2020 16th IEEE International Conference on Automation Science and Engineering (CASE)
August 20-21, 2020, Online Zoom Meeting

Object Detection based on Fusing Monocular Camera and Lidar Data in Decision Level Using D-S Evidence Theory

Qingzhu Wu[1], Meng Zhou*[2], Biao Hu[1]

Abstract— **Visual sensors such as optical camera possess fatal shortcomings like vulnerable to the weather condition and light intensity, which may cause catastrophic disasters. Fortunately, the prevalent 3D sensors like Light Detection and Ranging (Lidar) can overcome these drawbacks. Fusing both monocular camera and 3D Lidar data can often achieve a better performance than solely using one of them. In this paper, we propose an object detection approach by fusing monocular camera data and Lidar data in decision level based on Dempster-Shafer evidence theory. First, the 3D point cloud collected from Lidar is projected onto the image plane and the upsampling method is adopted to receive dense depth maps. Then, the RGB images and dense depth maps are separately fed into YOLOv3 detection framework to extract the object features. Next, the final detection attributes, the class confidence and bounding box, are generated by using Dempster's combination rule and extracting intersection of the associated bounding boxes. Finally, the mean Average Precision (mAP) performance of the proposed method is improved by 0.5% and 3.1% compared with sigle camera method and Lidar method, respectively.**

I. INTRODUCTION

Environment awareness has been broadly studied in intelligent mobile robots and autonomous driving in recent decades, particularly object detection task is critically vital in environment awareness. RGB based visual sensors such as monocular camera have widely been implemented in recently state-of-the-art object detection or classification tasks. Among the current results, object detection methods can be simply separated into deep learning based methods and classic machine learning techniques [1]. Before the recently boosted neural network algorithms, the main methods presented in literature rely on classic machine learning methods which use handmade features. In the last few years, Convolutional Neural Network (CNN) based object detection algorithms which automatically extract object features have made extraordinary progress. Currently the most prevalent object detectors use CNN including region proposal based CNN like Faster R-CNN [2], and single shot object detectors like YOLO [3].

However, some intricate drawbacks such as vulnerable to the weather condition and illumination make it incapable to obtain the best performance, additionally in some extreme environment can even cause catastrophic disasters. 3D sensors like Lidar is unaffected by light and measures distance of objects precisely but without color and texture information [4]. Fusing multiple sensors can overcome the intricate drawbacks of one single sensor. Optical cameras and Lidar sensors nowadays have been commonly installed on mobile robots and autonomous vehicles, boosting more extraordinary work on navigation [5], object detection and classification [6], and tracking tasks [7]. Currently, more studies concentrate on fusing multiple sensors to gain more reliable and stable detection. The easily available 3D sensors such as Lidar accelerate the research of detection methods, transforming the pure vision based detection to depth inference and fusion detection [8].

There are three strategies of multi-sensor fusion according to the stage to fuse data: data fusion, feature fusion and decision fusion [8]. Data fusion is also recognized as early fusion [9], where in the input stage original data collected from multiple heterogeneous sensors is combined or connected using external parameters to obtain a composite data with abundant and complementary information. In work of road detection [4], a hybrid Conditional Random Field (CRF) model is presented to integrate the images and point clouds in a probabilistic manner. The labels of image pixels and point clouds are regarded as random variables, and three types of edges are performed to express the neighboring relationship. Qu et al. [10] present a composite system for target recognition and tracking task where three heterogeneous sensor data is combined using external rotation and translation matrixes.

Feature fusion is generally based on CNN methods, where the middle network layers share features from previous network layers to produce multi-scale object features. Schlosser el al. [7] propose a CNN based pedestrian detection approach of fusing Lidar and RGB image, by extracting Horizontal disparity, Height and Angle data (HHA) from a dense depth map, and transmit the HHA and RGB data into network in different layers, to explore the optimal data fusion way. Moreover, multi-view or view aggregation methods have also been presented in order to generate more reliable and stable object features. In [11], a multi-view representation of the point clouds is created and both RGB images and multi-view point clouds are sent to the framework to predict 3D Bounding Boxes (BBs). This method acquires multiple interactions between different network layers and multi-scale features from different views. In [12], Lidar point clouds is transformed into Bird's Eye View (BEV) and both BEV map and original image are fed into a region proposal network to produce trustworthy object proposals. Benefited from the

*This work is supported by the Beijing Leading Talents Program (Z191100006119031) and National Natural Science Foundation of China under (Grant No. 51805021 and No. 91848103).

