Proceedings of the 35th Chinese Control Conference
July 27-29, 2016, Chengdu, China

Frontal Object Perception for Intelligent Vehicles Based on Radar and Camera Fusion

HAN Siyang, WANG Xiao, XU Linhai, SUN Hongbin, and ZHENG Nanning
Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, Xi'an, Shaanxi 710049, P. R. China
E-mail: wangxiao.0312@stu.xjtu.edu.cn

Abstract: This paper addresses the issue of frontal object perception in real-world traffic scenarios. Accurate and real-time frontal object perception plays a key role in Advanced Driver Assistance Systems (ADAS) and Intelligent Vehicles (IV). However, perceiving complex traffic environments, which contain multiple classes of on-road objects with various visual appearances from different viewpoints and partial observations, is still a challenging task. In this paper, a perception system fusing a millimeter-wave (MMW) radar and a monocular camera is proposed. Firstly, the detections of MMW radar are converted to regions of interest (ROIs) on the image. Then, these ROIs are classified by four classifiers using Deformable Part Model (DPM). Finally, a mixer module is used to combine all the classification results and infer the final result for each ROI. The computation intensity of the DPM algorithm can be efficiently reduced through this mechanism. Meanwhile, high detection precision is achievable. Experiment results show that the proposed frontal object perception system can detect and classify on-road objects in complex urban traffic scenarios with 98.4% detection rate at nearly real-time performance (29Hz).

Key Words: Sensor Fusion, MMW Radar, Monocular Camera, Object Detection, Scene Perception

1 Introduction

Frontal object perception is one of the key technologies for ADAS or IV to perceive traffic environments. It can directly impact the safety and smartness of the entire system. MMW radar and monocular camera are two sensors normally used for vehicular active-safety systems. The popularity of these two sensors is caused by their individual advantages: MMW radar has all-weather working capacity and good detection performance; Monocular camera benefits from low cost and small size. However, MMW radar and monocular camera both have their own shortcomings when working individually. Due to the position uncertainty of the microwave reflection point on the detected object, MMW radar has poor lateral detection accuracy [1]. For camera, those complicated algorithms used for vision based object detection and classification require high computational intensity to guarantee good accuracy and thus has poor practicability [2]. Fusion of these two sensors is thus a promising approach to complement one another efficiently.

There have been many studies about MMW radar and monocular camera fusion. To our knowledge, Amir et al. [3] firstly proposed a radar and vision fusion scheme. After matching radar and vision targets, vertical and horizontal edges from image are used to judge whether the detection is solid, ghost or occluded. Similar methods are further proposed, such as using edge symmetry features [4] or shadow features [5], to locate and verify the detected vehicle. However, these methods assume the detected objects to be vehicles, and there was no classification step to determine the type of the object. Methods using machine learning algorithms, such as AdaBoost [6], SVM [7] [8], or neural network [9], are also proposed to validate the object detected by MMW radar. However, the types of identified objects are limited to vehicles, and these methods are vulnerable when dealing with challenging visual appearances of objects, such as partial observations or occluded situations.

Recently, part model based object detection algorithms like DPM [10] and deep learning based algorithms like Fast R-CNN [11] have high classification precision and the ability to detect multiview and occluded objects. This is a crucial feature for frontal object perception under complex traffic conditions. Although many studies have been proposed to push CNN based algorithms toward real-time implementations, their work always require high performance computing platform with multiple GPUs, of which the tremendous power consumption are inappropriate for vehicular applications. By comparing the performance of both algorithms on our platform, we finally choose DPM for monocular camera based objects detection. However, it is difficult to achieve an accurate and fast vision-only frontal object perception system for the following reasons: the construction and using of multi-layer Gaussian pyramids along with scanning window searching strategy in DPM algorithm lead to intense computational workloads. Meanwhile, the false-negative instances caused by irrelevant background is more than acceptable.

