

Research on Automatic Driving Object Detection based on Point Cloud and Image Data Fusion

Jian Di, Yedan Wang*

¹School of Control and Computer Engineering, North China Electric Power University, Baoding 071000, China.

*Corresponding author Email: yedanwang503@163.com

Abstract

Researchers focus on using multiple sensors to improve the accuracy of object detection model in automatic driving. Therefore, the research on data fusion method in object detection has important academic and application value. Therefore, this paper expounds from three aspects: point cloud and image data fusion object detection, the level of data fusion and the calculation method of data fusion, and comprehensively shows the cutting-edge progress in this field; Finally, the challenges, strategies and prospects of integration practice are summarized.

Keywords

Automatic Driving; Point Cloud; Image; Fusion.

1. Introduction

As one of the tasks in the field of computer vision, object detection is the basis of many other visual tasks, such as instance segmentation [1] and object tracking [2], which aims to detect each instance of different types of objects. Considering the availability of data and the richness of data features, it generally refers to object detection with RGB image as the main data, and usually uses bounding box on the image to locate objects and give the probability of object category attributes. According to the definition, the object detection algorithm is usually required to search the region that may contain the object on the image, and then classify it. This model is a multi-stage model. With the development of deep learning, one-stage detection model is proposed, which can classify while detecting, so as to improve the detection speed. In addition, the detection methods are different for different modal data. Due to the application of laser radar, radar, depth camera (RGB-Depth camera) and other sensors in automatic driving vehicles, automatic driving requires attention to point clouds, depth images or other modal data, including point clouds provided by Lidar or radar. For object detection on images, the accuracy of traditional methods is often lower than that of deep learning methods, while the latter often needs large data sets and long-time training to learn features. For the object detection on the point cloud, the advantage is that it can use the three-dimensional spatial information for detection, but the disadvantage is that the increase of spatial dimension leads to the sparse point cloud data, resulting in the poor effect of model fitting [3]. Although the existing object detection methods based on computer vision and deep learning have achieved excellent results on specific data sets, the performance of the existing models is still insufficient in the face of specific scenes, especially those with high requirements for robustness, detection speed and accuracy, such as automatic driving.

2. Object detection

2.1. Based on RGB image

In 2012, the emergence of AlexNet [4] became the beginning of the development of deep learning and computer vision. Since then, the method based on convolutional neural network

(CNN) has gradually become the mainstream of computer vision tasks. Since 2014, girshick and others have successively proposed r-cnn and fast r-cnn. The neural network developed based on r-cnn has become the main benchmark model of two-stage detection. R-cnn first generates candidate bounding boxes through selective search, then extracts features through CNN, and finally classifies them. In addition, in order to improve the computational performance, Ren et al. [5] Proposed Faster R-CNN in 2015. The region proposal network (RPN) used has become the design idea of the classic object detection model. Since then, feature pyramid networks (FPN) have fully mined image information by integrating different levels of semantics. Many models adopt this design combining shallow and deep semantics.

Compared with two-step detection, one stage detection, whose classical models are YOLO (you only look once) [6] and SSD (single shot detector) [7]. It can be seen from the name that one-step detection is different from the two-step detection mode of proposing candidate bounding box for classification, but predicting the bounding box and classification on a part of the image at the same time. Redmon and others have successively put forward three versions of Yolo models: Yolo, yolo9000 and yorov3. When predicting the boundary box, yolo9000 and yov3 successively use the idea of anchor to replace the full connection layer, adopt multi-scale model and carry out logical regression on the boundary box. Finally, they improve the accuracy and maintain a high detection speed, And the detection performance of yolov3 for small objects is significantly improved. In addition to the Yolo series network, SSD proposed in 2016 and retina net in 2017 have also been widely studied and used. Since then, there are still various models that have made breakthroughs in the benchmark data set. However, these models still follow the classical design of one-step or two-step models.

2.2. Lidar based point cloud

Lidar has been widely used in automatic driving sensors because of its active sensing characteristics and providing more dense point clouds than radar [8-10]. Lidar can not only provide a large sensing range with a scanning radius of 50 ~ 70 m or even further, but also provide the depth information and reflectivity of the environment without being affected by the ambient light. Lidar data is in the format of three-dimensional point cloud, so the coordinates of point cloud naturally provide the three-dimensional spatial coordinates of the object relative to Lidar, which can be used for object detection in three-dimensional space; The reflection intensity value of the point cloud reflects the material of the object surface, so different objects can be reflected according to the reflection.

