PhD Forum Abstract: Sensor Fusion for Vehicle-side and Roadside 3D Object Detection and Tracking

Yao Li
University of Science and Technology of China
Hefei, China
zkdly@mail.ustc.edu.cn

ABSTRACT

We focus on the sensor fusion for vehicle-side and roadside 3D object detection and tracking. Although quite a few sensor fusion algorithms have been proposed, some of which are top-ranked on various leaderboards, a systematic study on how to integrate three crucial sensors (LiDAR, camera and millimeter-wave Radar sensors) to develop effective multi-modal 3D object detection and tracking for vehicle-side perception is still missing. Therefore, we first study the three sensors' strengths and weaknesses carefully, then compare several different fusion strategies to maximize their utility. Finally, based on the lessons learnt, we propose a simple yet effective multi-modal 3D object detection and tracking framework (namely EZFusion). Without fancy network modules, our proposed EZFusion makes remarkable improvements over the LiDAR-only baseline, and achieves comparable performance. For intelligent transportation, far-range perception with roadside sensors is vital. The main challenge of far-range perception is performing accurate object detection and tracking under far distances (e.g., > 150m) at a low cost. To cope with such challenges, deploying both millimeter wave Radars and high-definition cameras, and fusing their data has become a common practice. Towards this goal, the first question is to conduct the association on the 2D image plane or the BEV plane. We argue that the former is more suitable because the magnitude of location errors in the perspective projection points is smaller at far distances on the 2D plane, leading to more accurate association. Thus, we first project the Radar points to the 2D plane and then associate them with the camera-based 2D object locations. Subsequently, we map the camera-based object locations to the BEV plane through inverse projection mapping (IPM) with the corresponding depth information from the Radar data. Finally, we engage a BEV tracking module to generate target trajectories. Our system is capable of achieving an average location accuracy of 1.3m when we extend the detection range up to 500m.

1 INTRODUCTION

1.1 A close look at the integration of LiDAR, millimeter-wave Radar, and camera for accurate 3D object detection and tracking

We first introduce the integration of three typical sensors LiDAR, millimeter-wave Radar and camera for autonomous driving. 3D object perception has attracted a great deal of research interest in autonomous systems. Meanwhile, 3D detection and tracking provides accurate location information, which can enhance navigation and safety in autonomous vehicles. Recent trends use multiple data sources to improve detection performance. For examples, combining camera images with LiDAR point clouds [4] or Radar point clouds can enhance 3D object detection and tracking [3], and obtain more accurate depth information [2] respectively. These techniques significantly boost detection and tracking performance.

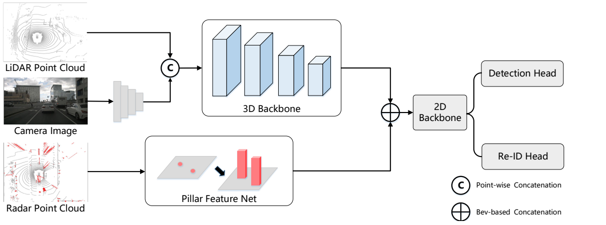

Figure 1: Overview of the EZFusion pipeline. We employ both point-wise and BEV-based fusion strategies.

However, how to effectively integrate these sensors has largely remained a hit-or-miss process. Several important questions remain unanswered in this space. To name a few, how much does each type of sensor data contribute to the detection and tracking performances? How should the fusion among the three modalities be carried out? Should one fuse LiDAR and camera data first and then add Radar, or in a different order? Should one employ point-wise fusion between LiDAR and Radar points or convert them to the bird's eye view (BEV) plane and perform fusion there?

To answer these questions, we take a close look at the integration of these three types of sensor data by conducting a systematic comparison. In particular, we focus our scope on the two important design questions: (1) what to fuse, and (2) how to fuse. For the what-to-fuse question, we first provide a comparison between the LiDAR, Radar, and camera sensors from several perspectives. Then we study several different fusion input configurations: LiDAR only, LiDAR-camera fusion, LiDAR-Radar fusion, and LiDAR-Radar-camera fusion, and quantify their detection and tracking performance carefully. While the three-way fusion (LiDAR-Radar-camera) provides the best performance as expected, each modality contributes to different metrics with a different scale. Extensive experiments reveal the unique effects of different modalities with their own physical properties on detection and tracking systems. For the how-to-fuse question, we compare three fusion strategies with all three types of sensor data, and identify the most efficient fusion strategy – first painting each LiDAR point with corresponding image features, and then combining the LiDAR-camera features based on painted points and the Radar features on the BEV plane.

After exploring the above design choices, we've developed a simple yet effective 3D detection and tracking pipeline EZFusion, integrating three modalities, as shown in Fig. 1. EZFusion only contains the bare minimum network modules to demonstrate the effectiveness of our fusion framework. More sophisticated modules

979-8-3503-6201-5/24/$31.00 ©2024 IEEE
DOI 10.1109/IPSN61024.2024.00061

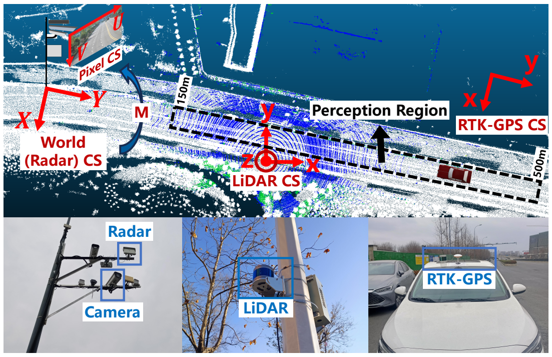

Figure 2: Our OWL testbed.

can be easily added later. Our EZFusion has an improvement of 4.7% in mAP, and 6.7% in AMOTA compared to the LiDAR-only baseline on the nuScenes validation set. Our results are comparable to state-of-the-art fusion methods, while our framework doesn't need much bulkier operations.

