A LiDAR - camera fusion detection method based on weight allocation

Haotian Kang[1], *Tianshu* Wang[1*]
[1]The Graduate School, Northwestern University, Evanston, IL 60208, United States

Abstract. In automatic driving target detection problem, the neural network is applied to two methods, vision, and LiDAR. These two methods have some relatively mature models based on neural networks. Combining the two to complement each other has become a hot topic. At present, most autonomous driving sensor fusion methods focus on fusion strategy and feature alignment, and there are few studies on the weight ratio of the two sensors after fusion in different environments. In this paper, a fusion target detection model of camera and LiDAR is proposed based on the weighted weight allocation method. The weighted fusion method is adopted, image feature points are extracted by Fast RCNN, and then LiDAR point cloud data is fused into the model by the weighted method, environment variables are introduced, and different weight allocation methods are output under different environments through full connection layer preprocessing. The results on the Nuscenes dataset show that compared with the results without weight assignment, the model can effectively achieve targeted weight assignment in different situations, and the performance is due to the single-sensor method.

1 Introduction

Neural network technology plays an important role in every component of the automatic driving system. There are some applications of neural networks in automatic driving environment perception, decision planning, and object behavior prediction. At present, the perception of automatic driving mainly has two methods according to the different sensors, the camera method and the LiDAR method. The camera method uses the camera as the only sensor and conducts target detection according to the RGB image data collected by the camera. The LiDAR (Light Laser Detection and Ranging) method takes LiDAR as the only sensor and uses point cloud data for target detection. There have been some studies on both methods before. For example, Alex et al. established the PointPillars model for feature extraction of LiDAR point cloud data to achieve effective target detection with point cloud data [1]. Carion et al. build a DETR model based on Transformer to complete feature extraction of camera RGB image data, convert 2D RGB images into 3D images, and realize 3D target detection of vision methods [2]. For the dual-sensor fusion strategy, there have been certain studies before. For example, Jason et al. established the AVOD model to extract

* Corresponding author:qinlei@Idy.edu.rs

the features of the two types of sensors respectively and fused them to improve the accuracy of target detection [3].

Works done on a single sensor provide the basis for sensor fusion. Previous works on sensor fusion usually focused on the selection of the fusion stage and the optimization of the fusion mode, and the fusion weight was usually 50% each, and relatively little work was done on the specific weight allocation. In practical applications, under the influence of different external environments, the performance of LiDAR and the camera has its advantages and disadvantages, and it is not appropriate to adopt the same weight at any time.

The purpose of this paper is to propose a direction for improving the performance and robustness of the sensor fusion model, that is, the adaptive allocation of different sensor weights after sensor fusion. Aiming at the weight distribution problem after sensor fusion, a pre-processing multi-sensor data weight model is established based on Fast RCNN technology.

2 Sensor data

LiDAR data generally includes a table head, variable length recording block (VLR), point data recording block, and extension. The header is the beginning of the LiDAR data file, which records the sensor type, acquisition time, acquisition location, data format, and data version. Variable length record blocks are usually immediately following the table header, and these blocks are used to store additional metadata or user-defined extended information. The length of the VLR can vary, so it can contain different types of additional information, such as projection information, additional sensor configuration data, and so on. The point data recording section is the core part of the LiDAR data file, storing the specific data of each collection point. Point data usually includes the three-dimensional coordinates of each point, as well as reflection intensity, classification information, and so on. An extension section is a subsequent section of a file that is usually used to contain additional or user-defined extension data.

In these modules, the most important data is generally stored in the point data recording section, and the LiDAR point cloud data after general processing includes x,y,z coordinate information, target object reflectivity information, time stamp, reflected light intensity, and other data. Compared with vision information, LiDAR point cloud information is sparse and 3D data.

The principle of the automatic driving vision method is to collect image data through the camera, and the processor can detect and recognize the target and predict the movement behavior of the target through the algorithm. The vision method data obtained from the video captured by the camera is usually the RGB image taken out of the video in a single frame. RGB image is a combination of red, green, and blue channels to achieve different color display of data, which is usually converted into a series of pixel values containing three channels, each pixel value 0 to 255, representing 256 colors under the channel.