In this paper, a frontal object perception system based on fusion of MMW radar and monocular camera is proposed. At the beginning of our approach, the positions of objects detected by MMW radar are projected onto the image plane to get corresponding ROIs. Then, four well trained DPM based classifiers are deployed to recognize the different classes of objects within the ROIs. Finally, the classification results are mixed through a mixer module. This module combines each classification result in a probabilistic inference framework to generate the final detection result. Meanwhile, we designed a negative sample generating procedure with the help of MMW radar and dynamic Gaussian background model. This procedure can efficiently improve the detection performance of the proposed system.

To evaluate the proposed system, datasets consisted of several typical urban traffic scenarios in Xi'an are created by

This work is supported by National Natural Science Foundation of China (NSFC) under Grant 61231018.

our intelligent vehicle platform. Experiments and comparisons show that proposed system can perceive four classes of frontal on-road objects with a 98.4% detection rate at almost real-time performance (29 Hz). The proposed system had also been installed in our intelligent vehicle Challenger for the 6th autonomous vehicle competition Intelligent Vehicle Future Challenge' (IVFC) in China, and successfully detected all on-road objects during the match.

The remainder of this paper is organized as follows. Section 2 briefly describes the system platform and sensor spatial alignment method. Section 3 presents the proposed training sample generating procedure, together with trained DPM classifiers. Section 4 presents the details of each step in our fusion approach. The performance of the proposed system is extensively evaluated in Section 5. Section 6 summarizes this paper.

2 System Platform and Sensor Alignment

2.1 System Platform

The self-driving vehicle platform for the presented study and experiments is based on a 1.6 L Tiggo SUV manufactured by Cherry Automobile Co. and called 'Challenger'. 'Challenger' is the successor of our former platform 'Kuafu-II'. It had participated in the 6th autonomous vehicle competition 'Intelligent Vehicle Future Challenge' (IVFC) in 2015, organized by Natural Science Foundation of China (NSFC). The MMW radar is a Delphi ESR bi-mode radar mounted on the front bumper, and the camera is a PointGray FMUV-03MTC color camera mounted behind the front windshield, as shown in Fig. 1 respectively. The detailed specifications of these two sensors and the computational platform are the same as presented in [12], and are not further described.

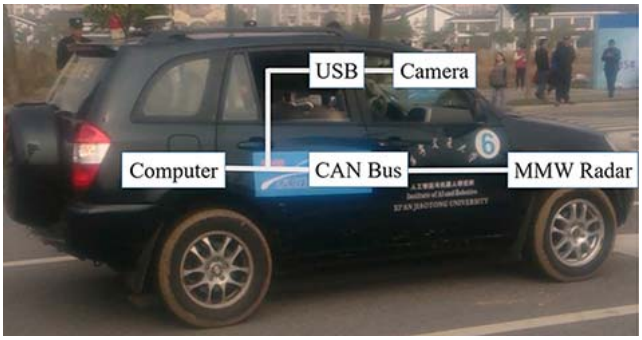


Fig. 1: The hardware platform of our 'Challenger' intelligent vehicle during the 6th IVFC competition.

2.2 Spatio-Temporal Alignment Between Sensors

The first step of fusing MMW radar and monocular camera is to transform the MMW radar's detection to a pixel point on the image plane. Perspective transformation is commonly used to describe this relationship. This step is also called spatial alignment between sensors. Similar to [13], we also adopt a multi-model approach to achieve good spatial alignment precision. The position of detected object by MMW radar can be described as (x, y) on a horizontal plane. The position of detected object on the image captured by monocular camera can be regarded as (u, v). In the beginning, the detection range of MMW radar is divided with a step of $10m$ from $0m$ to $100m$. Then 10 transformation matrixes are estimated respectively for the n-th (1st-10th) division through following equation.

$$\begin{cases} u^n = & \dfrac{a_1^n x + a_2^n y + a_3}{a_7^n x + a_8^n y + a_9^n} \\ v^n = & \dfrac{a_4^n x + a_5^n y + a_6}{a_7^n x + a_8^n y + a_9^n} \end{cases} \tag{1}$$

Please refer to [12] for more details about the calibration procedure. Finally, a ROI on the image is chosen as a $4m \times 4m$ square around the calculated pixel position of the detected object.

The frame rate of camera (60 Hz) is the same as the detection speed of MMW radar. Temporal alignment is performed by externally triggering the camera to allow for synchronous detection with MMW radar.

