Indian Institute of Information Technology
Allahabad

DEPARTMENT OF INFORMATION TECHNOLOGY(IT)

Seventh Semester Project Report on

3D Multi-Object Tracker

Under the supervision of
Dr. Anshu S. Anand (Asst. Professor)

Submitted by:
Joel Swapnil Singh, (ITM2017002)
Pratham Singh, (IRM2017006)

Contents

1 Certificate from the Supervisor

I hereby declare that the project work entitled **"3D Multi-Object Tracker"** submitted at Indian Institute of Information Technology, Allahabad, is the bonafide work of **Pratham Singh (IRM2017006) and Joel Swapnil Singh (ITM2017002)**. It is an authentic record of our study carried out from June 2020 till September 2020 under my guidance. Due acknowledgments have been made in the text to all the materials used. The project was done in full compliance with the requirements and constraints of the prescribed curriculum.

Dr. Anshu S. Anand
(Assistant Professor)
Department of Information Technology
Indian Institute of Information Technology Allahabad

2 Abstract

3D multi-object tracking (MOT) is an essential component for many applications such as autonomous driving and assistive robotics. Recent work on 3D MOT focuses on developing accurate systems giving less attention to practical considerations such as computational cost and system complexity. In contrast, this work proposes a simple real-time 3D MOT system.

3 Introduction

MOT is an essential component for many real-time applications such as autonomous driving and assistive robotics. Due to advancements in object detection, there has been much progress on MOT. Although 2D object detection is relatively mature and has been widely used in the industry, 3D object detection from 2D imagery is a challenging problem, due to the lack of data and diversity of appearances and shapes of objects within a category. While 2D prediction only provides 2D bounding boxes, therefore by extending prediction to 3D, the main objective is to capture the object's size, position and orientation in the world, leading to a variety of applications in robotics, self-driving vehicles, image retrieval, and augmented reality. Unlike other filter based MOT systems which define the state space of the filter in the 2D space or bird's eye view, we extend the state space of the objects to the 3D space, including 3D location, 3D size, 3D velocity and heading orientation.

4 Motivation

MOT being an important component for many real-time applications, the greatest being the driving automation which can help reduce the number of crashes on our roads. Government data identifies driver behavior or error as a factor in 94 percent of crashes, and self-driving vehicles can help reduce driver error. Higher levels of autonomy have the potential to reduce risky and dangerous driver behaviors. The greatest promise may be reducing the devastation of impaired driving, drugged driving, unbelted vehicle occupants, speeding and distraction. Also, People with disabilities, like the blind, are capable of self-sufficiency, and highly automated vehicles can help them live the life they want. Automated driving systems could impact our pocketbooks in many ways viz. They can help avoid the costs of crashes, including medical bills, lost work time and vehicle repair. Fewer crashes may reduce the costs of insurance. Which brings us to work on this project more enthusiastically.

5 Problem Definition

One of the key tasks in autonomous driving is objet detetion. Autonomous ars are often fitted with many sensors like a amera, LiDAR. Although Convolutional Neural Networks is a state-of-the-art 2D objet detetion tehnology, it does not perform well in the 3D point loud due to the sparse sensor data, so new tehniques are needed. 3D objet detetion networks operate in a 3D point loud provided by a range distane sensor. For autonomous vehiles to work, it is very important for the pereption omponent to detet the real world objets with both high auray and fast inferene. Being a vital omponent for many real-time appliations. Our main goal is to implement a novel neural network arhiteture along with the training and optimization details for deteting 3D objets in point loud data.