There are many methods for object detection on point cloud, which are related to the data representation of point cloud. Due to the spatial characteristics of point cloud, it can not only perform three-dimensional detection in three-dimensional space, but also project to two-dimensional plane, and calculate by using the two-dimensional object detection model based on image. Specifically, the projection mainly includes front view projection and bird's eye view (BEV). On the one hand, the front view projection can obtain a dense projection image, on the other hand, it can be projected on the pixel plane of the vehicle camera, so it can be integrated with the camera image to perform object detection. However, the Bev view of point cloud in the model has been more used for three reasons:

- 1) when the object is projected to the Bev view, the physical size will be retained, while other projected views will produce perspective effect;
- 2) The objects in Bev view rarely have occlusion problem, and are distributed discretely in space;
- 3) In the road scene, because the object is usually located on the ground and the change of vertical position is very small, the aerial view is conducive to obtain an accurate 3D bounding box.

In addition, Lidar can also provide rich semantic information on the projected image, such as depth information, reflectivity information and object height information. According to the multiple semantic information, visual computing can be carried out more effectively on the point cloud.

2.3. Object detection based on image and point cloud fusion

In recent years, the first mock exam of object detection technology based on multimodal data fusion is based on the above considerations of single modal data, and on the other hand, the consideration of vehicle sensor stability. Table 1 summarizes the depth object detection methods based on multimodal data fusion in automatic driving scenes in recent years.

Table 1. Statistics of deep object detection data fusion methods

literature	sensor	Point cloud representation	Data fusion	Feature fusion	Result fusion
[11]	Lidar, RGB camera	Front view, bird view, voxelization	√	√	√
[12]	Lidar, RGB camera	3D point cloud, voxelization	√	√	
[13]	Lidar, RGB camera	bird view		√	
[14]	Lidar, RGB camera	Front view	√	√	
[15]	Lidar, RGB camera	Voxelized front view and bird view			√
[16]	Lidar, RGB camera	bird view		√	
[17]	Lidar, RGB camera	bird view		√	
[18]	Lidar, RGB camera	Front view		√	
[19]	Lidar, RGB camera	Generate 3 front views by depth		√	√

3. Data fusion arrangement

Since the representation of LIDAR point cloud in the fusion model affects the selection of fusion methods, the corresponding information is also listed in Table 3. According to Feng et al. [20], according to the data fusion steps that appear in different stages of the object detection model, the fusion methods can be divided into three types: early fusion, middle fusion or deep fusion and late fusion. Among them, the pre fusion aims at the fusion of original data or only preprocessed data. The post fusion fuses the operation results of multiple branches of the model to obtain the final result. The intermediate fusion combines the characteristics of pre fusion and post fusion to fuse the data of multiple modes or their corresponding feature maps, and then continue the operation. Intermediate fusion can exist only in the intermediate steps of the model or throughout the whole model. In some papers, the three kinds of fusion are also called data level fusion, feature level fusion and decision level fusion. However, in the deep learning model of multi-level and multi model combination, the result of model branching is also the intermediate feature of the whole model. Most fusion methods are classified as intermediate fusion, resulting in the lack of discrimination of stage based division methods. In order to distinguish the degree of model fusion more intuitively, this paper further distinguishes the fusion methods from the perspective of fusion level, which are data fusion, feature fusion, result fusion and auxiliary estimation. According to the division method proposed in this paper, the division of fusion method is no longer based on the sequential position of fusion steps in the model, but the role of fusion results in the model, and the division of fusion will not be limited by the form of fusion data, which more intuitively reflects the role of fusion operation on the model. Definition of different fusion levels.

