Multimodal Early Raw Data Fusion for Environment Sensing in Automotive Applications

M. E. Pederiva[1], J. M. De Martino[1] and A. Zimmer[2]

[1]School of Electric and Computer Engineering, UNICAMP, Brazil
[2]CARISSMA - Center of Automotive Research on Integrated Safety Systems and Measurement Area, Technische Hochschule Ingolstadt, Germany

Abstract

Autonomous Vehicles became every day closer to becoming a reality in ground transportation. Computational advancement has enabled powerful methods to process large amounts of data required to drive on streets safely. The fusion of multiple sensors presented in the vehicle allows building accurate world models to improve autonomous vehicles' navigation. Among the current techniques, the fusion of LIDAR, RADAR, and Camera data by Neural Networks has shown significant improvement in object detection and geometry and dynamic behavior estimation. Main methods propose using parallel networks to fuse the sensors' measurement, increasing complexity and demand for computational resources. The fusion of the data using a single neural network is still an open question and the project's main focus. The aim is to develop a single neural network architecture to fuse the three types of sensors and evaluate and compare the resulting approach with multi-neural network proposals.

CCS Concepts

• ***Computing methodologies*** → ***Object identification; Object detection;*** • ***Applied computing*** → ***Transportation;***

1. Introduction

Autonomous vehicles have been the target of great interest in universities, research centers, and industry. With the advance of computer technology and computational techniques, autonomous cars' implementation became increasingly viable. However, implementing autonomous vehicles on urban streets requires a thorough perception of the environment, including detecting objects and their movements. These tasks require exteroceptive sensors to measure the car's surroundings. Sensors in this category include Cameras, Radio Detection and Ranging sensors (RADAR), and Light Detection and Ranging sensors (LIDAR).

Current works combine the information from two of these three sensors. Methods based on LIDAR-Camera have shown convincing results concerning visual detection, distance, and geometry estimation of objects [PMR20]. However, the methods are not adequate to estimate objects' velocities [YZK21]. Velocity estimation is better tackled by LIDAR-RADAR fusion methods [SHL*20]. However, the absence of cameras in LIDAR-RADAR approaches precludes objects' visual identification, impacting autonomous decision-making. Finally, the RADAR-Camera detection shows gains in performance for detecting objects in low light and rainy/cloudy weather [KRBG20]. Although RADARs are reliable all-weather sensors, they can not provide a dense environment sampling as LIDARs.

The fusion of three sensors can cover the limitations of each sensor and provide a high-performance estimation of object characteristics. However, this merger has not yet been properly explored.

The principal approaches present a Late Fusion of the sensors. Based on parallel networks, each sensor passes through a different neural network, and in the end, the results are combined. This method provides a robust result. However, it requires complexity and demand for computational resources. On the other hand, the Early Fusion fuses the sensors' information before the network, implementing a single architecture for the prediction. This process produces a low computational demand, a relevant characteristic for autonomous vehicles' application. Therefore, our project proposes a new fusion approach based on a single Neural Network. This network will combine and analyze information acquired by LIDAR, RADAR, and Camera sensors to detect vehicles and pedestrians in the street. Moreover, it will be based on the latest detection techniques for fast predictions and onboard implementations.

2. Methodology

We unify the Camera image (RGB) and the LIDAR data (XYZ), in a single input tensor with $Width \times Height \times 6$ (channels RGB-XYZ). The input tensor is divided into an Sx×Sy×Sz grid. Our model uses $Sx = 10$, $Sz = 10$, and $Sy = 4$.

In the grid, the object is presented in one of these grid cells. The grid cell that has the center of this object will be responsible

DOI: 10.2312/egp.20221006

for detecting it (Figure 1). Each grid cell stores an 8 length array prediction: $class, x, y, z, w, h, l, P$. Where the first value represents the object class, the following six values represent the 3D bounding box (center of the object and its dimensions), and the last is the confidence of its estimation.

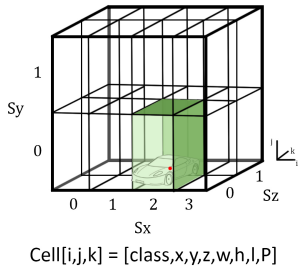

Figure 1: *Illustration of the model's approaches with Sx=4, Sy=2, and Sz=2.*

2.1. Network

Our proposed network is composed of 2 types of convolutional layers, 2D and 3D. First, an EfficientNet Backbone with a sequence of 2D convolutional layers receives the input tensor (W×H×6). Then, the last backbone's layer's output passes by a reshape, introducing one dimension to the architecture. The result of this process passes by two 3D convolutional layers with $kernel\ size = 1$ and $stride = 1$. After Flatting, the output pass by a Fully Connected (FC) layer. Finally, the output of the FC layer is reshaped into an Sx×Sy×Sz×8 tensor of predictions.