The current two sensor methods have their own advantages and disadvantages. In terms of data information, LiDAR point cloud data has distance information compared with image data, so the 3D position of the detected target can be obtained. However, image data does not have distance information, so it lacks detailed coordinate position, which may lead to wrong segmentation in semantic segmentation. Besides, the resolution of RGB image data is better in medium and long-distance target detection, and its color information helps better feature extraction. When there is an occlusion between the target and the sensor, the image can also better restore the occlusion relationship. In terms of data noise, LiDAR point cloud data currently has the problem of too much noise, and the point cloud data is sparse, which is more obvious in long-distance targets, which makes it difficult to use point cloud data to extract target features, while RGB images basically do not have this problem. In terms of

environmental robustness, LiDAR point cloud data quality is less affected by environmental factors, while RGB image data is more affected by low-visibility environments such as heavy rain and fog, and the effect is not good in harsh environments.

3 Previous works

There has been some previous work on sensor fusion issues. The AVOD model uses the RPN network to extract the features of 3D point cloud data and RGB data respectively to obtain rough 3D contours and then uses the fully connected layer for fusion to finally obtain the target detection results [3]. Different from previous one-to-one fusion, the LIF-Seg model first uses LiDAR point cloud data to project to camera images, then uses the splicing data to obtain the coarse features of LiDAR point cloud data, and then performs fusion [4]. The RoIFusion model simplifies data fusion issues from previous work. Previous work has typically fused all point data in LiDAR point cloud data with RGB images, which is computation-based and slow and involves unimportant background point data. RoIFusion uses a layer of key point extraction to generate RoI (Region of Interest), which is pooled after blending 3D and 2D RoI to obtain the fused features and finally completes 3D target detection and classification based on the fused features [5]. The sem-Aug model adopts an occlusion sensing mask generation method to solve the problem that it is difficult to detect the occluded line in LiDAR point cloud data. At the same time, the model uses a simple automatic labeling strategy to generate a dense vehicle segmentation mask without image-based data labeling and then learns the semantic features of the vehicle from the auto-labeled segmentation mask and fuses it with the original LiDAR point cloud [6]. Different from the other fusion models, CL3D extracts the features of RGB data to generate the fusion of pseudo-LiDAR point cloud and original point cloud data to reduce the impact of sparse LiDAR point cloud data, and then extracts the features of RGB image data and enhanced LiDAR point cloud data respectively, and performs the fusion [7]. The biggest highlight of the DTFI model is that it can be operated at a higher operating frequency of 30hz to be suitable for highway scenarios [8]. FUTR3D model adopts the Transformer model, which combines the features of a multi-eye camera, high-resolution LiDAR, low-resolution LiDAR, and millimeter-wave radar at the same time [9]. The polymerization is carried out in a medium stage and is realized through a multi-layer Transformer structure.

Most of these works focus on fusion mode, and pay little attention to the weight distribution of two types of sensor features in feature fusion. However, in different scenarios, LiDAR and camera have their advantages. For example, under good lighting conditions, camera data can better distinguish objects and judge signal lights through complete color information, while LiDAR can still output point cloud data under low visibility conditions such as heavy rain or fog. The same and equal weight distribution in different scenarios may affect the target detection effect after fusion. There are a few research on the sensor weight assignment. AdaFusion is a model for robot scene recognition, which is divided into feature extraction branches and weight generation branches [10]. The weight generation branch consists of three parts: multi-scale attention, intra-modal fusion and inter-modal fusion. First, attention is learned from different layers of the network, and then attention is fused in two stages to output the weight value.

4 Model

The model used the NuScenes dataset, a publicly available multimodal data set for autonomous driving developed by Aptiv and nuTonomy. The data is collected by cars with different sensors on the road, the main sensor types are cameras, LiDAR, millimeter wave

radar and IMU (Inertial Measurement Unit). Nuscenes' scenes consist of 1,000 complex urban road traffic scenes, each lasting about 20 seconds, including busy intersections, crowded pedestrian areas, and nighttime driving. In the annotation of the data set, the 3D boundary box is used to represent the position and size of the object recognized by the camera image or LiDAR, and the object category is used to show the object category represented by each boundary box. NuScenes data set has a description of the environment, but it is not uniform, that is, there is no systematic summary of the weather conditions of the photographed environment, such as temperature, visibility, ambient light intensity and other information, and these environmental information will actually affect the recognition quality of the camera image. Therefore, this paper mainly adopts the idea of antagonism, and increases the blurring degree to some images manually to optimize and update the model simply.

In the image processing of NuScenes, Fast RCNN is selected for processing. The model extracts feature maps by convolutional neural network, then generates candidate regions by selective search, pools the candidate regions, and then sends the processed feature maps to FPN and the full connection layer for classification, achieving the effect of target detection.

The result of Fast RCNN prediction for Nuscenes' images is shown in Fig. 1. The number on the left of each Label in the Fig. 1 represents the type of prediction target, which is determined by the pre-processed data set COCO of Fast RCNN, and the right is the prediction probability.