3.1. Data fusion

Data fusion refers to the fusion operation in which the output data will be used as the input of subsequent models. On the one hand, the model can be a complete model or a sub model as a branch or part of the integration model; On the other hand, the input data, that is, the fused data, can be either the original data from the sensor or only the preprocessed data, or the feature

vector, feature map or calculation result output by the sub model of the integrated model. Document [21] fused the RGB image and depth image provided by the depth camera after processing respectively. The document [22] provides multiple possible fusion positions. When the model is connected through vectors before extracting features, the fusion results are used as the input of the subsequent complete model, which belongs to data fusion, that is, the data level fusion mentioned in document [23].

3.2. Feature fusion

Feature fusion refers to the fusion operation that the output data will be used as the feature vector or feature map of the model (hereinafter referred to as feature map). Specifically, feature fusion usually refers to the fusion operation between the feature maps input by the model in the process of model calculation, and the fusion result continues to participate in the calculation as the feature map of the current model. However, in some models, the original data may also be feature fused with the feature map after preprocessing. The same method which point clouds and images are processed by separate SSD branches, and their feature images are cross fused appears in [24] on lane line detection.

3.3. Result fusion

Result fusion refers to the fusion operation in which the output data is the final object output of the model, or the input data itself is the object output and the fusion result is used for result correction. By definition, result fusion mainly includes the fusion operation of calculation results of parallel model branches and the optimization calculation for the same object output of subsequent sub models on the existing results in serial model. In reference [25], the model fuses the extracted region proposal based on point cloud and two branches of image to obtain all candidate detection frames. Since the candidate frames of object detection model are obtained by Region recommendation regression, it can be regarded as the result fusion of candidate frames. Similar fusion is also reflected in literature [26].

4. Calculation method of data fusion

4.1. Connection method

As one of the most direct fusion methods, the connection method directly connects the data of different modes or their feature maps, and the increase of features is directly reflected in the increase of dimension of data or feature maps [27]. The first mock exam (data) is spliced to the first modal feature (data), and the number of layers is constant, but the size of each layer is larger. This method is also called concatenate; the 1 way is to connect the two mode feature data. the first mock exam is used to extend the second modal feature map (data). The fused feature map is more varied and the size of each channel is unchanged. The former fusion method is more common in the feature fusion stage, and the second fusion method is more common in the data fusion stage. However, in order to use the second fusion method for feature fusion, 1 can be used after fusion 1×1 convolution to adjust the number of channels. The second method can be understood as extended channel. The PointFusion model[28] uses concatenate to connect the feature images extracted from point clouds and RGB images respectively.

4.2. Merger method

The merging method is applicable to multiple similar candidate elements, such as the fusion of similar object outputs of multiple branches and the fusion of candidate boxes.

Literature [29] adopts the form of weighted summation for the boundary box and classification results output by the model branch, and literature [30] adopts the camera with long and short focal length, so the image with two visual fields, in which the image of the long focal length camera is a subset of the image of the short focal length camera, and also has higher resolution.

Therefore, the model superimposes the boundary frame in the long focus camera image on the corresponding position of the short focus camera image, so as to improve the detection performance of the model for distant small objects while maintaining the field of view.

4.3. Additive method

The addition method is usually applied to the fusion of feature maps and the processing of region of interest (ROI). The latter includes the merging of ROI and the superposition of ROI on the feature map as the constraint condition of detection. Some models consider fusing features through weighted summation of feature graphs. ROI regions can be directly merged as ordinary feature graphs or multi-level constraints of models. As proposed in reference [31], the method continuously reduces the search range according to the point cloud filtering, projects the point cloud onto the image, and detects and classifies within the point cloud range.

5. Challenge and strategy

Despite the emergence of automatic driving object detection model based on data fusion, there is still a lack of systematic description and in-depth exploration. On the one hand, there is a lack of theoretical research on data fusion methods in object detection; On the other hand, the current fusion method is still too simple. From the rationality analysis of fusion, even if only the existing methods are considered, there is still a lot of work to be completed. Not only there is a lack of comparative experiments between different methods, but also the interpretability of fusion model is an important problem. This paper suggests that the interpretability of fusion should be considered from the perspective of model robustness and data redundancy.

According to the existing paper methods, the main challenges in fusion include but are not limited to: interpretable fusion calculation, estimation of the optimal fusion stage, anomaly detection of fusion data, adaptive fusion mechanism, information gain estimation of cross modal data fusion and estimation of the change degree of model uncertainty by data fusion. At present, there have been some validation experiments and comparative experiments, but further research is needed.