3 Sample Generation and Classifiers Training

The frontal objects in real-world traffic scenarios have following challenging characteristics:

- Vehicles are not the only type of on-road objects. In our dataset, the frontal objects also include vehicles, pedestrians, two wheels, and traffic cones.
- The visual appearances of frontal objects change dramatically across different viewpoints. Partial observations always happen due to occluded situations or limited field of view of camera.
- Frontal objects have different relative distances and speeds. The sizes of them on image change dramatically according to their distances.

To our knowledge, conventional machine learning algorithm, such as NN or SVM, using Hog or Haar-like features are short of handling above challenges [7–9]. Recently, detection methods using model based machine learning algorithms like DPM [10] or deep learning algorithms like Fast R-CNN [11] have achieved good performance and pratical potentials especially when dealing with multi-class recognition tasks and partial viewed objects. After comparing these two algorithms on our platform which is shown in Section 5, we finally choose DPM for monocular camera based objects detection. However, a monocular camera based object perception system based on DPM algorithm still requires heavy computation workloads, which is short of practicality. Hence, a radar guided system is further introduced to handle this problem efficiently in this paper. In the following subsections, our procedure to choose training samples is proposed which is essential for the overall detection performance. The trained classifiers are also presented.

3.1 Training Sample Generation

In this study, the on-road objects are divided into 4 classes: vehicles, pedestrians, two wheels and traffic cones. Firstly, we adopt the positive samples from well-known datasets, such as PASCAL VOC2010 dataset [14], INRIA Person dataset [15]. The positive samples from our datasets are further added and mixed. In the training process of DPM, bounding boxes are used as latent information: The ratios between height and width of the bounding boxes are used to divide positive samples into several sub-classes. These subclasses represent different views of the same type of object.

However, in real-world traffic scenarios, this separation is error prone due to the various postures and shapes of objects. By additionally assigning a view label for each sample manually, positive samples can be separated more properly for each sub-model. This method also makes the training results more representative.

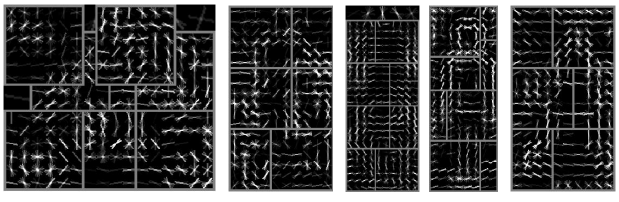
As a matter of fact, a better choose of negative samples can efficiently decrease the number of false-negative instances of the detection system. By using the coherent information from adjacent frames and detection results from MMW radar, a dynamic background is generated and used to provide sufficient amount of negative samples for classifiers training. The algorithm to detect such a dynamic background is presented here. For an arbitrary pixel P_n on the n-th image frame, if it falls in the ROIs O_n provided by MMW radar, this pixel will be considered as an object pixel, otherwise it is regarded as a background pixel. We collect all background pixels in a number of adjacent frames in a histogram manner and take the Gaussian expectation of them to make up the corresponding background image for these periods of frames. This procedure is briefly shown in Fig. 2. The negative samples are chosen randomly from all of the dynamic background images later.

(a) (b) (c)

Fig. 2: The procedure of dynamic background detection (a) the original image frame, (b) the ROI of detected object is cut off, and (c) the background is added to the ROI.

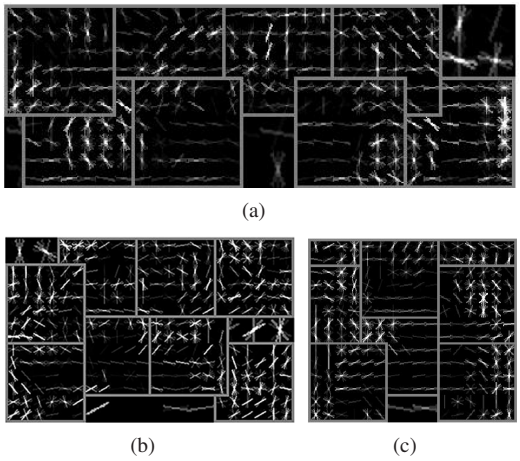
3.2 Trained Classifiers

lSVM with HOG features are used to train four DPMs based on image patch samples. Fig. 3-4 illustrate all the models trained for different classes of objects in our dataset. As illustrated, three sub-models are trained for vehicles with regard to three different viewpoints. Three sub-models are trained for two wheels for the same purpose. One sub-model for pedestrians and another one for traffic cones are trained respectively.



