
Car Driving Without Cameras

By

JONES MABEA AGWATA



Department of Computer Science
UNIVERSITY OF BRISTOL

A dissertation submitted to the University of Bristol in accordance with the requirements of the degree of MASTER OF SCIENCE in the Faculty of Engineering.

SEPTEMBER 2018

Word count: ten thousand and four

EXECUTIVE SUMMARY

The need for robust object detection in 3D point clouds has greatly increased with the ongoing push for autonomous vehicles (AVs). Most of these systems use Light Detection and Ranging (LiDAR), cameras or a combination of both in order to perform object detection. LiDAR presents objects as point clouds in a 3D space thus offering critical shape information of objects in view. However, this representation is sparse. As a result, LiDAR-based detection performs poorly as compared to multimodal methods that have helped overcome this. Nonetheless, multimodal methods are often complex to set up and synchronise with the cost of components running into thousands of pounds. In light of this, reducing the number of sensors while still maintaining or even improving the accuracy has become a topic of interest by industry players who are developing AVs. With the cost of LiDAR declining following recent improvements in solid-state LiDAR technology, dropping cameras in favour of LiDAR has become more plausible. As compared to cameras that suffer drawbacks such as visibility issues in extreme weather, LiDAR is extremely robust and accurate in various conditions.

The aim of this project is to replicate and improve on VoxelNet, A Region Proposal Network for point cloud object detection that was recently released by Apple Inc. Doing so will be of added value as the implementation is currently proprietary to Apple Inc. If successful, my implementation will be made open source and can be used for further research in the field of object detection from LiDAR point clouds. In addition, through the evaluation and analysis, I hope to propose ways in which my implementation can be used in making cheaper AVs through the use of solid-state LiDAR without having to use cameras at all.

In consideration of this, two main objectives emerge. The first is to develop a region proposal network for object detection from LiDAR point clouds. This will be implemented using the TensorFlow machine learning framework. The second will be a detailed evaluation report containing research into this topic area as well as an analysis and comparison of the performance of my implementation. Finally, an evaluation of the VoxelNet implementation will be performed to establish its limitations and possible ways to improve it will be discussed.

This project will be 65% type I (Software Development) and 35% type II (Theoretical). Software development of the deep learning framework will involve obtaining AV videos and point cloud data from public datasets such as KITTI. This data will then be used in the development and training of the RPN. On completion of development, the model will be evaluated against the KITTI benchmark. Tentatively, datasets from the University of Bristol Robotics department from their AV project will be tested on the implementation and a sanitised LiDAR dataset released.

DEDICATION AND ACKNOWLEDGEMENTS

Here goes the dedication.

AUTHOR'S DECLARATION

I declare that the work in this dissertation was carried out in accordance with the requirements of the University's Regulations and Code of Practice for Research Degree Programmes and that it has not been submitted for any other academic award. Except where indicated by specific reference in the text, the work is the candidate's own work. Work done in collaboration with, or with the assistance of, others, is indicated as such. Any views expressed in the dissertation are those of the author.

SIGNED: DATE:

TABLE OF CONTENTS

	Page
List of Tables	ix
List of Figures	xi
1 Introduction	1
1.1 Aims and Objectives	2
1.2 Deliverables	2
1.3 Report structure	3
2 Background	5
2.1 Components of an AV	5
2.2 Industrial Approaches	8
2.2.1 Camera	9
2.2.2 Camera and Radar	10
2.2.3 LiDAR, Camera and Radar	10
2.3 Related Research	10
2.3.1 Classic Computer Vision	10
2.4 Legal, Ethical and Economic Considerations	10
2.4.1 Deep Neural Networks	11
3 Methodology	15
3.1 Datasets	15
3.2 Context Detection and Characterisation	15
3.3 Context-Dependent Model Selection	16
3.3.1 Conclusion	16
3.4 Validation with different datasets	16
4 Results and Analysis	17
5 Conclusion	19

TABLE OF CONTENTS

A Appendix A	21
Bibliography	23

LIST OF TABLES

TABLE	Page
A1 Velodyne LiDAR Family	22

LIST OF FIGURES

FIGURE	Page
21 Velodyne LiDAR family. From left: Velodyne HDL-64E , HDL-32E , PUCK	6
22 Components of Waymo’s Self Driving Car [35]	7
23 AV Leaderboard. Navigant Research[31]	9

INTRODUCTION

Accelerated by recent advancements in technology, the prospect of Autonomous Vehicles (AVs) driving in public roads is becoming more and more a reality. As this is an emerging field, there are numerous variations of implementations by different companies. Arguably, a key characteristic of these implementations is a large number of perception sensors including cameras, radars and light and detection ranging sensors(LiDAR). This is necessary for perceiving the environment around the vehicles in order to safely maneuver around surrounding objects. However, most of these sensors are quite expensive and also energy inefficient thus making AVs unviable. In an effort to reduce the cost, companies are exploring different ways to reduce the number of sensors while still achieving a high level of navigational accuracy and safety.