Although the existing models may favor a certain method in fusion calculation, there is still no quantitative analysis or derivation to prove the superiority of a certain method. However, in the paper recorded in Table 4, the experimental results of fusion level comparison basically support the view that intermediate fusion is better than pre fusion and post fusion, and corresponding to the classification method in this paper, feature fusion is better than data fusion and result fusion.

In recent years, there have been fusion modes that go beyond "specific stage fusion" to a certain extent, such as adaptive stage fusion, CO modal data fusion and so on.

Reorganize the data form to achieve more effective fusion. Further, from the perspective of bionics, cognitive information processing theory understands the human brain activity as an information processing system. External stimuli lead to a variety of receptors to produce instantaneous memory, which is then transformed into working memory, and combined with long-term memory to understand and judge the external information. Aiming at the basis of vehicle environment perception system based on multi-source sensors, this paper proposes a multi-modal cognitive fusion framework for automatic driving: improve the accuracy and robustness of object detection and semantic segmentation in road scene by means of specific processing and multi-stage and multi-level fusion; Improve the accuracy of object tracking, and then enhance the accuracy and reliability of vehicle environment perception system, so as to provide strong theoretical basis and technical support for autonomous vehicles to move towards reality.

Figure 1 shows a multimodal cognitive fusion framework for automatic driving. Considering the inconsistency of feature space of multimodal data, the objected feature extraction is carried out for multimodal data, and then the multi-channel features are fused in the multimodal integration module; The fusion process will combine the prior knowledge from continuous frames, which can enhance the cognitive ability of the whole system to the environment, and greatly improve the performance of algorithms such as object detection, semantic segmentation and object tracking in abnormal exposure, rain and fog weather, etc.

6. Conclusion

Data fusion is an important trend in the perception task of autopilot. Therefore, this paper provides an overview of multimodal data fusion methods in object detection in autopilot scene. This paper introduces the background of object detection and data fusion, and summarizes the current research from two aspects: fusion level and fusion calculation method. We believe that the traditional "front, middle and back" division method does not distinguish the fusion methods in the integration model. Therefore, this paper adopts a new hierarchical definition method and gives a clear definition, which can help researchers better understand the motivation and function of fusion method design. In addition, this paper puts forward the rationality analysis of data fusion, summarizes the existing research, discusses the challenges and Strategies of current fusion methods, and puts forward some open problems.

Although new fusion methods have been proposed, there is still a lack of theoretical analysis and in-depth comparative experiments. The development of automatic driving technology depends on efficient and robust environment perception. The corresponding data fusion methods should be developed as soon as possible to ensure the performance of vehicle perception technology. Our follow-up work includes the research on data representation and the information gain measurement of cross modal data fusion. At the same time, we expect that the new research work can solve the challenges in the fusion methods discussed in this paper.

References

- [1] Hariharan B, Arbeláez P, Girshick R, et al. Simultaneous detection and segmentation [C]// European conference on computer vision. Springer, Cham, 2014: 297-312.
- [2] Kang K, Li H, Yan J, et al. T-cnn: Tubelets with convolutional neural networks for object detection from videos [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2017, 28(10): 2896-2907.
- [3] Zou Z, Shi Z, Guo Y, et al. Object detection in 20 years: A survey[J]. arXiv preprint arXiv:1905.05055, 2019.
- [4] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks [J]. Advances in neural information processing systems, 2012, 25: 1097-1105.
- [5] Ren S, He K, Girshick R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks [J]. Advances in neural information processing systems, 2015, 28: 91-99.
- [6] Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection [C]// Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 779-788.
- [7] Liu W, Anguelov D, Erhan D, et al. Ssd: Single shot multibox detector [C]// European conference on computer vision. Springer, Cham, 2016: 21-37.
- [8] Meyer G P, Laddha A, Kee E, et al. Lasernet: An efficient probabilistic 3d object detector for autonomous driving [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 12677-12686.
- [9] Qi C R, Su H, Mo K, et al. Pointnet: Deep learning on point sets for 3d classification and segmentation [C]// Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 652-660.