2.2. Loss

As presented in Equation 1, the loss considers the error between center position, dimension, confidence, and class of each cell prediction. The λ values represents variables to implement weights in each loss calculus. The 1_i^{obj} results to 1 if the cell is responsible for detecting an object, otherwise 1_i^{obj} results 0. The λ values was defined as: $\lambda_x = 20, \lambda_y = 5, \lambda_z = 20, \lambda_{dim} = 5, \lambda_{obj} = 5, \lambda_{noobj} = 1$.

$$
\begin{aligned}
&\lambda_x \sum_{i=0}^{S_X} 1_i^{obj}(x_i - \hat{x}_i)^2 + \lambda_y \sum_{i=0}^{S_Y} 1_i^{obj}(y_i - \hat{y}_i)^2 + \lambda_z \sum_{i=0}^{S_Z} 1_i^{obj}(z_i - \hat{z}_i)^2 + \\
&\lambda_{dim} \sum_{i=0}^{S_X * S_Y * S_Z} 1_i^{obj}[(\sqrt{w_i} - \sqrt{\hat{w}_i})^2 + (\sqrt{h_i} - \sqrt{\hat{h}_i})^2 + (\sqrt{l_i} - \sqrt{\hat{l}_i})^2] + \\
&\lambda_{obj} \sum_{i=0}^{S_X * S_Y * S_Z} 1_i^{obj}(C_i - \hat{C}_i)^2 + \lambda_{noobj} \sum_{i=0}^{S_X * S_Y * S_Z} 1_i^{noobj}(C_i - \hat{C}_i)^2 + \\
&\sum_{i=0}^{S_X * S_Y * S_Z} 1_i^{obj} \sum_{c \in classes} (class_i - \hat{class}_i)^2
\end{aligned}
\tag{1}
$$

3. Initial Results and Analysis

In Table 1, *mAp*, *mIoU*, and *MaxIoU* represent the mean Average Precision ($IoU > 0.5$), the mean Intersection over Union, and the maximum Intersection over Union, respectively. Furthermore, the *Obj. Recognition* shows the percentage of vehicles detected by the model which present a minimum IoU.

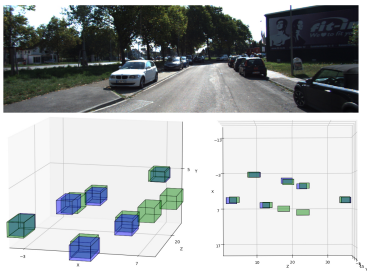

Figure 2: *Example of model detection. The top image is the camera input. The lower images are the comparison between the prediction (blue) and the ground truth (green).*

Table 1: *Model's Performance*

	mAP	mIoU	MaxIoU	Obj. Recognition
3D	12.46%	0.25	0.71	68.35%
2D	17.75%	0.28	0.84	69.12%

Besides the mAP performance, the model correctly identifies the mean position of fully and partially visible vehicles, recognizing around 68% of the vehicles on the street. Figure 2 shows an example of the model's detection, detecting most of the cars on the street. The model approach presents a potential candidate to achieve high performance in 3D object detection tasks.

4. Conclusions and Remaining Work

This study presents a different approach to fuse sensors' data. This paper shows the use of a single network to predict the 3D bounding box of objects for autonomous vehicles. Besides the model's actual performance, it showed itself a candidate detector with high potential to achieve high performance.

For the remaining work, we aim to test new architectures and optimize the system's variables to extract all the model's potential. Next, we aim to add RADAR data in the input tensor and estimate the velocity of the objects. Finally, the project seeks to build and publish a new dataset to contribute to research in the Self-Driving cars field.

References

[KRBG20] KOWOL K., ROTTMANN M., BRACKE S., GOTTSCHALK H.: YOdar: Uncertainty-based Sensor Fusion for Vehicle Detection with Camera and Radar Sensors. http://arxiv.org/abs/2010.03320, 2020. arXiv:2010.03320. 1

[PMR20] PANG S., MORRIS D., RADHA H.: Clocs: Camera-lidar object candidates fusion for 3d object detection. *CoRR abs/2009.00784* (2020). URL: https://arxiv.org/abs/2009.00784, arXiv:2009.00784. 1

[SHL*20] SHAH M., HUANG Z., LADDHA A., LANGFORD M., BARBER B., ZHANG S., VALLESPI-GONZALEZ C., URTASUN R.: LiRaNet: End-to-End Trajectory Prediction using Spatio-Temporal Radar Fusion. https://arxiv.org/abs/2010.00731, 2020. URL: http://arxiv.org/abs/2010.00731, arXiv:2010.00731. 1

[YZK21] YIN T., ZHOU X., KRÄHENBÜHL P.: Center-based 3d object detection and tracking. https://arxiv.org/abs/2006.11275, 2021. arXiv:2006.11275. 1