[1]Qingzhu Wu and Biao Hu are with the College of Information Science and Technology, Beijing University of Chemical Technology, Beijing, 100029, China.

[2]Meng Zhou is with the School of Electrical and Control Engineering, North China University of Technology, Beijing, China. Meng Zhou is the Corresponding author (e-mail: zhoumeng@mail.buct.edu.cn).

multimodal features extracted from aggregation view, the architecture in [12] detects more small objects.

In decision fusion or named late fusion, some simple methods like weighted voting [13], probability models like Evidence Theory (ET) [6], as well as learnable methods like Multi-Layer Perceptron (MLP) [1] and CNN model [6] have been broadly studied. Chavez-Garcia et al. [14] propose an evidence theory based framework where mass functions of different sensors are created according to character of that sensor, and Yager's combination rule is used to increase the reliability. In [1], three modalities are extracted from camera, Lidar range data and Lidar reflectance data, and MLP is employed to find the mapping between object attributes detected from three modalities data and ground truth.

According to [8] and [7] decision fusion methods possess better performance. Early fusion is easily affected by the sensor restrictions such as the conflict information from different sensors. Meanwhile, in decision fusion techniques the simple methods are not reliable enough while the learnable method are time consuming, therefore in this work the evidence theory is utilized. Furthermore, the KITTI benchmark suit [15] contains plenty of RGB and Lidar data which makes it much convenient to carry out object detection and sensor fusion related work, thus in this work we would utilize KITTI public dataset for training and testing.

In this paper, our main contributions are as follows:

- D-S evidence theory is employed on the class confidence scores in fusion procedure for decreasing impact of conflict or incomplete information.
- In data association process, Intersection over Union (IoU) is used as distance metric to represent the similarity between bounding boxes.
- The proposed detection technique is faster than most of state-of-the-art methods listed on the KITTI benchmark. The excellent speed makes it possible to plant the algorithm into practical mobile robots and autonomous vehicles.

The rest of this paper is organized as follows. In section II, techniques used in this paper are introduced first. They involve the generation of dense depth map, and decision fusion using D-S evidence theory. Then the main structure of the proposed system is given. Section III presents experiment details and results. Finally, section IV concludes our work and presents some feasible way to improve our work.

II. Methodology

In this section the fusion system architecture and techniques used in this work are presented in details. The system architecture is comprised of three parts: data preprocess, basic object detection based on YOLOv3 framework, and detection fusion using D-S evidence theory. In data preprocess, the upsampling method and the generation of dense depth maps are clarified. Then the RGB images and generated dense depth maps are fed into YOLOv3 detection framework [3]. The class confidence extracted from YOLOv3 output is fused by applying Dempster's combination rule. In the meanwhile, the predicted bounding box is obtained by extracting the intersection area of the associated BBs. Here is the detailed description below.

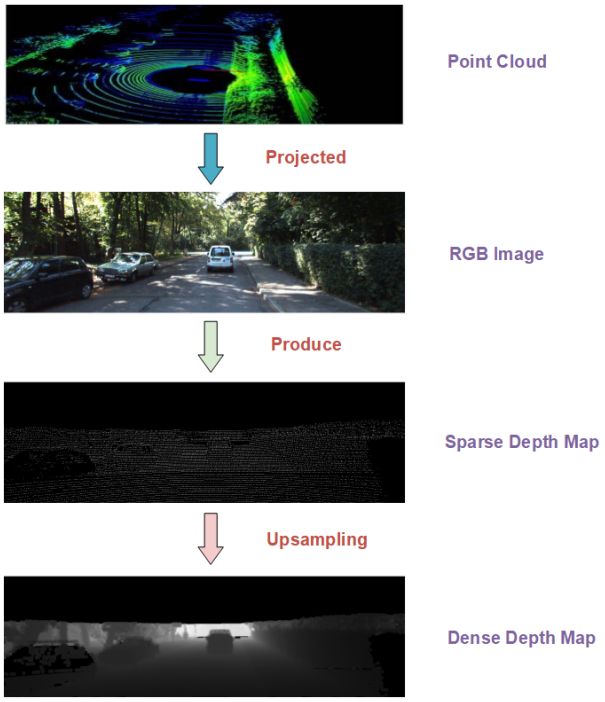

Fig. 1: Schematic of dense depth map generation.