Fig. 1. Predicts Nuscenes images using Fast RCNN (Original).

Based on the same sample, if it is blurred to simulate heavy rain or low visibility in the driving environment, it can observe whether the simple camera-based method can still have such a good effect. As shown in Fig. 2, when simulating a rainy environment, the number of targets collected and the corresponding accuracy rate are greatly reduced.

Fig. 2. After image blurring, Fast RCNN is used to predict Nuscenes images. It can be seen that only two targets are successfully predicted, and the probability of correct prediction is not 90% (Original).

At the same time, when the rainfall changes, it can be found from the confidence score of the model shown in Fig. 3 that the higher the rainfall, the lower the confidence score of the model, and the decreasing trend is obvious. The horizontal axis is rainfall and the vertical axis is average confidence.

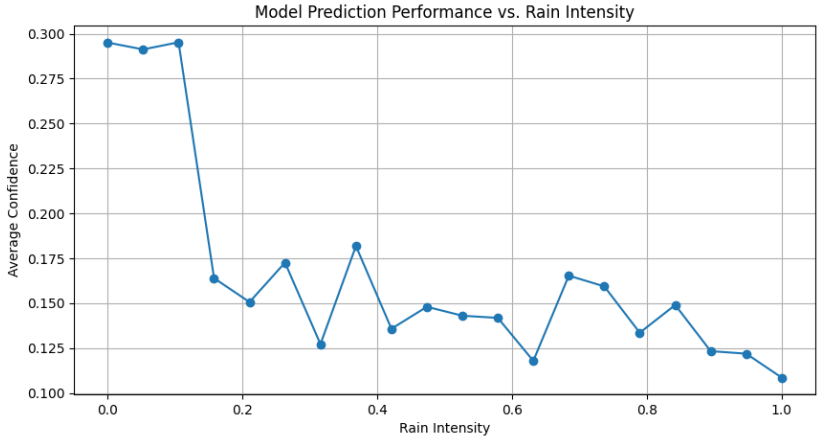

Fig. 3. Relationship of model confidence scores with rainfall (Original).

Compared with the camera, LiDAR has a better tolerance for visibility, and can also output point clouds to judge the objects in front when visibility is low. Therefore, at present, most car companies fuse LiDAR data and camera data to achieve the effect of target detection, but there is no good conclusion on how to choose the weight of the two (Fig. 4).

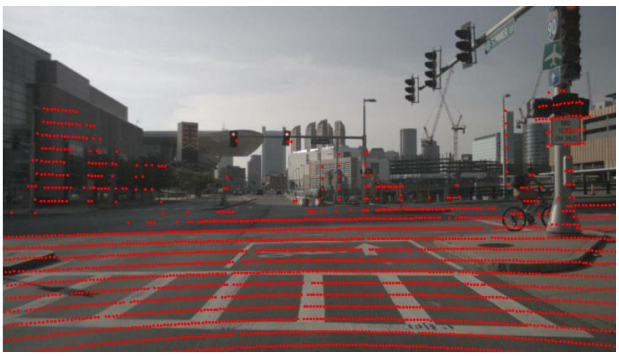

Fig. 4. LiDAR point cloud data (Original).

When there is a LiDAR point cloud, a simple weighted fusion method is carried out on it. Weighted fusion fuses data by assigning different weights to different sensors. Generally speaking, different weights represent the reliability and importance of different sensors. Here, we first chose the method of randomly assigning weights without adding neural network to fuse LIDAR point cloud data into the previous model, and the results are shown in Fig. 5. It can be seen that the model with LIDAR data has relatively more advantages in the performance of target detection, but in many cases, the simple image detection has better effect. Then dynamic selection weights are very necessary. The blue line represents the mixed model, the red line represents the vision model, the horizontal axis is rainfall, and the vertical axis is average confidence.

