color or surface normals, producing estimations in a single-shot manner while operating natively in 3D at all stages. In contrast to the recently developed approaches, our method does not require auxiliary representations, pose refinement and is able to handle objects with arbitrary poses.

Contents

1 Introduction

Advancements in the learning based AI algorithms in computer vision and demonstration of their impressive performance in the recent years resulted in increased demand in a world where cars can drive autonomously, robots are able to clean houses with no supervision and augmented/virtual reality experiences are ubiquitous in our daily lives.

One of the key challenges towards these goals is object detection. For example, in modern autonomous systems and AR/VR applications, commonly a perception module is required for the discovery of the relevant structures in the environment where the agent is situated, see Figure 1.1. In this particular domain, the location and the orientation of the relevant objects is a crucial piece of information in order to successfully carry out remaining sub-tasks such as localization and route planning. Consequently, such a central problem receives more and more attention from scholars nowadays.

In this thesis, we address a particular instance of this problem, however before we elaborate on details, we find it beneficial to formalize and classify the different object detection problems in order to position and facilitate our approach.

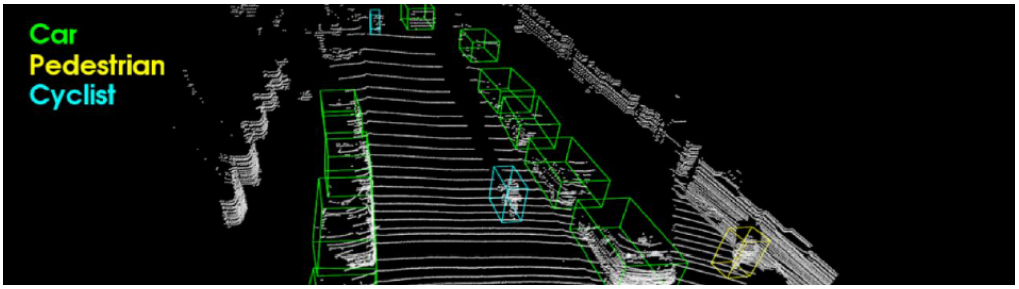


Figure 1.1: Perception in autonomous driving. Source: [31]

1.1 Object Detection Problems

The object detection is a broadly defined problem and throughout recent decades, different variations of it have been studied. We classify those into 3 main categories.

1. **2D detection in 2D modalities**
   Estimation of the 2D bounding box around the object given an image.
2. **3D detection in 2D modalities**
   Estimation of the 3D bounding box of the object in the 3D coordinate system given an image.
3. **3D detection in 3D modalities**
   Estimation of the 3D bounding box of the object in the 3D coordinate system given a representation of the 3D scene.

Any of these categories could be further split into 2 different sub-categories. The proposed approach could perform:

A. **Single-class prediction**
   1. Without intra-class variations (*instance detection*) (*6D pose estimation*)
   2. With intra-class variations

B. **Multi-class prediction**
   1. Without intra-class variations (*multi-instance detection*)
   2. With intra-class variations

Using this kind of grouping and our naming convention, we refer, for example, to the problem of *instance detection* by **Problem (*A1)**, to the problem of *2D instance detection in 3D data* by **Problem (2A1)**, and so on. See Figure 1.3 for an example on the inputs and output of the problems **(1B2)** and **(2B1)**.

Clearly, the single-class case is a simplification compared to the multi-class case and in like manner, assuming no intra-class variations is a simplification compared to assuming intra-class variations. With this reasoning, although the *instance detection* **(*A1)** is considered the most constrained and the easiest setup among the categories we mentioned, it is certainly a relevant problem to study since it emerges frequently in various real-life scenarios. On the other hand, the object detectors proposed nowadays that address the problem **(*B2)** are still not scalable to a large number of object classes.

Some of the earlier works in the field, such as [14], attack the problem **(2A1)** using grayscale/RGB images as the input modality. They often require the extraction of relevant key points with useful descriptors and the acquisition of the 2D-3D correspondences. On top of that, these approaches either involve computation of a closed form solution originating from geometrical constraints or running an iterative optimization algorithm to obtain the object pose. See Figure 1.2. Performance of such methods often is very sensitive to factors such as:

- Multi-stage design and propagation of errors

- Textureless objects
- Quality of the keypoint detector
- Size of the inlier correspondences set
- Ratio of outlier correspondences to inliers
- Sensitivity of the optimization to the initialization

Figure 1.2: Detection using a complex pipeline with handcrafted components **(2A1)**.

Unlike the traditional approaches, many of the recent works [24, 6, 22, 13, 19] are targeted towards the problem **(1B2)**, also named as *object classification and localization*. See Figure 1.3a. Often, they propose a powerful deep neural network model that is able to cope with the intra-class variations. Although the training phase of the algorithms is fairly tricky, the test time performances are in terms of accuracy and inference time are remarkable.

Given the demonstrated performance of neural networks, the ubiquity of RGB images and the need for 3D detections gave birth to approaches such as [18, 9, 28] where the authors propose algorithms attacking the problem **(2B1)**. See Figure 1.3b. It has been shown that, compared to the traditional methods, learning based methods have significantly higher performances, simpler procedures and are less sensitive to imaging variations.

Many object detectors rely on the exploitation of RGB images, but the advancements in the 3D sensing technologies and RGB-D cameras, rendered them inexpensive and accessible in many real-life scenarios, including the autonomous driving and AR/VR applications, which in turn, encouraged the researchers to make use of these new

(a) Classification and localization **(1B2)**. Source: [21]

(b) Instance detection and pose estimation **(2B1)**. Source: [28]

Figure 1.3: Detection outputs in problems **(1B2)** and **(2B1)**.

data modalities in developing their approaches. Although RGB images provide rich, compact and easy-to-process raw information, they are lacking structural information about the surroundings that RGB-D cameras and 3D scanners could offer.