6 Literature Review

Rapid development of 3D sensor technology has motivated researchers to develop efficient representations to detect and localize objects in point clouds. These hand-crafted features yield satisfactory results when rich and detailed 3D shape information is available. However their inability to adapt to more complex shapes and scenes, and learn required in variances from data resulted in limited success for uncontrolled scenarios such as autonomous navigation. Given that images provide detailed texture information, many algorithms inferred the 3D bounding boxes from 2D images . However, the accuracy of image-based 3D detection approaches are bounded by the accuracy of the depth estimation. Several LIDAR based 3D object detection techniques utilize a voxel grid representation. Some encode each nonempty voxel with 6 statistical quantities that are derived from all the points contained within the voxel. Some fuses multiple local statistics to represent each voxel. Some computes the truncated signed distance on the voxel grid. Some uses binary encoding for the 3D voxel grid. Some introduces a multi-view representation for a LiDAR point cloud by computing a multi-channel feature map in the bird's eye view and the cylindral coordinates in the frontal view. Several other studies project point clouds onto a perspective view and then use image-based feature encoding schemes . There are also several multi-modal fusion methods that combine images and LiDAR to improve detection accuracy. These methods provide improved performance compared to LiDAR-only 3D detection, particularly for small objects (pedestrians, cyclists) or when the objects are far, since cameras provide an order of magnitude more measurements than LiDAR. However the need for an additional camera that is time synchronized and calibrated with the LiDAR restricts their use and makes the solution more sensitive to sensor failure modes. In this work we focus on LiDAR-only detection.

7 Proposed Methodology

Before we plunge into the methodology we should understand what LiDAR and Point clouds are and how do they correlate to each other: Basically, LiDAR is a remote sensing process which collects measurements used to create 3D models and maps of objects and environments. Using ultraviolet, visible, or near-infrared light, LiDAR gauges spatial relationships and shapes by measuring the time it takes for signals to bounce off objects and return to the scanner where Point clouds are a powerful and dynamic information storage technology. They are used as a middleman to turn the raw data collected by LiDAR processes into 3D models. However, they can be used to store and manipulate any spatial information. A 3D point cloud is a collection of data points analogous to the real world in three dimensions. Each point cloud is an unordered set of lidar points. The data format of each returned lidar point is a 4-tuple formed by its coordinate with respect to the lidar coordinate frame as well as its intensity p the KITTI dataset, p is a normalized value between 0 and 1, and it depends on the characteristics of the surface the lidar beam reflects from.

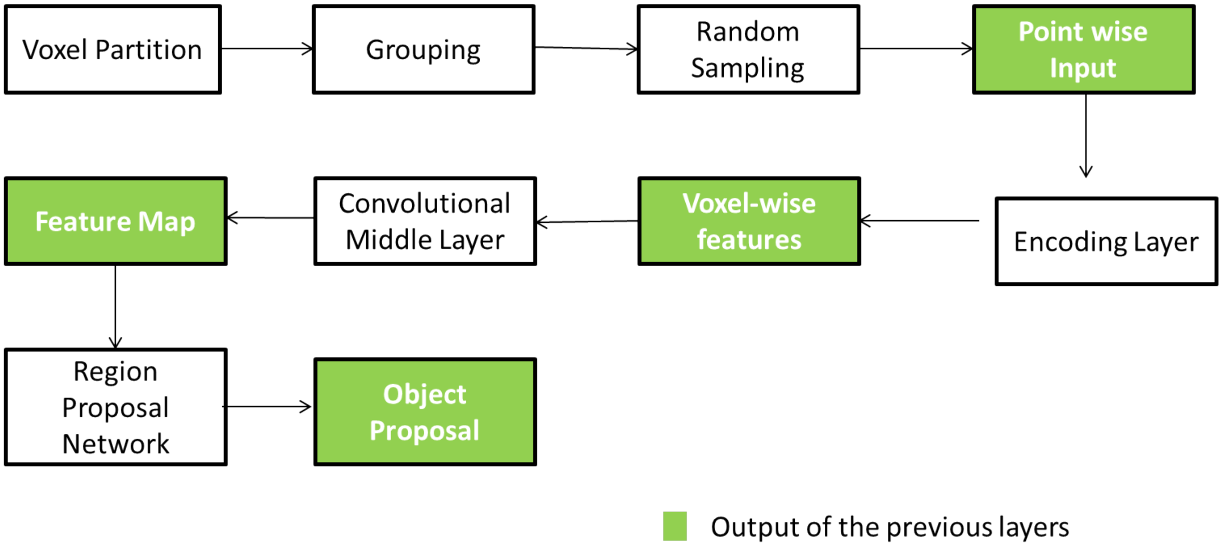


Figure 1: Neural Net Architecture

7.1 Dataset

KITTI is one of the well known benchmarks for 3D Object detection. Working with this dataset requires some understanding of what the different files and their contents are. Goal here is to do some basic manipulation and sanity checks to get a general understanding of the data. 4 different types of files from the KITTI 3D Objection Detection dataset as follows are used here: camera-2 image (.png), camera-2 label (.txt), calibration (.txt), velodyne point cloud (.bin).