In the multimodal approach, sensors are combined in various ways to achieve cost effective yet accurate configurations. Common configurations include Camera only, camera and radar, camera, LiDAR and radar. This is influenced by the different strengths and weaknesses of the sensors in terms of performance. Despite extensive research on these perception sensors, their performance in urban and non-urban contexts has not been exhaustively explored. This presents an opportunity to redesign the sensor configurations from a top down approach by first understanding their performance in different contexts then designing models that are efficient and feasible.

In order to investigate the performance of these sensors, state of the art object detection models will be considered. Specifically, two models will be used as the templates in this evaluation. Firstly, VoxelNet [38], is a LiDAR only model that uses point clouds as input. Secondly, Aggregated View Object Detection(AVOD) [18] is multimodal model that fuses image and point cloud data. Both model implementations were available on GitHub and were heavily modified in order to align with the aims of this project. Nonetheless, they formed an essential foundation that cannot be understated.

1.1 Aims and Objectives

Following the motivations in the presented discussion , the performance of LiDAR only and multimodal(LiDAR and Camera) models in different contexts will be investigated with the aim of reducing the number of sensors in AVs. To achieve this aim, the following objectives will need to be achieved.:

1. Detect and characterise the context of images and point clouds.
2. Evaluate the performance of sensors in different contexts.
3. Evaluate the performance of both models in different contexts.
4. Validate performance of single sensor model on custom dataset.
5. Legal, social and economic analysis of current implementations and proposed improvements.

1.2 Deliverables

The deliverables are split into two groups.

- **Image and LiDAR Context Classifier.** Available as Jupyter interactive notebooks including pre-trained models.
- **Custom VoxelNet Model** Modified model of the VoxelNet unofficial implementation including interactive notebooks for training, testing and validating the model.
- **Custom AVOD Model** Modified model of the AVOD official implementation including interactive notebooks for training, testing and validating the model.
- **Validation Dataset** Point Cloud dataset obtained from the .
- **Evaluation report.** The following topics will be discussed.
 1. A review of related research and implementations tackling object detection using LiDAR cloud points.
 2. Performance analysis of implementation and analysis criteria.
 3. A comparison between the implemented system and other state-of-the-art detection systems, potentially through a public benchmark.
 4. The ethical and safety implications of the system and its viability in a real world setting.

5. Economic analysis of the implementation and its potential impact on the development of AVs.
6. Validation of implementation performance against university or public datasets containing data from AVs.

1.3 Report structure

This report will consist of five main chapters. Chapter 2 discusses the different components of AVs, current implementations in the industry, a background on the research that has been undertaken in the field of object detection and finally a brief overview of deep learning frameworks.

Chapter 3 is the final discussion that forms the conclusion of the review. In this chapter, I will evaluate what I hope to achieve and the assumptions and hypotheses that I have formulated.

BACKGROUND

The idea of autonomous vehicles(AVs) is not new. As early as 2005, DARPA had invested heavily in the creation of unmanned trucks and organised for the Urban Challenge [6] to allow for different teams to showcase their unmanned vehicles. However, due to the challenges such as low computational power and underdeveloped AI and ML systems, the resulting implementations were not practical and had a high fault rate of 1 fault in 100 miles compared to the human fault rate of around 1 in 100 million miles. Nonetheless, from this challenge, it was clear that the prospect of AVs was plausible and indeed possible.

Currently, AVs are divided into five levels of autonomy as defined by the Society of Automotive Engineers(SAE):

SAE level	Description
0	No AV control systems. An example is blind spot indicators.
1	Basic driver assistance built into vehicle design. An example is the cruise control function.
2	Basic AV control systems. Driver required to monitor the environment and take back control if need be. A good example is the Tesla Autopilot.
3	AVs can safely navigate and drive within mapped environment. Driver required to monitor the environment and take back control if need be.
4	Highly autonomous control capable of handling most conditions but the driver has the option to take control. Renault ISymbioz is an example of a level 4 AV.
5	Completely autonomous with zero human involvement.

2.1 Components of an AV

- **LiDAR** - LiDAR provides highly detailed 3D information about the environment around the vehicle and objects in it. LiDAR operates by sending out pulses of lasers and recording

the reflections of the pulses from objects. By comparing this with the time taken for the lasers to be reflected (time of flight) and their direction, the distance of these objects can be calculated and mapped in a point cloud.

LiDAR units require complex optical systems that are expensive to build. As a result, they are the most expensive sensors in AVs with the top end such as Velodyne HDL-64E shown in figure 21 costing more than 50,000\$. In a bid to reduce the cost of LiDAR units, different companies are exploring different design methods that are cheaper but still able to offer the same performance as the top end LiDAR units such as solid state LiDAR¹.



