3D Multi-modal Multi-object Tracking via Machine Learning and Analytic Collision Risk Calculation for Autonomous Vehicles Navigation: The Feature Exctraction Module

Spathoulas Dimitrios
Department of Industrial Engineering and Management
Democritus University of Thrace

CONTENTS

LIST OF FIGURES

3D Multi-modal Multi-object Tracking via Machine Learning and Analytic Collision Risk Calculation for Autonomous Vehicles Navigation: The Feature Exctraction Module

***Abstract*—This report focuses on the Feature Extraction Module, which is the first of the three main components in my master's thesis [1] [2] [3] [4]. The module is divided into two subcomponents: the 3D detection and geometric feature extractor, and the appearance feature extractor. The appearance feature extractor itself is split into two main processes: the projection and the appearance extraction process. To capture full environmental coverage, we utilize seven sensors—one LiDAR and six cameras—sourced from the Nuscenes dataset [1]. The point cloud data from the LiDAR are processed by the 3D detector CenterPoint [2], while the camera data are processed by the 2D detector and segmentor Mask R-CNN [3].**

I. Introduction

CenterPoint is a 3D detector designed to process point cloud data. Its name originates from a center-based, anchor-free detection paradigm, where object centers are represented as peaks on a Gaussian heatmap, thus eliminating the need for traditional anchor boxes. The architectural design incorporates parallel multi-head neural network components that systematically predict detection attributes through specialized predictive heads. Mask R-CNN innovates beyond Faster R-CNN [4] by introducing a Fully Convolutional Network (FCN) for end-to-end pixel-wise segmentation and the RoI Align layer, which uses bilinear interpolation to precisely sample feature maps, overcoming previous limitations in object detection and instance segmentation accuracy.

II. 3D Detection and Geometric Feature Extractor

First step is the 3D object detection based on the point cloud data. The detections are relative to the LiDar's reference frame. Our implementation functions for all CenterPoint's pre-trained Nuscenes models (different voxel sizes, 0.075 is preferred). For the extraction of geometric features we transform each x,y detection's center back to the discrete feature map created by the instance's point cloud. The transformation is as follows:

$$x_s = \frac{x_c - x_{min}}{s_x \cdot v_x}, \quad y_s = \frac{y_c - y_{min}}{s_y \cdot v_y} \tag{1}$$

Where x_c, y_c object center in LiDAR reference frame, x_{min}, y_{min} point cloud range, v_x, v_y voxel size and s_x, s_y feature map stride.
We then extract a 3x3 region of 512 layers around each center, encompassing the spatial pattern of the area. For better feature preservation and stability, we use bilinear interpolation for each point in the grid.

III. Projection and Camera Feature Extraction

To determine if each detection projects onto a camera's plane, we first retrieve the extrinsic transformation matrices for the lidar-to-ego-to-world-to-ego-to-camera transformations. Subsequently, we apply the respective intrinsic camera matrices to each detection.
A projection is valid if and only if:

1) A rectangular area can be formulated within the image's pixel bounds.
2) The projection lies in front of the camera, effectively filtering out projections behind the camera.

In the Nuscenes dataset, the cameras' Fields of View have spatial overlaps, making it possible for the same object to be validly projected onto multiple camera planes. To address this, we retain only the projection with the largest rectangular area in an image among consecutive valid projections of the same object.
For appearance feature extraction, we modify Mask R-CNN's Region Proposal Network to propose only the projected molded regions on the feature maps. This process generates a 1024-dimensional vector that encapsulates the visual information of a specific region within the image for each projected detection.

IV. Manifold Learning

Using manifold learning, particularly UMAP [5], we can demonstrate that the features are stable across multiple validation runs, and interpret various class-specific attributes. Specifically Fig. 1, shows that the "bus" class exhibits at least four distinct principal components in its feature space. This indicates the presence of multiple sub-classes within the "bus" category, each exhibiting unique characteristics. The UMAP parameters used were 20 neighbors, cosine metric, and a minimum distance of 0.1.

[1] https://github.com/DimSpathoulas/PointCloud_Feature_Extractor
[2] https://github.com/DimSpathoulas/2D_FEATURE_EXTRACTOR
[3] https://github.com/DimSpathoulas/GeomApp_3MOT
[4] https://github.com/DimSpathoulas/Collision_Risk_Calculation

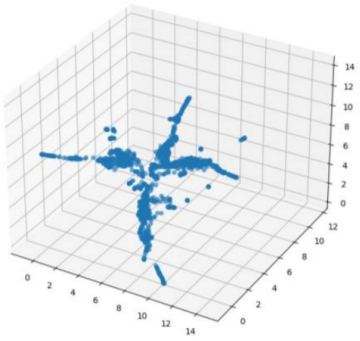

Fig. 1. UMAP representation for bus point cloud features

APPENDIX A
TECHNICALITIES

Both extraction processes have been optimized for resource efficiency. However, even with a setup consisting of 125 GB RAM, 20 CPU cores, and two GPUs (RTX 3090 24 GB), we can only process detections with a prediction confidence of 57% or higher. This limitation significantly reduces interpretability and severely impacts tracking performance. Multiprocessing optimization is a promising avenue, particularly for the appearance extraction module.

REFERENCES

[1] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020.

[2] Tianwei Yin, Xingyi Zhou, and Philipp Krahenbuhl. Center-based 3d object detection and tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11784–11793, 2021.

[3] Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2961–2969, 2017.

[4] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1137–1149, 2016.

[5] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.