For each frame , there is one of these files with same name but different extensions. The image files are regular png file and can be displayed by any PNG aware software. The label files contains the bounding box for objects in 2D and 3D in text. Each row of the file is one object and contains 15 values , including the tag (e.g. Car, Pedestrian, Cyclist). The 2D bounding boxes are in terms of pixels in the camera image . The 3D bounding boxes are in 2 co-ordinates. The size (height, weight, and length) are in the object co-ordinate , and the center on the bounding box is in the camera co-ordinate. The point cloud file contains the location of a point and its reflectance in the lidar co-ordinate. Two tests need to be done here. The first test is to project 3D bounding boxes from label file onto image. Second test is to project a point in point cloud coordinate to image. The algebra is simple as follows. The first equation is for projecting the 3D bouding boxes in reference camera co-ordinate to camera-2 image.

The second equation projects a velodyne co-ordinate point into the camera-2 image.

y-image = P2 * R0-rect * R0-rot * xRefCoord

y-image = P2 * R0-rect * TrVeloToCam * xVeloCoord

In the above, R0-rot is the rotation matrix to map from object coordinate to reference coordinate.

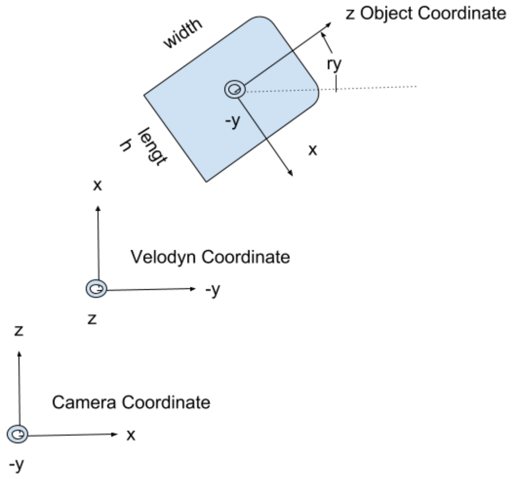


Figure 2: The many coordinate system used in dataset

7.2 3D Object Detection Task

The main characteristic of point clouds that prevent us from readily adopting CNN object detection neural networks is their permutation invariance property whereas CNN object detection networks assume their inputs are ordered data structures represented in grid forms. Examples of grid ordered data structures are images meaning that changing orders of pixels of images modifies the contents of images. We focus on a category of 3D object detection networks that rely on ordered grid tensors to represent point clouds in order to remove their permutation invariant constraint. In particular, these detection networks instead of directly consuming raw point clouds, take intermediate representations of point clouds in the form of image-like ordered grids as their inputs. The main characteristic of this category of 3D object detection networks is to represent point clouds similar to images as structured grids so that they can benefit from the existing image-based CNN object detection networks. One such representation of point clouds is 3D voxelization which is realized by quantization of 3D cuboid subspaces surrounding lidar sensors. 3D cuboids are chosen because of their geometric shape compatibility with cubical shapes of tensors.

7.3 Working

First, the 3D space is divided into equally spaced voxels. The points are grouped according to the voxel they belong to. The first layer of the Network is an encoding layer which transforms a group of points within each voxel into a feature representation, a 4D tensor. The layer is called the Voxel Feature Encoding layer. Then a Convolutional Neural Networks extracts the complex features and outputs the confidence values and the regression values of the bounding boxes.. The output of the convolutional middle layer is the input of the Region Proposal Network (RPN) layer. The RPN produces a probability score map and a regression map. The loss is the sum of the classification loss and the regression loss.