Figure 21: Velodyne LiDAR family. From left: Velodyne HDL-64E , HDL-32E , PUCK

- **Cameras** - Cameras mounted on the vehicle are used for classification and identification of various objects on the road. Cameras can also be used to create 3D maps of the surrounding environment. By combining two cameras, a stereo image can be captured that provides depth information. Alternatively, by combining a camera and IR Laser sensor for depth estimation, RGB-D [15] images are obtained and mapped in a point cloud.
- **Position Estimators** - Position estimators are a group of sensors used for navigation of the vehicle. These include GPS systems, odometers and gyroscopes.

¹See appendix ??

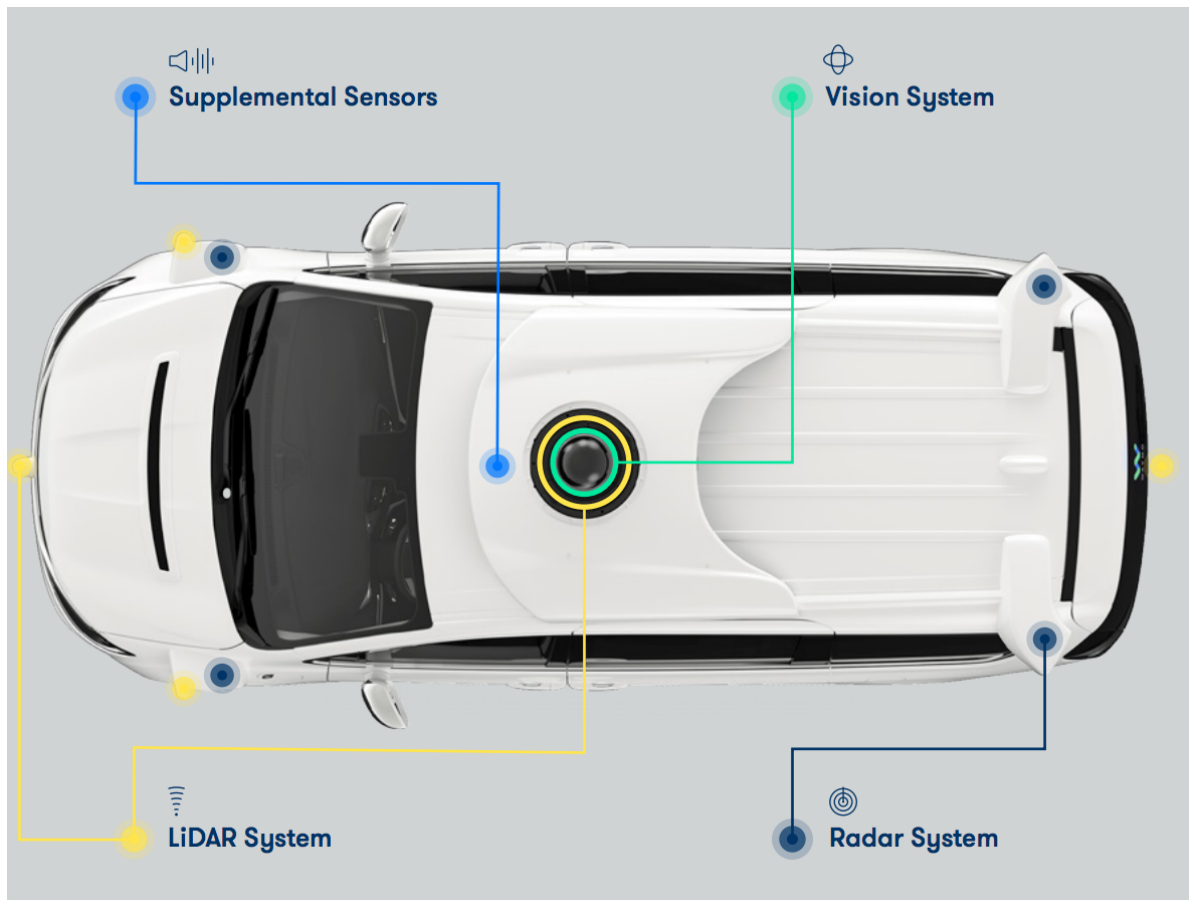


Figure 22: Components of Waymo's Self Driving Car [35]

- **Distance Sensors** - Distance sensors such as radars and sonars are important for gauging the distance of objects on the road. Radars are the most commonly used distance sensors and they work by transmitting radio waves and recording the reflected radio waves from objects. As compared to cameras and LiDARs, radars work well in a variety of low visibility scenarios such as poor weather. However, the reflectivity of these radio waves depends on the nature of objects, their size, absorption characteristics and the transmitting power. As such, it may not be effective for detecting objects with low absorption characteristics such as pedestrians and animals.
- **Processing Unit** - In order to process all the data from sensors on the vehicle, AVs require powerful processing units to do so in real time. Most ML/AI algorithms used for detecting and identifying objects from LiDAR and camera data demand large amounts of processing power. This is achieved through the use of CPUs, GPUs, Field Programmable Gate Arrays(FPGA)[5], Application Specific Integrated Circuits(ASICs)[33] or combinations with each other.