A. Dense depth map generation

In order to generate the aligned dense depth map, we utilize the upsampling technique proposed in [8]. The complete schematic is shown in Fig. 1. The method with Bilateral filtering formalism uses only range data to produce depth map. First the 3D Lidar data is projected onto the 2D image plane and aligned with the image through coordinate transform using calibration matrix. However, the produced depth map is too sparse for detection task thus straightforward leading to a upsampling process. The formula of upsampling is given by:

$$D_p = \frac{1}{W_p} \sum_{q \in N} I_q G_{\sigma_r}(|I_q|) G_{\sigma_s}(\|p-q\|) \qquad (1)$$

where D_p is the output dense map generated by the upsampling approach, I_q is the sparse depth map which is produced by projecting the original 3D Lidar point cloud data onto the image plane, here we assume the camera and Lidar have already been calibrated. N represents the image space, while index $()_p$ and index $()_q$ denote the pixel coordinate of D and I, respectively. $G_{\sigma_r}(|I_q|)$ is a penalty function of the impact of points of range values, and $G_{\sigma s}(\|p-q\|)$ denotes the reciprocal Euclidean distance of pixel coordinate p and q. At last, W_q is a normalization factor which weights sum to one.

B. Class fusion using evidence theory

The typical implementation of Evidence Theory (ET) is in homogeneous sensors, but it is also capable in

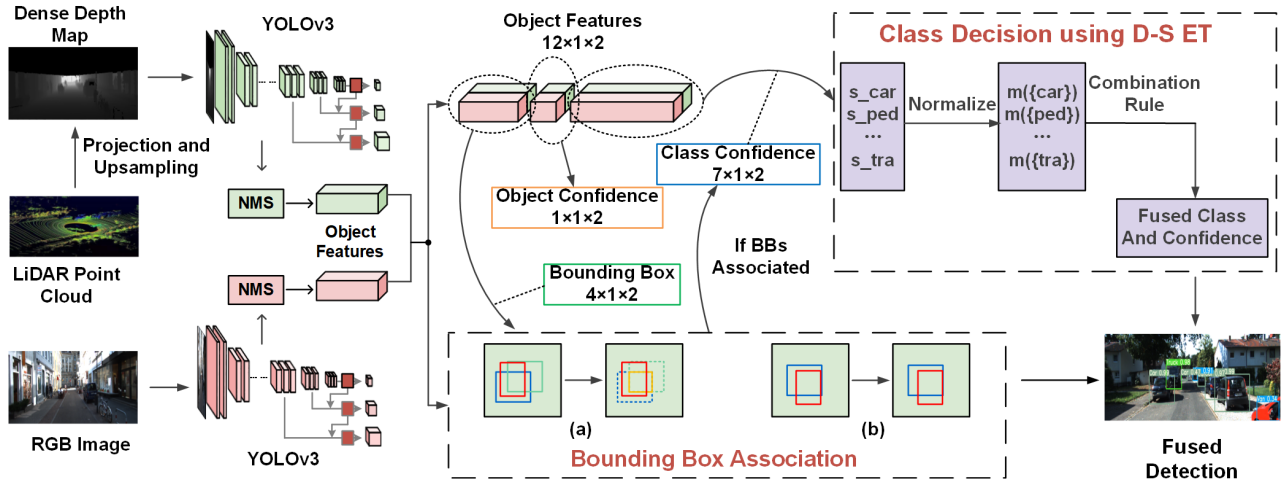

Fig. 2: Architecture of our proposed method. The leftmost presents data preprocess, the middle shows detail of object features extracted from YOLOv3 framework, the two dashed rectangular boxes express the detection fusion process, respectively. More details are described in section II.