7.4 Training and Testing

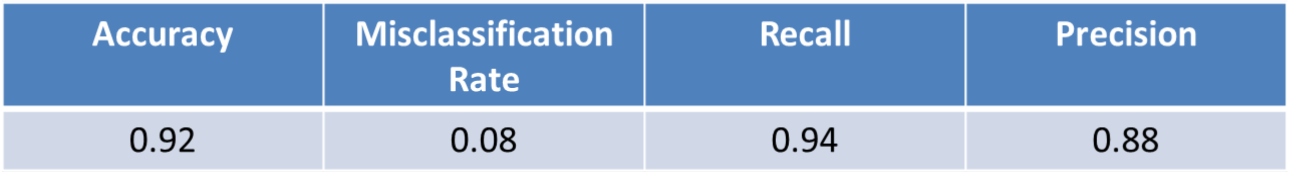
The hyperparameters are modified to have better experimental results. An anchor box is considered positive if it has the highest IOU with any ground truth or the IOU is above 0.6 with any ground truth boxes. An anchor box is evaluated as negative if the IOU with all ground truth box is less than 0.45. Those anchor boxes are considered as don't cares which have their IOU value between 0.45 and 0.6.

8 Problem Faced

8.1 Representation transformation

Cameras are usually mounted on car roofs on some prototype self-driving cars, or behind rear-view mirrors like a normal dash-cam. Therefore camera images typically have perspective views of the world. This view is easy to understand for human drivers as it resembles what we see during driving and but poses two challenges for computer vision: occlusion and scale variation due to distance.