AVs have three main modes of operation namely,

- **Perception** - This is the first step which involves processing the input from the sensors. In this mode tasks such as object detection and tracking, lane detection, traffic sign detection and recognition are performed.
- **Planning** - This is the next step after detection and recognition tasks are performed . In this stage route and trajectory planning algorithms are run to plan how the vehicle should navigate in the immediate environment as well as a route to a target location. These algorithms are required to handle complex situations to ensure safety of the passengers and other road users.
- **Control** - This stage involves the execution of plans created in the planning stage. This stage is crucial as the actuators involved in steering and movement have to be able to be able to accurately follow the plans. This involves calculation of energy and forces. At this stage the trajectories and movement of other road users and objects have to be calculated in order to anticipate and avoid any accidents.

2.2 Industrial Approaches

In order to be viable, AVs need to meet a few constraints as discussed in [22].

- **Performance**

At the moment there are no clear regulations as to how fast the perception, planning and control pipeline should be. However according to research by Brown et al. [4], humans take around 600 ms to respond and brake when expecting an interruption, however this figure shoots to 850 ms when an unexpected situation arises. In addition, Newell et al [26], established that the fastest human response time is between 100-150ms. These figures can be used as a baseline while developing a pipeline. In this pipeline, there are two factors to consider, namely the frame rate (frequency of the data from sensors) and the processing latency (time taken to process data). As such, an AV system should be able to react within 100ms, that is, faster than the human response time.

- **Storage**

AVs should be able to store maps of different areas with fine granularity for accuracy during localisation. As a result, the storage of these maps can run into tens of Terabytes. Despite recent advances in cloud technology and the emergence of 5G connectivity that is significantly fast. Downloading these maps would take significant time and would also render the car unusable in case of no internet connectivity. As such, the vehicles should have enough storage to store these maps locally.

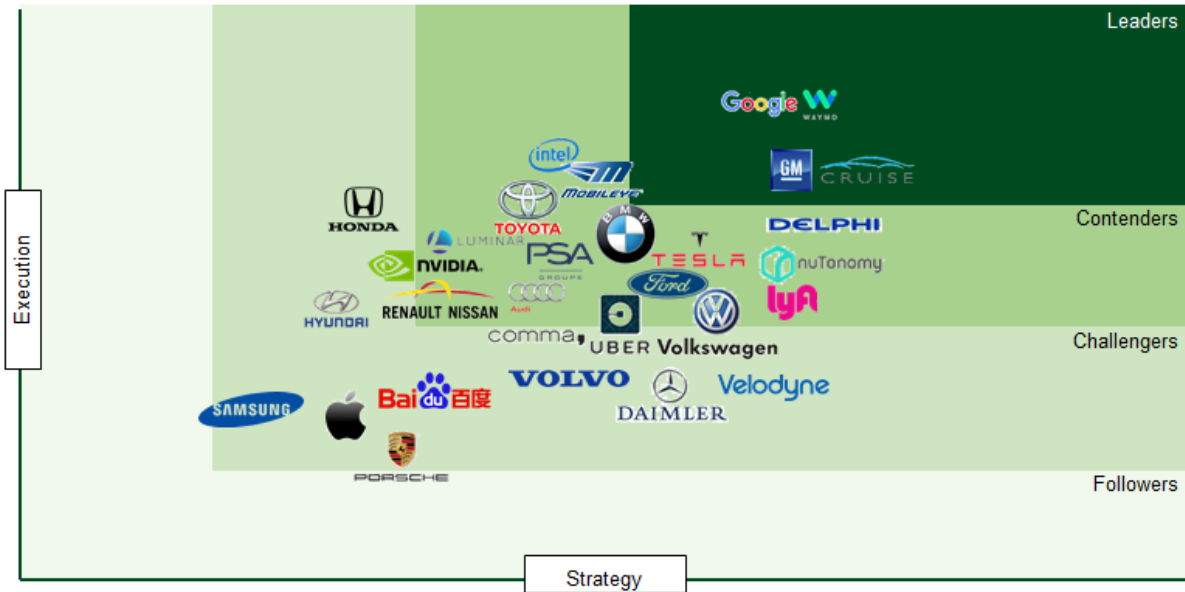


Figure 23: AV Leaderboard. Navigant Research[31]

- **Power**

Most of the major industry players have moved to electric-vehicles for their AV systems. Depending on the equipment and sensor configurations used in these systems, the power usage can range from 500 watts to 1.5 kilowatts. Given that the AVs have limited battery capacity, heavy power consumption by these systems can lead to poor driving ranges hence making the cars less viable. As such the configuration of these systems has to be carefully considered to ensure a reasonable driving range.