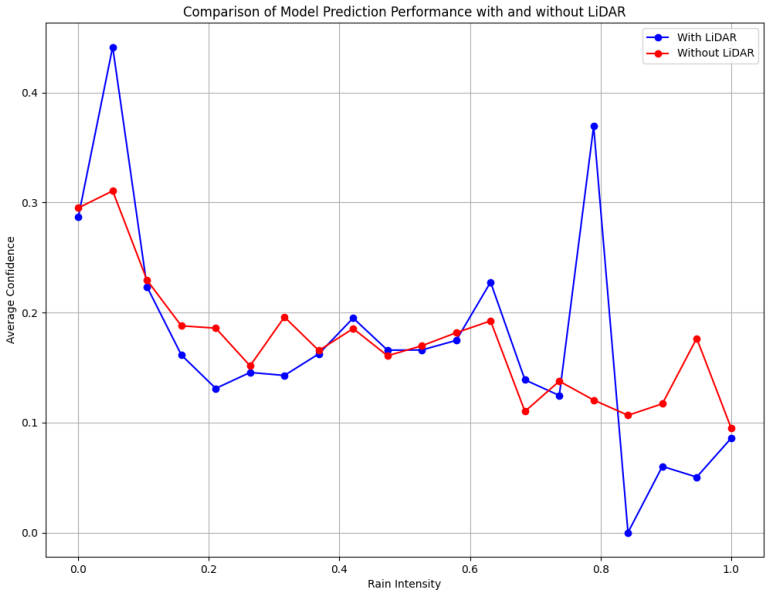

Fig. 5. Performance comparison between the vision method model and LiDAR/vision method mixed model under different rainfall (Original).

By adding a pre-processing model in front of the Fast RCNN model, which is used to automatically adjust the weights of different sensors according to the raindrop density, a relatively simple fully connected layer is selected here. By providing environmental parameters as input vectors to the neural network, non-linear features of input data are captured through multiple hidden layers. Finally, Softmax function is used to ensure that the sum of weights of all sensors is 1, and a simple dynamic distribution method is obtained, the

results of which are shown in Fig. 6. The blue line represents the fusion model, the red line represents the pure vision model, the green is the fusion model with adaptive dynamic adjustment weights, the horizontal axis is the rainfall, and the vertical axis is the average confidence.

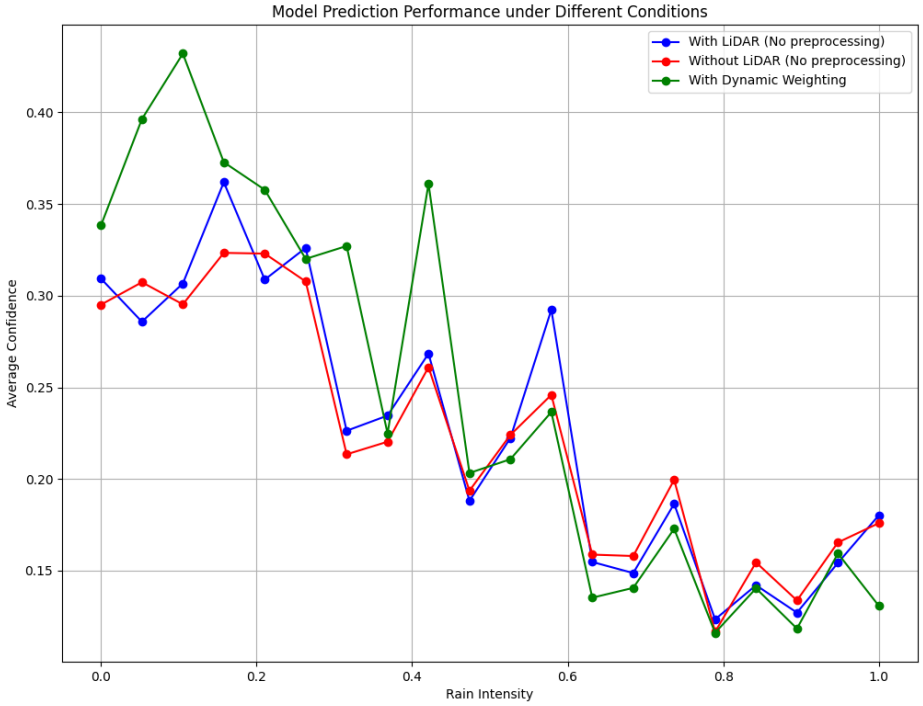

Fig. 6. Comparison of performance of pure vision method model, LiDAR - vision method fusion model and fusion weight adaptive adjustment fusion under different rainfall (Original).

It can be seen that the recognition ability of the model under different rainfall has been improved after including the preprocessing of dynamic allocation of the most reasonable weights.

5 Future work

In the previous part, a simple model is established to realize the multi-sensor fusion and weight allocation under the introduction of weather variables. The results show that different fusion weights in different environments can improve the robustness of fusion 3D object detection. Compared with the previous work, the main contribution of this model is to propose a fusion sensor weight allocation strategy for automotive autonomous driving scenarios. However, the model still has the following limitations:

1. The fusion method is relatively simple. We use a simple weighted fusion to fuse LiDAR and camera. Compared with recent work, this fusion method is simple in structure and easy to implement, but the accuracy of feature alignment is relatively low.