(a) (b) (c) (d) (e)

Fig. 3: Three sub-models trained for bicycle class (a) the lateral side model, (b) the tilt rear side model, and (c) the rear side model. Two models trained for pedestrian class and traffic cones (d) the pedestrian model, (e) the cone model.



(a) (b) (c)

Fig. 4: Three sub-models trained for the vehicle class (a) the lateral side model, (b) the tilt rear side model, and (c) the rear side model.

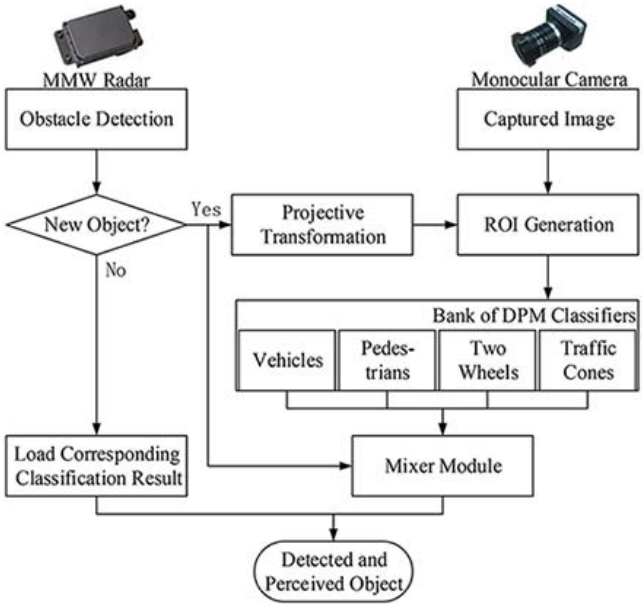


Fig. 5: Overall architecture of the proposed frontal object perception system.

4 Proposed Frontal Object Perception System

In this section, the overall architecture of the proposed frontal object perception system is described in detail. In addition, a mixer module using probabilistic inference is designed to combine outputs from all classifiers and output the final recognition results.

4.1 Overall Architecture of Sensor Fusion

The overall architecture of the proposed detection system is illustrated in Fig 5. Firstly, MMW radar detects on-road objects and transmits the location and size of ROI to the image sequences captured by monocular camera. Then, four parallel DPM based classifiers are deployed to recognize what the object is in the ROI. Finally, the recognition results from all classifiers and the detections of MMW radar are altogether input into a mixer module to determine what the object is. The type and position of the object are outputted as the final perception result. It is worth mentioning that the

used MMW radar also has the ability to track each individual detection. So the recognition step only takes place when a new object is detected. The classification result is kept until the object is changed or missed by MMW radar.

4.2 Mixer Module

For four classes of objects: C_1 for vehicles, C_2 for pedestrians, C_3 for two wheels and C_4 for traffic cones, four parallel DPM classifiers $D_n(n = 1..4)$ are deployed in the proposed system. Each of them can detect object O within the ROI and give out the detected region R_n and the corresponding confidence score s_n. If $s_n > s_0$, the result is stored, otherwise it is discarded. The s_0 is set as -0.8 in the experiment. The mixer module is in charge of judging which recognition result is more reasonable especially when classifiers get conflicting results. The judgement is realized in a probabilistic inference manner. The posterior possibility of the class C_n over the object O is written on Bayesian rules as following equation.

$$P(C_n|O) = \frac{P(O|C_n)P(C_n)}{P(O)} \tag{2}$$

The $P(O)$ is a normalized factor which can be calculated as follows.