- **Thermal**

Processing components in the AV systems such as CPUs and GPUs require a significant amount of energy to cool. This is necessary to ensure that they operate within their recommended thermal operating range. Failure to do so could result in the failure of these systems. As such, additional cooling systems have to be installed in the vehicle in order to ensure this.

Following the discussion above, the next subsections will review how different industry players are implementing their AV systems with special focus to the issue of perception

2.2.1 Camera

Mobileye[25], is one of the top contenders in the development of camera only AVs. It was recently acquired by Intel and believes that AVs should be able to accurately and safely navigate using cameras only given the fact that humans are able to do so with vision only. Previously, MobilEye was a supplier of vision systems for Advanced Driver Assistance Systems and had a partnership

with Tesla to supply their vision systems prior being acquired by Intel. Given their extensive background in developing these vision systems, the partnership with Intel aims to develop a complete autonomous driving package.[16]

2.2.2 Camera and Radar

Another major contender that is using the camera and radar configuration is Tesla. Elon Musk, the founder, believes that LiDAR is not necessary for AV perception. However, just recently, a Tesla vehicle was involved in a crash while in 'autopilot' mode that resulted in the loss of life. A major attributing factor to this accident was caused by the fact that bright sunlight blinded the camera thus causing the vehicle not to detect a white truck that was ahead of it. The radar was also unable to detect it thus leading to a collision. LiDAR on the other hand, would have been able to detect the truck despite the poor visibility. This illustrates how unreliable this configuration can be.

2.2.3 LiDAR, Camera and Radar

This configuration is used by most of the leading industry players such as Waymo and GM Cruise. This configuration has proven to be robust and accurate. However, in order to process the amount of fused data, they require large amounts of processing power which in turn may lead to increased power usage. In addition due to the number of sensors, the cost of these AVs is quite high and therefore not economically feasible.

2.3 Related Research

2.3.1 Classic Computer Vision

With regard to perception, classic computer vision methods have advanced over the years from simple algorithms such as SIFT[23] and SURF[2] which use descriptors to detect interesting points in images for object detection to much higher level representations such as the Histogram of Gradient(HOG)[9] or Haar Features[34] that were used in classifiers. However, these methods have been unreliably slow and not as accurate. In addition these features are only able to model 2D data and therefore when applied to 3D data such as point clouds they are unusable. Furthermore, these feature detectors have to be hand-crafted manually which is a tedious process. As a result, such classical methods can not be used in AVs as they are not as robust and accurate.

2.4 Legal, Ethical and Economic Considerations

The classic ethical dilemma for self driving cars poses a scenario whereby AVs are presented with a situation whereby a fatal accident is inevitable. For example, the AV either has to crash into a

group of people in order to save the life of a passenger or to crash itself and sacrifice the life of the passenger. This dilemma highlights important legal, ethical and economic considerations to be considered by companies involved in the production of AVs and their corresponding systems.

According to [11], there are no public laws that cover the use of independent autonomous vehicles in public spaces. In his article, he cites the fundamental issue as "understanding and accepting the effect of AVs as an independent action by a machine". Due to the lack of public laws on their use, he proposes extending fundamental rights such as the right to life into the framework for creating laws that cover emerging technologies that have an effect on the public, that are otherwise not accounted for in traditional laws.

Bearing this in mind, it is important to note that despite improvements in traffic safety over the years, 94% are still caused by human beings with a large majority of them being fatal with the current human error rate being 1 in 100 million miles. This creates a realistic baseline that can then be extended in evaluating the performance of AV systems. As such, if the risk of automation is lesser than the risk of human vehicle control then the AV would be beneficial. Consequently, the AVs are not to be considered as perfect systems.

With regard to economic considerations, car manufacturers involved in the production of AVs have to ensure that their AVs are able to handle numerous scenarios even if they are rare or considered statistically impossible. They should also be required to take legal liability in case any of their system components are defective and result in failures or accidents. This is important for customer trust without which they will not be able to convince customers to shift to AVs. In addition, for the adoption of AVs to be widespread, they have to be reasonably priced and efficient. With the current prices of the various components and their power consumption, the adoption of AVs at the moment is not viable.

2.4.1 Deep Neural Networks

The emergence and development of Deep Neural Networks such as Convolutional Neural Networks (CNNs)[19] has offset the reliance of classical methods of object detection in images by allowing for features to be automatically learnt by the network without the need for manual feature extraction. This is due to the fact that these networks have numerous number of deep layers that are able to capture these features.