2. Insufficient environmental variables cannot reflect the common scenarios in the process of automatic driving. We introduced rainfall as an environmental variable, but these variables cannot fully reflect the typical scenario in reality. The environment also contains other variables that can affect the data acquisition effect of the sensor, such as the ambient light intensity.

3. The pre-processing model of weight allocation is too simple, which simply adopts a full-connection layer model to read the target features, which may not be ideal in the result of weight allocation.

4. Small scale of dataset. This paper only uses the mini version of Nuscenes for simple processing, and the amount of data is not large, which may affect the final effect.

In response to the above problems, further research can be carried out in the following directions:

1. Adopt a more effective fusion method and import more data as a training set to obtain a model with higher accuracy;

2. for complex urban or suburban scenes, introduce more environmental variables for learning. The environmental variables that can be considered are: building density, lane width, traffic flow, number of pedestrians, distance between the front and the road speed limit, etc. These environmental variables help us to construct a more differentiated vehicle driving scene, and adopt different weight distribution ratios for different scenes. It should be noted that these environmental variables may require more information transmission between sensors and urban transportation systems, and this work can be further extended to the target detection work of multi-sensor and multi-information source data fusion in the Internet of Things environment.

3. Introduce self-attention mechanism to dynamically adjust the proportion of weight allocation. Some previous works have realized target detection of sensor data fusion based on self-attention mechanism of Transformer model. However, these works also focus on optimization of fusion mode and lack discussion on fusion weight.

6 Conclusion

Based on the development of multi-sensor target detection in automatic driving, this paper first introduces the pure vision method data structure and LiDAR data structure to provide theoretical basis for the content of this paper. Then, the previous fusion models of LiDAR/vision methods are reviewed, and the overall architecture and fusion strategies of the models are summarized. Some of the fusion strategies are used in the subsequent simple models established in this paper. This paper summarizes the advantages and disadvantages of previous fusion models, and finds that there is a lack of research on sensor fusion weights in previous models. Therefore, the direction of adjusting sensor proportion weights to improve the robustness of models against environmental impacts is proposed, and the possibility of weight allocation under sensor fusion target detection is explored through a simple model. When precipitation is introduced as environmental disturbance, the model performs better in some scenarios than the simple fusion with no weight. Based on this model, future work can further explore the weight allocation strategy and environmental disturbance factors.

Authors Contribution

All the authors contributed equally and their names were listed in alphabetical order.

References

1. A. H. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang, O. Beijbom, PointPillars: Fast encoders for object detection from point clouds. 2019 IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 12689-12697 (2019).

2. N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, S. Zagoruyko, End-to-End Object Detection with Transformers. Lect. Notes Comput. Sci., **12346**, 213-229 (2020).
3. J. Ku, M. Mozifian, J. Lee, A. Harakeh, S. L. Waslander, Joint 3D Proposal Generation and Object Detection from View Aggregation. 2018 IEEE/RSJ Int. Conf. Intell. Robots Syst., 1-8 (2018).
4. L. Zhao, H. Zhou, X. Zhu, X. Song, H. Li, W. Tao, LIF-Seg: LiDAR and Camera Image Fusion for 3D LiDAR Semantic Segmentation. IEEE Trans. Multimedia, **26**, 1158-1168 (2024).
5. C. Chen, L. Z. Fragonara, A. Tsourdos, RoIFusion: 3D Object Detection from LiDAR and Vision. IEEE Access, **9**, 51710-51721 (2021).
6. L. Zhao, M. Wang, Y. Yue, Sem-Aug: Improving Camera-LiDAR Feature Fusion with Semantic Augmentation for 3D Vehicle Detection. IEEE Robot. Autom. Lett., **7(**4), 9358-9365 (2022).
7. C. Lin, D. Tian, X. Duan, et al., CL3D: Camera-LiDAR 3D Object Detection with Point Feature Enhancement and Point-Guided Fusion. IEEE Trans. Intell. Transp. Syst., **23**(10), 18040-18050 (2022).
8. C. Nie, Z. Ju, Z. Sun, et al., 3D Object Detection and Tracking Based on LiDAR-Camera Fusion and IMM-UKF Algorithm Towards Highway Driving. IEEE Trans. Emerg. Topics Comput. Intell., **7**(4), 1242-1252 (2023).
9. X. Chen, T. Zhang, Y. Wang, et al., FUTR3D: A Unified Sensor Fusion Framework for 3D Detection. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 172-181 (2023).
10. H. Lai, P. Yin, S. Scherer, Adafusion: Vision-LiDAR Fusion with Adaptive Weights for Place Recognition. IEEE Robot. Autom. Lett., **7**(4), 12038-12045 (2022).