9 Results



Accuracy	Misclassification Rate	Recall	Precision
0.92	0.08	0.94	0.88

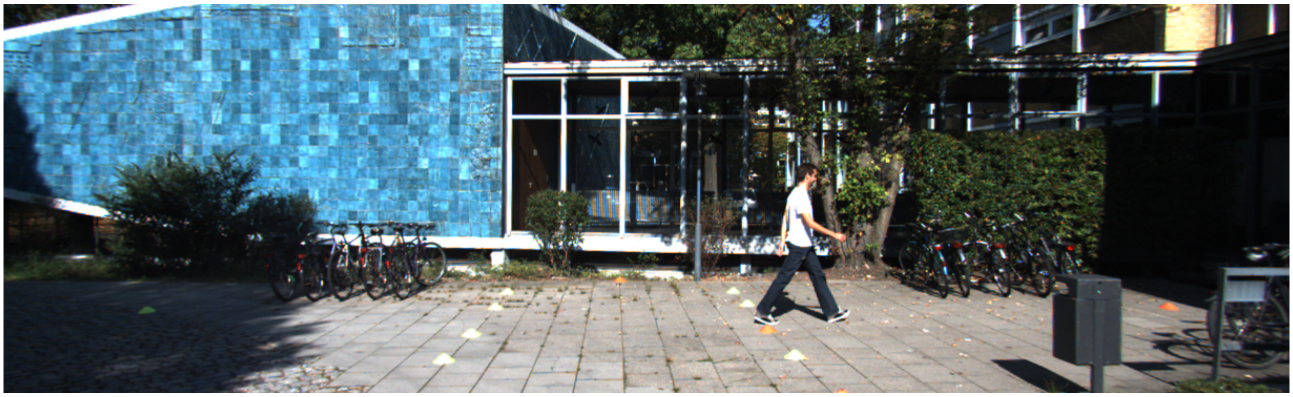


Figure 3: Input Image

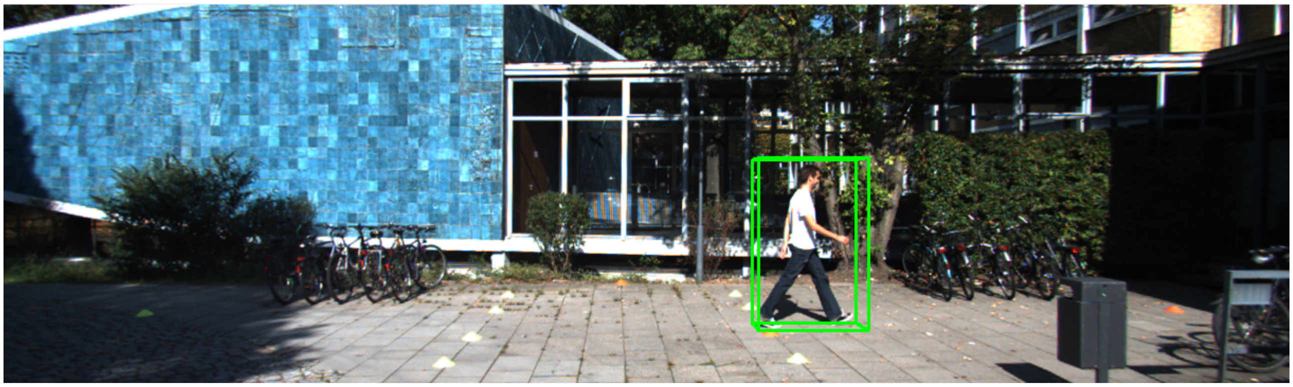


Figure 4: Output Image

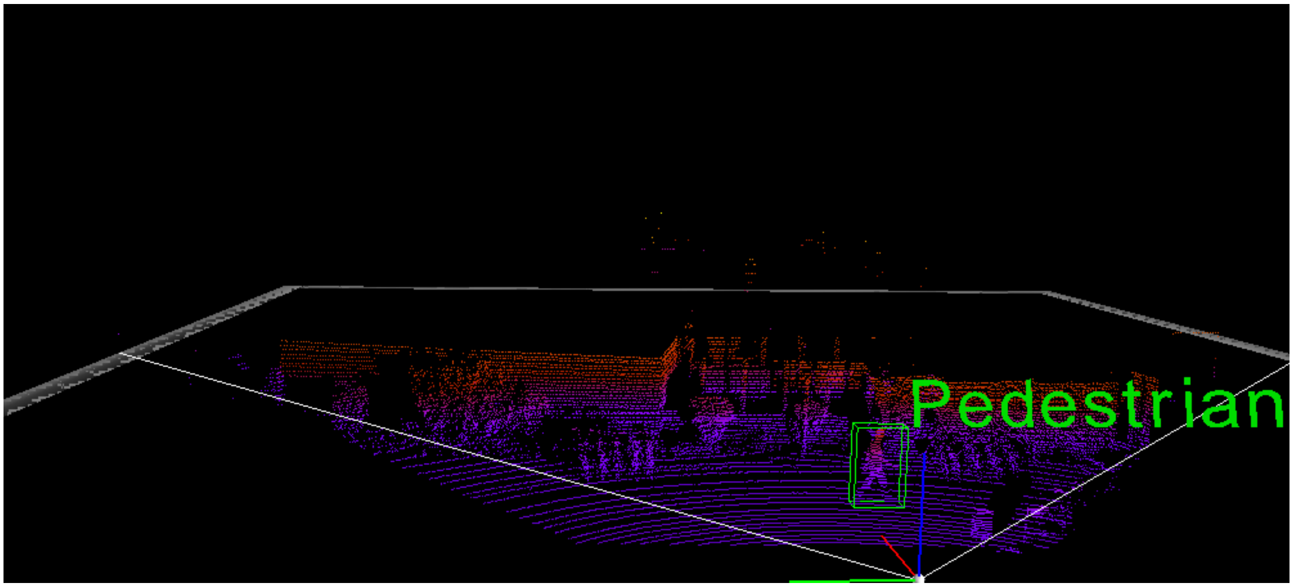


Figure 5: Bird-Eye View

10 Conclusion

We proposed an accurate, simple and real-time system for online 3D MOT. Through extensive experiments on the KITTI 3D MOT dataset, our system establishes new state-of-the-art 3D MOT performance while achieving the fastest speed. We hope that our system will serve as a solid baseline on which others can easily build on to advance the state-of-the-art in 3D MOT.

11 References

[1] S. Wang, D. Jia, and X. Weng, “Deep Reinforcement Learning for Autonomous Driving,” arXiv:1811.11329, 2018.

[2] A. Geiger, P. Lenz, and R. Urtasun, “Are We Ready for Autonomous Driving? the KITTI Vision Benchmark Suite,” CVPR, 2012.

[3] H. Karunasekera, H. Wang, and H. Zhang, “Multiple Object Tracking with Attention to Appearance, Structure, Motion and Size,” IEEE Access, 2019.

[4] W. Tian, M. Lauer, and L. Chen, “Online Multi-Object Tracking Using Joint Domain Information in Traffic Scenarios,” IEEE Transactions on Intelligent Transportation Systems, 2019.

[5] Xinshuo Weng, Jianren Wang, David Held and Kris Kitani, ”3D Multi-Object Tracking: A Baseline and New Evaluation Metrics”, arXiv 22 July 2020.