Monocular and Stereo Vision

Due to the high prices of LiDAR units, there were many attempts to develop detection methods that would use standard cameras that were cheap and readily available. Chen et al [7] of Baidu Inc. developed a 3D proposal method from monocular images that would then be used as input to CNNs. In this method, 2D objects were detected in the images and then extruded into 3D using class segmentation, instance level segmentation, shape, contextual features and location priors. This method was one of the best performing monocular object detection method according to the KITTI Benchmark. However, this method only performed well in well lit environments. It's

performance declined greatly in dark scenes. In addition, due to the lack of depth information, the accuracy of the method is entirely dependent on the segmentation of detected 2D objects. Furthermore, the source code of this paper was never released thus making it difficult to validate the results with other datasets

Another attempt at developing lower cost detection systems was through the use of stereo cameras. This involved deploying and calibrating two cameras and through stereo matching and reprojection, a 3D image could be obtained that contains depth information. However as with monocular systems, they suffered the same issue of poor performance in dark scenes. In light of this, such methods would not be viable for use in AVs.

Multimodal

In an effort to improve the accuracy and robustness of AVs, multimodal methods were developed. This involves combining both LiDAR and camera or a camera and a range finder(RGB-D). Both these configurations provide depth information in the form of point clouds. Chen et al further reiterated their monocular approach into MV3D[8] a multiview 3D object detection network. In their work, they were able to fuse LiDAR and camera input to create a bird's eye view of the surrounding environment and further perform object detection of features. Similarly as before, the source code was never released and therefore their results cannot be validated with other datasets.

LiDAR

Point cloud object detection is particularly challenging as established 3D object detection methods for images cannot be applied to point clouds. Point clouds are a set of geometric points in a euclidean space. They are unordered in nature and this presents a particular problem to DNNs as they need to be invariant to permutations of the input set.

Pointnet[28] developed by Qi et al was able to overcome this challenge on small sets of point clouds. By using point clouds as direct input into Recurrent Neural Networks[24] They were able to create a global point cloud signature that could classify and segment 3D objects in the point cloud. They further improved this model into Pointnet++[29] which recursively applied PointNet on nested partitions of the point clouds using a hierarchical neural network. In doing so they were able to capture the local structures of the point clouds as a result of their metric nature and thus achieve better performance than PointNet. For both of these publications, the source code was released for public use. Consequently, they have formed a fundamental foundation for further development of point cloud object detection methods.

An attempt at detecting objects in point clouds from a LiDAR mounted in a vehicle was performed by [20]. In their work, they used a Fully Convolutional Network (FCN) to detect and localise objects in a point cloud as 3D bounding boxes. They were able to achieve this by discretising the point clouds into square grids represented by a 4D array containing the dimensions and a channel indicating if there was a point at that space in the grid. It performed significantly well in the KITTI benchmark. However, the source code was not released.

Finally, [38] were able to create a point cloud based 3D object detection network by adding a Voxel Feature Extraction layer before a RPN that was able to divide the point cloud into voxels that were used as input for the RPN. The published results were well above the state of the art LiDAR object detection methods. As is the case with other publications, the source code was not released and therefore cannot be validated.

METHODOLOGY

3.1 Datasets

While executing the project, three main datasets were used.

- **KITTI** - KITTI is derived from a VW station wagon that was driven in different traffic situations. The car was fitted with a number of sensors including a high resolution greyscale and colour cameras, Velodyne LiDAR and a GPS/IMU inertial navigation system. The dataset includes scenes from urban and rural contexts.
- **CityScapes** - Cityscapes contains multiple street scenes from 50 different cities. In these scenes, 5000 are finely annotated at a pixel-level with 20000 coarsely annotated.
- **University of Bristol Smart Internet Lab** - Derived from the connected and autonomous vehicles project, this dataset contains LiDAR capture files from a stationary vehicle positioned at the Millennium Square in Bristol.

3.2 Context Detection and Characterisation

Determining the context is trivial in a mapped environment where you are able to infer from a map. However in a lab setting where you are only using image data from datasets this becomes quite difficult. This is mostly due to loss of spatio-temporal information if not prior specified or if so, done coarsely. Both VoxelNet and AVOD models were trained using the KITTI dataset using image frames, point clouds. The frames were discontinuous and shuffled into training and testing directories. As such it was difficult to infer the context of the frames prior as no information was provided about the location they were captured in. Notably, this was also evident in the

CityScapes dataset. From this observation, it was necessary to create a context detector for images and furthermore point clouds.

In order to develop, test and train an image and point cloud context detector, I visually classified 3715 images from the KITTI dataset. The corresponding point clouds shared the same context as the classified images.

Image Context Detection

Depending on the number of cars, vegetation, road -Segmentation using Deeplab extracting labelled regions Used to extract regions of interest

PointCloud Context Detection

Intrinsic Shape Signatures

Signature of Histogram of Orientations

Feature Matching

Once the training dataset was classified into to the different contexts. I was able to train a classifier to determine the context of a pointcloud input.