$$P(O) = \sum_{n=1}^{4} P(O|C_n)P(C_n) \tag{3}$$

$P(O|C_n)$ shows the probability of the object O belonging to the class C_n. As confidence score s_n outputted by DPM classifiers is a real value among (s_0, ∞), $P(O|C_n)$ is calculated according to following equation.

$$P(O|C_n) = \begin{cases} e^{s_n}, & s_n < 0 \\ 1, & otherwise \end{cases} \tag{4}$$

The prior possibility $P(C_n)$ is designed with help of the object's distance y and speed v detected by MMW radar for the following reasons: Each kind of object has its unique size and shape, but the resolution of camera is limited. When the distance of the detected object is farther than a maximum distance, its region on the image will be too small to be recognized. Such maximum distance is defined as y_n^{max}. On the other hand, the speed of the traffic cones is always zero. Only the vehicles can have high moving speed. Pedestrians usually move slowly and two wheels move faster. The speed information can provide an important cue for recognitions. The maximum speed of each kind of object is defined as v_n^{max}. Thus, the $P(C_n)$ is defined as a piecewise function as follows.

$$P(C_n) = \begin{cases} 1, & y < y_n^{max} \; and \; v < v_n^{max} \\ 0, & otherwise \end{cases} \tag{5}$$

Thus the maximum posterior possibility solution is found by solving:

$$n\_max = \arg\max_n P(C_n|O) \tag{6}$$

Finally, if the posterior possibility $P(C_{n\_max})$ is above 0.5, the corresponding object is classified into n_max-th class. Otherwise, the object is regarded as a non-object detection.

5 Experimental Results and Discussions

The proposed frontal object perception system was implemented in our intelligent vehicle 'Challenger'. Real-world MMW radar data and videos were captured by 'Challenger' for the experiment. Our dataset consists of four kinds of representative urban traffic scenarios in Xi'an, including normal streets, narrow roads, expressways and viaducts, as shown in Fig 6. There are more than 1 hour videos and corresponding MMW radar detections for each scenario. The detected objects in our datasets are manually labeled as ground truth to evaluate the performance of the proposed system.

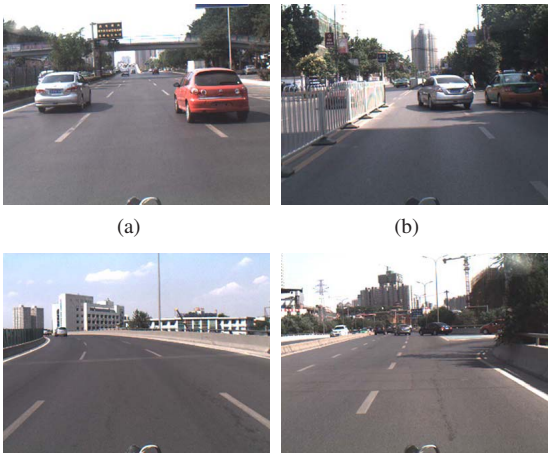


Fig. 6: Four representative urban traffic scenarios in our datasets (a) normal streets, (b) narrow roads, (c) expressways, and (d) viaducts.

5.1 Frontal Object Perception Result

Instances of detected results for four classes of objects are demonstrated in Fig. 7, 8 and 9. It is worth mentioning that all the figures after Fig. 7 are given under an uniform style in order to simplify the description: the image patches themselves depict the ROIs guided by MMW radar detections, and the red rectangles inside these figures show the detection results from the proposed fusion system. Fig. 10 illustrates that our system can easily handle difficult situations even when an object is heavily occluded.

5.2 Performance Analysis of the Overall System

The evaluation criteria used here is detection precision along with detection recall. Actually, detection rate is the same as detection recall in this study. Table 1 illustrates the detection performance in each traffic scenario. In our experiment, not only on-road objects, but also road curbs, guard rails or trees can be detected by MMW radar. These objects are not considered as on-road objects in this study, thus they are described as non-objects. Once the posterior possibility $P(C_{n\_max})$ in formula (6) is below 0.5, the corresponding object will be regarded as a detected non-object. The number of detected non-objects is the same as the amount of true-negative instances. According to the experiment results, the average precision of the proposed system is at 98.3%, and the average recall (detection rate) is at 98.4%.