Thereupon, several methods [2, 10, 8, 15] attempt to tackle the problem **(2B1)** by making use of the availability of RGB-D images also called 2.5D data in the literature. Although they have proven to be significantly robust in terms of accuracy, power-hungry RBG-D sensors limit the feasibility of such real-life applications that inherently require mobility of the devices.

Similar to the case of RGB-D sensors, LiDAR sensors became more and more applicable to autonomous driving scenarios and many object detection algorithms has been proposed that exploits point cloud data [4, 31, 16, 26]. Predominantly, these methods apply deep learning algorithms to some representation of 3D data. While some of them convert the point cloud to voxel data and simply adapt the common 2D neural network layers to 3D, some others propose new layers that are able to operate on the point cloud as it is. The issue with the former is the dramatically increased memory and computation power requirements. To remedy this issue, some methods project their data into 2D through the first several layers of the network. On the other hand, the fundamental issue with the latter method is the fact that the time complexity of the algorithms scales linearly with the size of the point clouds. Knowing that, in real-life scenarios point clouds grow tremendously, typically a pre-processing step is executed to extract a reasonably large and relevant point cloud from the original one.

1.2 Motivation

We draw the observation that in 3D object detection unlike in the case of 2D detection:

- **No method aims for full 3D rotation estimation**

  Works such as [4, 16, 31, 26] evaluate their networks on KITTI dataset where the rotation of the objects are to be predicted along z-axis only.

- **No method operates natively in 3D representation and produces a single-shot prediction**

  Works such as [4, 31, 26] project the 3D data into a 2D representation at a certain stage before the prediction. Works such as [9, 18, 28, 2, 10, 15] operate on 2D data entirely.

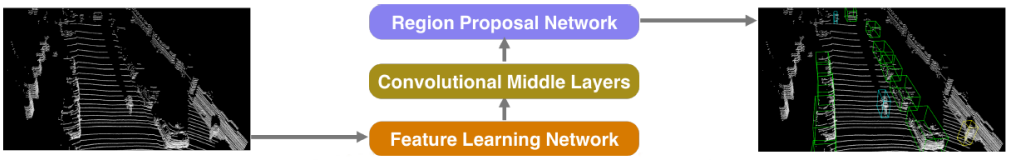
- **Many methods need cumbersome processing**

  Approaches such as [16, 9] employ a multi-stage prediction pipeline either involving a pre or post-processing such as non-linear pose refinement.

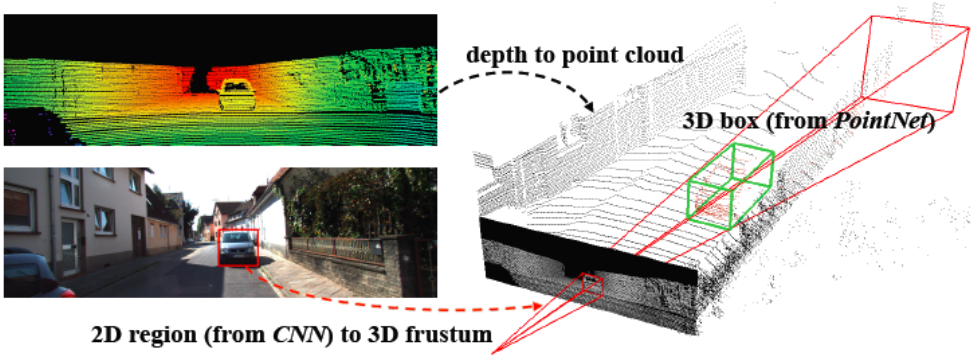
- **Many methods require auxiliary representations**

  Approaches such as [16] require RGB images in order to be able to operate.

Specifically, in this thesis, we address the problem of *multi-class 3D instance detection in 3D data* while addressing the drawbacks in the present techniques, see Figure 1.4. Our algorithm produces *6 DoF* estimations for each object to be detected in the 3D scenes. For this purpose, we develop a deep neural network algorithm that operates on 3D representations of data at all stages throughout the inference thus retaining the spatial information at every step. On top of that, our technique is able to address the datasets where the object rotations are unconstrained. We train our model end-to-end and produce the predictions in a single-shot manner. We base our approach on the *Octnet framework* [23] for the handling of the memory requirements and follow design choices similar to those of *YOLO object detectors* [19, 20, 21]. We evaluate the performance of our approach on *synthetic* and *KITTI datasets* and demonstrate it's characteristics in the ablation studies we perform.



(a) VoxelNet: Projects voxel data into 2D through the layers of the network. Source: [31]



(b) Frustum PointNet: Runs the network only on the relevant point cloud patch. Source: [16]

Figure 1.4: Prominent object detectors in the context of autonomous driving and their drawbacks.

1.3 Thesis Outline

In chapter 2, we review the recent work on 3D instance detection in 3D data and discuss the advantages/disadvantages of the selected methods. In particular we elaborate on Vote3Deep [4], VoxelNet [31], AVOD [11], Frustum PointNet [16], SECOND, PointRCNN. In the following chapter, we present our model and pose estimation algorithm, explain implementation details and the reasoning behind our design choices and their implications. In chapter 4, we present the preparation of the synthetic and KITTI datasets, the various experiments and ablation studies we conduct and the results we obtain. In addition, we discuss the findings from our experiments. In the last chapter, we summarize our work and draw conclusions together with improvement suggestions.