heterogeneous ones and can be utilized in classification tasks like [14]. In this work, our concentration is on the two heterogeneous sensors, i.e. the monocular camera and Lidar sensor. D-S evidence theory is an uncertain reasoning theory based on statistics that tells the degrees of belief of given problems. Generally, we denote $\Theta = \theta_1, \theta_2, ...\theta_N$ as the discernment framework. In this work, Θ includes seven different classes, i.e. $Car, Pedestrian, Cyclist, Van, Truck, Person\_sitting, Tram$. The D-S evidence theory uses mass function, also known as basic belief assignment (BBA), to describe how trustable the evidence is. In a frame of discernment, BBA or mass function denoted by $m()$ is defined as the map $2^\theta \rightarrow [0, 1]$ which contains the follow conditions:

$$\begin{cases} m(\phi) = 0, \\ \sum\limits_{A \subset \Theta} m(A) = 1 \end{cases} \tag{2}$$

where ϕ is an empty set. A is a subset from the discernment Θ. If set A only contains one class, then it is called singleton. $m(A)$ expresses degrees of belief of set A. The basic belief assignment in class confidence fusion is defined as follows:

$$\begin{cases} m_i(A) = \alpha_i S_i(A) & A \subset \Theta \\ m_i(\Theta) = 1 - \alpha_i \end{cases} \tag{3}$$

where $i \in \{1, 2\}$ represents two different sensors, i.e. camera and Lidar. S_i is the class confidence obtained from original detection network. $\alpha_i \in [0, 1]$ is evidence discounting factor. Here we assume there are two mass functions m_1 and m_2 gained from two different sensors, then the combined mass function is produced using Dempster's combination rule:

$$m(C) = (m_1 \oplus m_2)(C) = \begin{cases} 0 & C = \phi \\ \frac{\sum\limits_{A \cap B = C} m_1(A) m_2(B)}{1 - K} & C \neq \phi \end{cases} \tag{4}$$

where $1 - K$ denotes the normalization factor. It represents the conflict among evidence from different sources. K is defined as:

$$K = \sum_{A \cap B = \phi} m_1(A) m_2(B) \tag{5}$$

$K = 0$ indicates no conflict between A and B, while $K = 1$ represents totally conflict. The mass function after combination has the same frame of discernment.

In [14], the author holds the point that object with bigger size is hardly predicted as an smaller object, e.g. a truck won't be recognized as a pedestrian. However, the objects closer to the sensors seem to be bigger than farther ones after projected onto the image plane. Meanwhile, as the object class number grows the complexity increases quite fast. The method allocates handcrafted evidence distribution is unreasonable. For these reasons, in this work all classes of objects are considered to own the same weight. In order to utilize the theory in evaluating the final object detection confidence, [6] transforms the classification results of each unary classifier, which consists of the probability proportions about classes of interest, into a BBA mass function form using pignistic transformation. However, the mass functions in this work are obtained directly using the class confidence output of each detector. Since confidence scores of all classes are already separated, the pignistic transformation is then unnecessary.

C. Details of system framework

The architecture of our system illustrated in Fig. 2 is comprised of three parts: data preprocess, YOLOv3 detector and NMS, and decision fusion. First, we project the 3D Lidar data onto the 2D pixel plane where the camera and Lidar have been calibrated. After the projection, we get sparse depth map which contains inadequate information, therefore an upsampling method is used to produce the dense depth map. Since two types of data from different sensors are obtained, they can be fed into YOLOv3 to train the two separate models individually. The individual detector will

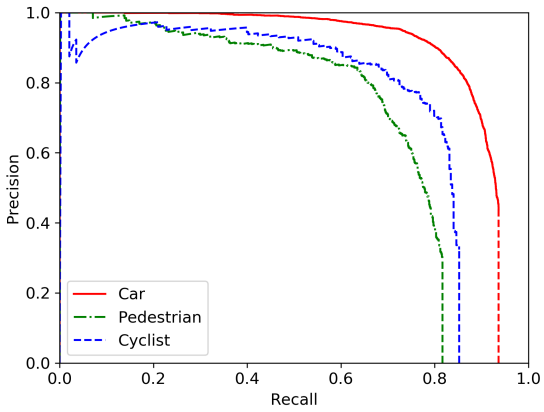

Fig. 3: P-R curves of our proposed method. The result of class Car is in red, Pedestrian in green, and Cyclist in blue, respectively. Only the three classes which contain enough sample labels are drawn in this picture.

create two different detection results which have their own bounding boxes and class confidence (or called scores). The BBs and scores from camera and Lidar are denoted by BB_C, s_C and BB_D, s_D, respectively.