Table 1: Experiment Results in Our Dataset. G. Represents the Total Number of Ground Truths. C. Represents the Total Number of Confirmed Objects. R. Represents the Detection Rate. AR. Represents the Average Detection Rate.

Scenarios	Normal Streets			Narrow Roads			Expressways			Viaducts			AR.
	G.	C.	R.	G.	C.	R.	G.	C.	R.	G.	C.	R.	
Vehicles	900	890	98.9%	709	706	99.6%	721	714	99.0%	635	631	99.4%	99.2%
Pedestrians	319	301	94.3%	262	252	96.2%	88	83	94.3%	0	0	-	95.5%
Two Wheels	100	99	99.0%	46	43	93.5%	51	51	100%	0	0	-	98.0%
Traffic Cones	51	50	98.0%	91	90	99.0%	46	45	97.8%	20	20	100%	98.6%
All Objects	1370	1340	97.8%	1108	1091	98.5%	906	893	98.6%	655	651	99.4%	98.4%
Non-Objects	1261	1243	-	1446	1417	-	929	921	-	928	916	-	-

(a) (b) (c)

Fig. 7: Three instances of vehicle detections with different views (a) a lateral side detection, (b) a tilt rear side detection, and (c) a rear side detection.

(a) (b) (c)

Fig. 8: Three instances of two wheel detection with different views (a) a lateral side detection, (b) a tilt rear side detection, and (c) a rear side detection.

(a) (b) (c)

Fig. 9: Instances of pedestrian detection with different views (a), (b) and traffic cone detection (c) respectively.

(a) (b)

Fig. 10: An instance of occlusion handling dealing with objects at different distances on the same frame. (a) The large ROI and the detection of the near vehicle. (b) The small ROI and the detection of the far and occluded vehicle.

(a) (b) (c)

(d) (e) (f)

Fig. 11: Detection results for different kinds of vehicles (a) sedan (b) minibus (c) SUV (d) Jeep (e) truck and (f) bus.

5.3 Efficiency Analysis of the Overall System

Table 2 shows the average detection time consuming for each class of object. As it illustrates, the average detection time of vehicles and two wheels are longer than pedestrians and traffic cones. The tracking ability of MMW radar can efficiently remove redundant classifications, because the tracked object has already been recognized when it is detected. After removing tracked objects and parallel programming, the average detection time per frame is almost the same as the average detection time per object. The proposed system can work at a framerate of 29Hz on our intelligent vehicle platform, which is ideal for real-world vehicular applications.

Table 2: Average Execution Time for Different Object Detection in the Proposed System

Task	Time Consuming
Average Vehicle Detection Time	$32ms$
Average Two Wheels Detection Time	$32ms$
Average Pedestrian Detection Time	$16ms$
Average Traffic Cones Detection Time	$16ms$
Detection Time per Object	$34ms$

5.4 Performance Comparison

We implemented both DPM and Fast R-CNN approach in order to compare their individual performance on our platform. For DPM, the original version with sliding window searching is implemented. For Fast R-CNN, a VGG-16 network is trained and tested using our datasets. For each image, 3k proposals are generated using EdgeBoxes and further

classified into five classes: vehicles, two wheels, pedestrians, traffic cones and background. Performance comparison between DPM and Fast R-CNN is shown in Table 3. As it illustrates, DPM based approach is more appropriate for our platform. In addition, this table also illustrates the improvement of detection performance and the decrease of detection time per object caused by fusion of MMW radar and monocular camera. The reason for this improvement is twofold: Firstly, the object detected at far distance usually has small ROI on the image which is negligible by DPM algorithm. The fusion approach can efficiently resize different ROIs for detection thus the recall is increased. Secondly, ROIs which contain no valid objects will not be provided by MMW radar. As a result, almost all false-negative instances of DPM algorithm will not appear after sensor fusion. Therefore, the detection precision is also increased.

Table 3: Performance Comparison between Different Approaches. P. Represents Detection Precision. R. Represents Detection Recall. ADT. Represents Average Detection Time per Object.