- [10] Qi C R, Yi L, Su H, et al. Pointnet++: Deep hierarchical feature learning on point sets in a metric space[J]. arXiv preprint arXiv:1706.02413, 2017.
- [11] Liang M, Yang B, Chen Y, et al. Multi-task multi-sensor fusion for 3d object detection [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 7345-7353.
- [12] Sindagi V A, Zhou Y, Tuzel O. Mvx-net: Multimodal voxelnet for 3d object detection[C]//2019 International Conference on Robotics and Automation (ICRA). IEEE, 2019: 7276-7282.
- [13] Bijelic M, Mannan F, Gruber T, et al. Seeing through fog without seeing fog: Deep sensor fusion in the absence of labeled training data [J]. arXiv preprint arXiv:1902.08913, 2019, 1030.
- [14] Dou J, Xue J, Fang J. SEG-VoxelNet for 3D vehicle detection from RGB and LiDAR data[C]//2019 International Conference on Robotics and Automation (ICRA). IEEE, 2019: 4362-4368.
- [15] Du X, Ang M H, Karaman S, et al. A general pipeline for 3d detection of vehicles [C]// 2018 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2018: 3194-3200.
- [16] Liang M, Yang B, Wang S, et al. Deep continuous fusion for multi-sensor 3d object detection[C]// Proceedings of the European Conference on Computer Vision (ECCV). 2018: 641-656.
- [17] Wang Z, Zhan W, Tomizuka M. Fusing bird's eye view lidar point cloud and front view camera image for 3d object detection [C]// 2018 IEEE Intelligent Vehicles Symposium (IV). IEEE, 2018: 1-6.
- [18] Kim J, Koh J, Kim Y, et al. Robust deep multi-modal learning based on gated information fusion network[C]//Asian Conference on Computer Vision. Springer, Cham, 2018: 90-106.
- [19] Du X, Ang M H, Rus D. Car detection for autonomous vehicle: LIDAR and vision fusion approach through deep learning framework [C]// 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2017: 749-754.
- [20] Feng D, Haase-Schütz C, Rosenbaum L, et al. Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges[J]. IEEE Transactions on Intelligent Transportation Systems, 2020, 22(3): 1341-1360.
- [21] Schneider L, Jasch M, Fröhlich B, et al. Multimodal neural networks: Rgb-d for semantic segmentation and object detection [C]// Scandinavian conference on image analysis. Springer, Cham, 2017: 98-109.
- [22] Luo Junhai, Yang Yang. Overview of target detection methods based on data fusion [J]. Control and decision making, 2020, 35 (1): 1-15
- [23] Zhou T, Jiang K, Xiao Z, et al. Object Detection Using Multi-Sensor Fusion Based on Deep Learning [M]// CICTP 2019. 2019: 5770-5782.
- [24] Caltagirone L, Bellone M, Svensson L, et al. LIDAR-camera fusion for road detection using fully convolutional neural networks[J]. Robotics and Autonomous Systems, 2019, 111: 125-131.
- [25] Kim T, Ghosh J. Robust detection of non-motorized road users using deep learning on optical and lidar data [C]// 2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC). IEEE, 2016: 271-276.
- [26] Liu J, Zhang S, Wang S, et al. Multispectral deep neural networks for pedestrian detection[J]. arXiv preprint arXiv:1611.02644, 2016.
- [27] Wagner J, Fischer V, Herman M, et al. Multispectral Pedestrian Detection using Deep Fusion Convolutional Neural Networks [C]// ESANN. 2016, 587: 509-514.
- [28] Xu D, Anguelov D, Jain A. Pointfusion: Deep sensor fusion for 3d bounding box estimation [C]// Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 244-253.
- [29] Guan D, Cao Y, Yang J, et al. Fusion of multispectral data through illumination-aware deep neural networks for pedestrian detection [J]. Information Fusion, 2019, 50: 148-157.
- [30] Chadwick S, Maddern W, Newman P. Distant vehicle detection using radar and vision[C]//2019 International Conference on Robotics and Automation (ICRA). IEEE, 2019: 8311-8317.
- [31] Matti D, Ekenel H K, Thiran J P. Combining LiDAR space clustering and convolutional neural networks for pedestrian detection [C]// 2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). IEEE, 2017: 1-6.