The original network of YOLOv3 predicts numbers of useless or duplicated bounding boxes in the output, which causes a powerful algorithm named NMS connected with it. Generally NMS generates predicted objects only with the highest class confidence, but in this work confidence of all classes is kept during NMS process so that the format is consistent with evidence theory. The format of these two confidence set is intuitively similar to the mass function of D-S evidence theory, therefore transforming the confidence set into mass function the Dempster's combination rule of evidence theory can be easily applied.

Each prediction produced by NMS consists of a 12-dimensional vector $(Box, Obj\_conf, Class\_conf)$, where Box is 4-dimension with the upper left and lower right corner of the predicted bounding box, Obj_conf which is a scalar denotes how possible the bounding box contains an object, and $Class\_conf$ with 7 dimension respectively presents confidence of all classes. Meanwhile, confidence of all the classes are normalized.

Before the class confidence fusion process, the detection results from individual detectors need to be confirmed whether they are associated. We keep the same frame of discernment Θ of class similarity CS which is defined in Equ. (6). As for position similarity PS we use Intersection over Union (IoU) defined in Equ. (7) since IoU with scale invariance can directly reflect the coincidence degree of two BBs.

$$CS = 1 - \sum_{A \cap B = \phi} m_1(A) m_2(B) \tag{6}$$

$$PS = \frac{Area(BB_I) \bigcap Area(BB_G)}{Area(BB_I) \bigcup Area(BB_G)} \tag{7}$$

where BB_I denotes the bounding box predicted from RGB image BB_C or depth map BB_D, and BB_G from ground truth, respectively.

We denote two BBs as association only if the class and position similarity of them is bigger than threshold which is chosen through experiments ($CS > 0.68$, $PS > 0.80$). The detected bounding boxes from two individual detectors are denoted by a_i, b_j, where $i = 1, 2, ..., M$, $j = 1, 2, ..., N$. M, N represent the total number of BBs detected from camera and Lidar, respectively. An association matrix with the dimension of $M \times N$ is constructed to explicitly describe whether the two BBs are associated. The element $t_{i,j}$ from the association matrix will be set to one if a_i and b_j are associated, otherwise it will be set to zero. If a BB is not associated with any BBs predicted in another model, keep this complementary information if score of the BB is over a threshold, otherwise drop it.

In Fig. 2, the dashed rectangle at the bottom shows two situations. (a) in the left presents that if two detected BBs (blue and green boxes) are associated, then the intersection area between them (the yellow solid box) is extracted. (b) in the right presents that if only one BB (the blue box) is detected but the score is over the confidence threshold, then keep this BB reserved (in both situation red box represents the ground-truth). According to the association matrix, class confidence scores are fused using D-S evidence theory, and bounding boxes are extracted from the intersection area of the associated boxes, respectively.

III. Experiment and Evaluation

This section presents the elaborate implementation details of our experiments. For better evaluation of our method, we use public KITTI vision dataset [15] to train and test. The label files from KITTI dataset are transformed to Pascal VOC format [16] for convenience.

A. Experiment setup and KITTI dataset

The KITTI dataset consists of 7481 training images and 7581 test images which contain 51867 labeled objects. The majority of these images have spatial resolution of 1242×375 pixels. The original training images are separated into two parts: 80% for training, 20% for testing, i.e. 5985 images used in training process while 1496 in testing. There are nine different object categories in KITTI, but in our experiments we only use seven of them. The remained two categories "Misc" and "Don't Care" are abandoned since they are not actual object categories. We create three types of datasets extracted from original KITTI dataset and train them separately. One of the datasets is directly the RGB images from the monocular camera, one is the dense depth maps using upsampling technique [8], the last is a mixed dataset containing both RGB images and high resolution depth maps.

The experiments are conducted using one P6000 GPU. In the training process, the images are trained for 180000 iterations. We compute the Average Precision (AP) metric following the Pascal VOC dataset [16]. The AP index is calculated by area under the Precision-Recall (P-R) curve.

Fig. 4: Some examples of detection results. Detection results are marked in green boxes, while the red dashed rectangles represent ground truth.

TABLE I: Quantitative results on the KITTI dataset. The Precision, Recall, AP, and F1-score metrics with all classes are evaluated.