Approach	P.	R.	ADT.
Fast RCNN	98.3%	96.1%	$3.2s$
DPM	97.5%	93.0%	$0.22s$
DPM with MMW Radar Fusion	98.3%	98.4%	$0.034s$

6 Conclusions

In this paper, an accurate and fast frontal object perception system is proposed for intelligent vehicles to understand traffic environments. The system is based on a fusion approach which combines an MMW radar and a monocular camera. MMW radar detects the object and gives out the corresponding ROI on the image. Then, four DPM based classifiers dealing with four different classes of on-road objects are deployed to classify the object within the ROI. Finally, a mixer module based on probabilistic inference addresses the results given by each classifier and gives out the final detections. Experiments under real-world datasets show that the proposed system achieved a 98.4% frontal object detection rate with an almost real-time performance (29Hz). In the future, we are looking forward to fusing multiple MMW radars and cameras to fulfill a full perception of the traffic environment.

References

[1] S. Sugimoto, H. Tateda, H. Takahashi, and M. Okutomi, Obstacle detection using millimeter-wave radar and its visualization on image sequence, in *Proceedings of the 17th International Conference on Pattern Recognition*, 2004: 342–345.

[2] S. Sivaraman, and M. Trivedi, Looking at Vehicles on the Road: A Survey of Vision-Based Vehicle Detection, Tracking, and Behavior Analysis, *IEEE Transactions on Intelligent Transportation Systems*, 14(4): 1773–1795, 2013.

[3] A. Sole, O. Mano, G.P. Stein, H. Kumon, Y. Tamatsu, and A. Shashua, Solid or not solid: vision for radar target validation, in *Proceedings of IEEE Symposium on Intelligent Vehicles*, 2004: 819–824.

[4] G. Alessandretti, A. Broggi, and P. Cerri, Vehicle and guard rail detection using radar and vision data fusion, *IEEE Transactions on Intelligent Transportation Systems*, 8(1): 95–105, 2007.

[5] T. Wang, N. Zheng, J. Xin, and Z. Ma, Integrating millimeter wave radar with a monocular vision sensor for on-road obstacle detection applications, *Sensors*, 11(9): 8992–9008, 2011.

[6] U. Kadow, G. Schneider, and A. Vukotich, Radar-vision based vehicle recognition with evolutionary optimized and boosted features, in *Proceedings of IEEE Symposium on Intelligent Vehicles*, 2007: 749–754.

[7] Y. Tan, F. Han and F. Ibrahim, A Radar Guided Vision System for Vehicle Validation and Vehicle Motion Characterization, in *Proceedings of IEEE Intelligent Transportation Systems Conference*, 2007: 1059–1066.

[8] X. Liu, Z. Sun and H. He, On-road vehicle detection fusing radar and vision, in *Proceedings of IEEE International Conference on Vehicular Electronics and Safety*, 2011: 150–154.

[9] Z. Ji and D. Prokhorov, Radar-vision fusion for object classification, in *Proceedings of International Conference on Information Fusion*, 2008: 1–7.

[10] P.F. Felzenszwalb, R.B. Girshick, D. McAllester, and D. Ramanan, Object Detection with Discriminatively Trained Part-Based Models, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9): 1627–1645, 2010.

[11] R. Girshick, Fast R-CNN, in *IEEE International Conference on Computer Vision (ICCV)*, 2015.

[12] X. Wang, L. Xu, H. Sun, J. Xin and N. Zheng, Bionic vision inspired on-road obstacle detection and tracking using radar and visual information, in *Proceedings of IEEE Intelligent Transportation Systems Conference*, 2014: 39–44.

[13] L. Jin, M. Fu, Y. Yang and H. Zhu, Space alignment of camera and Millimeter Wave(MMW) radar, in *Proceedings of Chinese Control Conference*, 2013: 4979–4983.

[14] M. Everingham, L. Van Gool, C.K.I. Williams, J. Winn, and A. Zisserman, The PASCAL Visual Object Classes Challenge 2010 (VOC2010) Results, http://www.pascal-network.org/challenges/VOC/voc2010/workshop/index.html.

[15] N. Dalal, and B. Triggs, Histograms of oriented gradients for human detection, in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2005: 886-893.