1.2 A practical roadside Radar-camera fusion system for far-range perception

Nowadays, roadside perception is crucial for smart transport systems, which can provide real-time information about the road environment, traffic incidents, etc., potentially facilitating autonomous driving and traffic monitoring. However, due to cost constraints, it's not feasible to densely deploy roadside sensors. Consequently, far-range object detection and tracking (e.g. $> 150m$) is urgently needed. To achieve accurate detection at such distances, it's common to use both millimeter wave Radar and high-definition cameras and combine their data for joint perception.

Fig. 2 shows our OWL testbed with a Radar and HD camera installed next to each other. Meanwhile, aligning their data is challenging due to their different perspectives and measurement noises. Feature-level fusion methods use learnable models to align sensor data, enhancing detection performance but needing annotated data. Vehicle-side data is easy to annotate due to near-range perception, while roadside scenes require far-range detection, making annotation difficult. Therefore, we use target-level fusion, which aligns Radar and camera data on 2D or BEV planes.

Though BEV-based fusion [1] is increasingly used in vehicle sensor fusion, it's less effective for long-range roadside perception. BEV-based fusion typically uses Inverse Perspective Mapping (IPM) to map 2D pixels to the BEV plane and associate them with Radar points. However, the IPM process has limited precision due to the coarser resolution of image pixels at larger distances. Indeed, we find that the magnitude of location errors in IPM points is greater in far ranges on the BEV plane, whereas the magnitude of errors in perspective projection points is smaller on the 2D plane. Thus we believe that 2D-based association is well-suited for our application. Another key issue is the transformations between views, *i.e.*, the 2D-to-BEV IPM, and BEV-to-2D perspective projection, which heavily hinge on the accuracy of transformation parameters. Any unexpected small motions of the sensors caused by windy weather likely undermine these parameters' accuracy, especially in distant ranges. Although [5] proposes an automatic camera calibration method to refine the IPM, relying on special environmental cues and precise vanishing point detection, they often struggle in the far-range scenes. Besides, it is prohibitively costly to deploy precise GPS to adjust transformation parameters for each sensor motion.

To address these challenges, we propose a roadside Radar-camera fusion system FARFusion for far-range traffic monitoring. Our proposed system has two main ideas. Firstly, the 2D plane is more suitable for associating the two types of sensor data than the BEV plane for far-range perception. Specifically, we propose a 2-stage fusion strategy. The first fusion stage occurs on the 2D plane, in which we project the Radar points from the BEV to the 2D plane and then associate them with the camera-based object locations that are modeled as a point on each object. Upon the point-wise target association, we can accurately map the camera-based object locations to the BEV plane by leveraging the depth information from the corresponding Radar points. Subsequently, the second fusion stage occurs on the BEV plane, in which we perform a point-based target tracking module to combine the detected locations from the first stage with different modalities and generate continuous trajectories for traffic monitoring. Secondly, we propose a transformation parameters refining approach based on our depth scaling technique to refine the transformation parameters.

In the future, we plan to deploy more LiDARs on our OWL testbed to provide more annotations. With additional annotated data, we can exploit more effective camera-based depth estimation methods for far-range scenes. The current method of target-level fusion highly relies on the depth information provided by Radar, and it does not make full use of the camera image to determine the depth. Furthermore, we will investigate the feature-level methods for integrating roadside Radar and camera data. We attempt to achieve higher detection performance with more annotation data.

2 A SHORT INTRODUCTION TO THE PH.D. PROGRAM

I am currently pursuing a Ph.D. degree with the Computer Science and Technology, University of Science and Technology of China (USTC) in Hefei, China. I am advised by Prof. Yanyong Zhang at USTC. I began my Ph.D. studies in September 2021 and anticipate receiving my Ph.D. degree in December 2024. I am focused on studying the application of sensor fusion techniques in 3D object detection and tracking, specifically for enhancing perception in autonomous driving and intelligent transportation systems.

REFERENCES

[1] Zhijian Liu, Haotian Tang, Alexander Amini, Xinyu Yang, Huizi Mao, Daniela L Rus, and Song Han. 2023. Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. In *IEEE Int. Conf. Robot. Autom.* IEEE, 2774–2781.

[2] Ramin Nabati and Hairong Qi. 2021. Centerfusion: Center-based radar and camera fusion for 3d object detection. In *Proc. IEEE Winter Conf. Applications Comput. Vis.* 1527–1536.

[3] Abhijeet Shenoi, Mihir Patel, JunYoung Gwak, Patrick Goebel, Amir Sadeghian, Hamid Rezatofighi, Roberto Martin-Martin, and Silvio Savarese. 2020. Jrmot: A real-time 3d multi-object tracker and a new large-scale dataset. In *Proc. Int. Conf. Intell. Robots Syst.* IEEE, 10335–10342.

[4] Sourabh Vora, Alex H Lang, Bassam Helou, and Oscar Beijbom. 2020. Pointpainting: Sequential fusion for 3d object detection. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* 4604–4612.

[5] Zheng Yuan and Peng Silong. 2014. A practical roadside camera calibration method based on least squares optimization [J]. *IEEE trans. Intell. Transp. Syst.* 15, 2 (2014), 831–843.