CLASS	PRECISION			RECALL			AP			F1-score		
	RGB	Lidar	Fusion	RGB	Lidar	Fusion	RGB	Lidar	Fusion	RGB	Lidar	Fusion
All(mean)	0.175	0.108	**0.396**	**0.902**	0.887	0.898	0.816	0.790	**0.821**	0.289	0.192	**0.546**
Car	0.236	0.176	**0.444**	0.940	**0.942**	0.936	**0.896**	0.881	0.890	0.377	0.296	**0.602**
Pedestrian	0.0798	0.0728	**0.302**	0.814	0.800	**0.816**	0.679	0.673	**0.709**	0.145	0.134	**0.441**
Cyclist	0.0886	0.0835	**0.333**	**0.864**	0.855	0.846	0.751	**0.753**	**0.753**	0.161	0.152	**0.478**
Van	0.201	0.135	**0.480**	**0.979**	0.945	0.962	**0.942**	0.881	0.920	0.334	0.235	**0.640**
Truck	0.283	0.110	**0.558**	**0.981**	0.958	0.967	**0.965**	0.922	0.954	0.440	0.197	**0.708**
Person_sitting	0.124	0.0653	**0.261**	**0.769**	0.731	**0.769**	0.599	0.552	**0.616**	0.214	0.120	**0.390**
Tram	0.214	0.115	**0.395**	0.966	0.978	**0.989**	0.878	0.865	**0.903**	0.350	0.206	**0.564**

B. Results evaluation

Results of the three different models are presented in Table I. The first column indicates seven different object classes in which "All" represents average value of all classes of these metrics. The "RGB" and "Lidar" denote the original YOLOv3 framework with color images and dense depth maps fed into, respectively. "Fusion" represents the proposed fusion model with both camera and Lidar data fed into. From Table I we can see, after fusion process the model using fusion data reaches higher performance of precision, AP and F1-score (22.1%, 0.5% and 25.7%, respectively). This proves that using our fusion framework which benefits from both RGB and range data achieves performance enhancement in detection result. However, due to the inherent drawback of YOLOv3 which may detect objects with incorrect localizations, our proposed method simply extracting the intersection of the associated boxes can deviate from ground truth which cause the lightly decrease (0.4%) in recall metric. In addition,

detection results from range data achieve relatively lower performance comparing with RGB data, since converting 3D Lidar point cloud into 2D depth map will potentially lose some vital features. We believe a better expression of Lidar data would reach better performance.

Fig. 3 presents the P-R curves of our proposed method. Some detection results are presented in Fig. 4. The left column shows the detection examples using only RGB images, the middle column using dense depth maps, while the right column presents the fusion results in the same scene. Table II shows our fusion approach compared with some state-of-the-art results from KITTI benchmark. For fair comparison, results from KITTI benchmark containing object difficulty level of "moderate" which constrains that each bounding box is at least 25 px height and the maximum truncation ratio is 30%. As can be seen in Table II, our proposed method outperforms some of the methods from the benchmark and has the shortest runtime which makes it possible to be applied in the practical intelligent mobile robots and autonomous vehicles.

TABLE II: Fused detection performance compared with some state-of-the-art results on KITTI benchmark.

Method	Car	Pedestrian	Cyclist	Runtime
TuSimple [17]	94.47%	78.40%	75.22%	1.6s
RRC [18]	93.40%	76.61%	76.71%	3.6s
DeepStereoOP [19]	90.06%	68.46%	67.22%	3.4s
AVOD-FPN [12]	88.92%	57.87%	60.79%	0.1s
CLA [20]	88.86%	75.16%	75.48%	0.3s
Regionlets [21]	76.99%	60.83%	58.52%	1s
Fusion(Our Method)	89.00%	70.90%	75.30%	0.072s

IV. Conclusion and future work

In this study, a decision level fusion of multi-sensor based object detection method is proposed using D-S evidence theory. First, raw 3D Lidar point cloud is projected onto the image plane to generate the sparse maps, where the upsampling method is applied to produce high-resolution dense depth maps. Then the RGB images and depth maps are separately fed into a detection framework named YOLOv3 which outputs the object related features. The final bounding box and class are generated by using Dempster's combination rule and extracting intersection of the associated bounding boxes, respectively. The experiment results show that by fusing the composite information from the camera and Lidar, 0.5% and 3.1% promotion of mAP compared to single sensor camera and Lidar is achieved with only 0.072s runtime. In the future, the proposed method will be applied into real mobile robots, and a novel and reliable presentation of Lidar data will also be studies.