Conditional risk

3.2.0.1 Design

3.2.0.2 Evaluation

Include pictures of results

3.3 Context-Dependent Model Selection

Introduction as to different modes of operation? Multimodal vs Single model? Modes of operation?

3.3.0.1 VoxelNet

3.3.0.2 AVOD

3.3.1 Conclusion

Poor performance in terms of pedestrian and cyclists Good performance in non-urban areas

3.4 Validation with different datasets

RESULTS AND ANALYSIS

CONCLUSION

Following the discussions in the previous chapters, it is clear that there is lack of transparency between industry and academic research on AVs. This can be explained by the industrial competition with companies trying to establish themselves as the leaders in development of AVs. However, I believe that for expedited development of AVs there needs to be more collaboration between these two communities. This will provide an overall benefit as it will prompt the development of standards and regulations.

With regard to LiDAR-based object detection, new advances in solid state technology has allowed for the development of cheap LiDAR units such as Quanergy's 250\$ unit. This presents an opportunity for AV manufacturers to switch to these cheaper units with no fear of losing accuracy. It is therefore my wish to implement my solution to be able to be used in solid-state LiDARs.

In pursuing this project, I intend to engage in an exploratory investigation in order to establish various options in the implementation of the various tasks listed in chapter ???. In doing so I will be able to provide an informed evaluation of the different factors affecting the development of AVs. Furthermore, if time allows, the issue of lack of interpretability of deep learning models will be explored. Determinism of the AV systems is an important aspect in terms of safety.

APPENDIX



APPENDIX A

LiDAR	Hor FOV	Ver FOV	Range	Angular Resolution	Points/second	Channels
<i>VLS-128</i>	360°	+15°to -25°	300m	0.11°	~9.6 Million	128
<i>HDL-64E</i>	360°	26.9°	120m	~0.4°	~2.2 Million	64
<i>HDL-32E</i>	360°	+10°to -30°	80m-100m	0.1°	~1.39 Million	32
<i>VLP-32C</i>	360°	+15°to -25°	200m	0.1°	~1.2 Million	32
<i>VLP-16</i>	360°	± 15°	100m	0.1°	600,000	16

Table A1: Velodyne LiDAR Family

BIBLIOGRAPHY

- [1] M. ABADI, P. BARHAM, J. CHEN, Z. CHEN, A. DAVIS, J. DEAN, M. DEVIN, S. GHEMAWAT, G. IRVING, M. ISARD, ET AL., *Tensorflow: A system for large-scale machine learning.*, in OSDI, vol. 16, 2016, pp. 265–283.
- [2] H. BAY, T. TUYTELAARS, AND L. VAN GOOL, *Surf: Speeded up robust features*, in European conference on computer vision, Springer, 2006, pp. 404–417.
- [3] J.-F. BONNEFON, A. SHARIFF, AND I. RAHWAN, *Autonomous vehicles need experimental ethics: Are we ready for utilitarian cars?*, arXiv preprint arXiv:1510.03346, (2015).
- [4] I. BROWN, *Human factors: the journal of the human factors and ergonomics society*, Driver Fatigue, 36 (1994), pp. 298–314.
- [5] S. D. BROWN, R. J. FRANCIS, J. ROSE, AND Z. G. VRANESIC, *Field-programmable gate arrays*, vol. 180, Springer Science & Business Media, 2012.
- [6] M. BUEHLER, K. IAGNEMMA, AND S. SINGH, *The DARPA urban challenge: autonomous vehicles in city traffic*, vol. 56, springer, 2009.
- [7] X. CHEN, K. KUNDU, Z. ZHANG, H. MA, S. FIDLER, AND R. URTASUN, *Monocular 3d object detection for autonomous driving*, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2147–2156.
- [8] X. CHEN, H. MA, J. WAN, B. LI, AND T. XIA, *Multi-view 3d object detection network for autonomous driving*, in IEEE CVPR, vol. 1, 2017, p. 3.
- [9] N. DALAL AND B. TRIGGS, *Histograms of oriented gradients for human detection*, in Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, vol. 1, IEEE, 2005, pp. 886–893.
- [10] D. J. FAGNANT AND K. KOCKELMAN, *Preparing a nation for autonomous vehicles: opportunities, barriers and policy recommendations*, Transportation Research Part A: Policy and Practice, 77 (2015), pp. 167–181.
- [11] T. M. GASSER, *Fundamental and special legal questions for autonomous vehicles*, in Autonomous Driving, Springer, 2016, pp. 523–551.

- [12] A. GEIGER, P. LENZ, C. STILLER, AND R. URTASUN, *Vision meets robotics: The KITTI dataset*, International Journal of Robotics Research (IJRR), (2013).
- [13] A. GEIGER, P. LENZ, AND R. URTASUN, *Are we ready for autonomous driving? the kitti vision benchmark suite*, in Conference on Computer Vision and Pattern Recognition (CVPR), 2012.
- [14] R. GIRSHICK, *Fast r-cnn*, in International Conference on Computer Vision (ICCV), 2015.
- [15] P. HENRY, M. KRAININ, E. HERBST, X. REN, AND D. FOX, *Rgb-d mapping: Using depth cameras for dense 3d modeling of indoor environments*, in In the 12th International Symposium on Experimental Robotics (ISER, Citeseer, 2010.
- [16] INTEL, 2018.
<https://www.intel.co.uk/content/www/uk/en/automotive/go-automated-driving.html>.
- [17] J. KU, A. HARAKEH, AND S. L. WASLANDER, *In defense of classical image processing: Fast depth completion on the cpu*, arXiv preprint arXiv:1802.00036, (2018).
- [18] J. KU, M. MOZIFIAN, J. LEE, A. HARAKEH, AND S. WASLANDER, *Joint 3d proposal generation and object detection from view aggregation*, arXiv preprint arXiv:1712.02294, (2017).
- [19] S. LAWRENCE, C. L. GILES, A. C. TSOI, AND A. D. BACK, *Face recognition: A convolutional neural-network approach*, IEEE transactions on neural networks, 8 (1997), pp. 98–113.
- [20] B. LI, *3d fully convolutional network for vehicle detection in point cloud*, arXiv preprint arXiv:1611.08069, (2016).
- [21] B. LI, T. ZHANG, AND T. XIA, *Vehicle detection from 3d lidar using fully convolutional network*, arXiv preprint arXiv:1608.07916, (2016).
- [22] S.-C. LIN, Y. ZHANG, C.-H. HSU, M. SKACH, M. E. HAQUE, L. TANG, AND J. MARS, *The architectural implications of autonomous driving: Constraints and acceleration*, in Proceedings of the Twenty-Third International Conference on Architectural Support for Programming Languages and Operating Systems, ACM, 2018, pp. 751–766.
- [23] D. G. LOWE, *Object recognition from local scale-invariant features*, in Computer vision, 1999. The proceedings of the seventh IEEE international conference on, vol. 2, Ieee, 1999, pp. 1150–1157.
- [24] L. MEDSKER AND L. JAIN, *Recurrent neural networks*, Design and Applications, 5 (2001).
- [25] MOBILEYE, *Enabling autonomous - mobileeye*, 2018.
<https://www.mobileye.com/future-of-mobility/mobileeye-enabling-autonomous/>.

- [26] A. NEWELL AND S. K. CARD, *The prospects for psychological science in human-computer interaction*, Human-computer interaction, 1 (1985), pp. 209–242.
- [27] D. A. POMERLEAU, *Alvin: An autonomous land vehicle in a neural network*, in Advances in neural information processing systems, 1989, pp. 305–313.
- [28] C. R. QI, H. SU, K. MO, AND L. J. GUIBAS, *net: Deep learning on point sets for 3d classification and segmentation*, Proc. Computer Vision and Pattern Recognition (CVPR), IEEE, 1 (2017), p. 4.
- [29] C. R. QI, L. YI, H. SU, AND L. J. GUIBAS, *Pointnet++: Deep hierarchical feature learning on point sets in a metric space*, in Advances in Neural Information Processing Systems, 2017, pp. 5105–5114.
- [30] S. REN, K. HE, R. GIRSHICK, AND J. SUN, *Faster r-cnn: Towards real-time object detection with region proposal networks*, in Advances in neural information processing systems, 2015, pp. 91–99.
- [31] N. RESEARCH, *Navigant research leaderboard: Automated driving vehicles*, Jan 2018.
<https://www.navigantresearch.com/research/navigant-research-leaderboard-automated-driving-vehicles>.
- [32] M. SIMON, S. MILZ, K. AMENDE, AND H.-M. GROSS, *Complex-YOLO: Real-time 3D Object Detection on Point Clouds*, ArXiv e-prints, (2018).
- [33] M. J. S. SMITH, *Application-specific integrated circuits*, vol. 7, Addison-Wesley Reading, MA, 1997.
- [34] P. VIOLA AND M. JONES, *Rapid object detection using a boosted cascade of simple features*, in Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on, vol. 1, IEEE, 2001, pp. I–I.
- [35] WAYMO, *Waymo safety report*, 2018.
<https://waymo.com/safety/>.
- [36] F. YANG, W. CHOI, AND Y. LIN, *Exploit all the layers: Fast and accurate cnn object detector with scale dependent pooling and cascaded rejection classifiers*, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2129–2137.
- [37] M. D. ZEILER AND R. FERGUS, *Visualizing and understanding convolutional networks*, in European conference on computer vision, Springer, 2014, pp. 818–833.
- [38] Y. ZHOU AND O. TUZEL, *Voxelnet: End-to-end learning for point cloud based 3d object detection*, arXiv preprint arXiv:1711.06396, (2017).