References

[1] Alireza Asvadi, Luis Garrote, Cristiano Premebida, Paulo Peixoto, and Urbano J Nunes. Multimodal vehicle detection: fusing 3d-lidar and color camera data. *Pattern Recognition Letters*, 115:20–29, 2018.

[2] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems 28*, pages 91–99. 2015.

[3] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

[4] Liang Xiao, Ruili Wang, Bin Dai, Yuqiang Fang, Daxue Liu, and Tao Wu. Hybrid conditional random field based camera-lidar fusion for road detection. *Information Sciences*, 432:543–558, 2018.

[5] Andreas Pfrunder, Paulo VK Borges, Adrian R Romero, Gavin Catt, and Alberto Elfes. Real-time autonomous ground vehicle navigation in heterogeneous environments using a 3d lidar. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2601–2608, 2017.

[6] Sang-Il Oh and Hang-Bong Kang. Object detection and classification by decision-level fusion for intelligent vehicle systems. *Sensors*, 17(1):207, 2017.

[7] Joel Schlosser, Christopher K Chow, and Zsolt Kira. Fusing lidar and images for pedestrian detection using convolutional neural networks. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 2198–2205, 2016.

[8] C. Premebida, J. Carreira, J. Batista, and U. Nunes. Pedestrian detection combining rgb and dense lidar data. In *IEEE/RSJ International Conference on Intelligent Robots and Systems(IROS)*, pages 4112–4117, 2014.

[9] G. Melotti, C. Premebida, N. M. M. d. S. Goncalves, U. J. C. Nunes, and D. R. Faria. Multimodal cnn pedestrian classification: A study on combining lidar and camera data. In *International Conference on Intelligent Transportation Systems (ITSC)*, pages 3138–3143, 2018.

[10] Yufu Qu, Guirong Zhang, Zhaofan Zou, Ziyue Liu, and Jiansen Mao. Active multimodal sensor system for target recognition and tracking. *Sensors*, 17(7):1518, 2017.

[11] Xiaozhi Chen, Huimin Ma, Ji Wan, Bo Li, and Tian Xia. Multi-view 3d object detection network for autonomous driving. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1907–1915, 2017.

[12] Jason Ku, Melissa Mozifian, Jungwook Lee, Ali Harakeh, and Steven L Waslander. Joint 3d proposal generation and object detection from view aggregation. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1–8, 2018.

[13] Yanfei Zhong, Qiong Cao, Ji Zhao, Ailong Ma, Bei Zhao, and Liangpei Zhang. Optimal decision fusion for urban land-use/land-cover classification based on adaptive differential evolution using hyperspectral and lidar data. *Remote Sensing*, 9(8):868, 2017.

[14] R. O. Chavez-Garcia and O. Aycard. Multiple sensor fusion and classification for moving object detection and tracking. *IEEE Transactions on Intelligent Transportation Systems*, 17(2):525–534, 2016.

[15] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, pages 3354–3361, 2012.

[16] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010.

[17] Fan Yang, Wongun Choi, and Yuanqing Lin. Exploit all the layers: Fast and accurate cnn object detector with scale dependent pooling and cascaded rejection classifiers. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2129–2137, 2016.

[18] Jimmy Ren, Xiaohao Chen, Jianbo Liu, Wenxiu Sun, Jiahao Pang, Qiong Yan, Yu-Wing Tai, and Li Xu. Accurate single stage detector using recurrent rolling convolution. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5420–5428, 2017.

[19] Cuong Cao Pham and Jae Wook Jeon. Robust object proposals re-ranking for object detection in autonomous driving using convolutional neural networks. *Signal Processing: Image Communication*, 53:110–122, 2017.

[20] Chen Zhang and Joohee Kim. Object detection with location-aware deformable convolution and backward attention filtering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9452–9461, 2019.

[21] Xiaoyu Wang, Ming Yang, Shenghuo Zhu, and Yuanqing Lin. Regionlets for generic object detection. In *IEEE International Conference on Computer Vision*, pages 17–24, 2013.