2018-01-1608 Published 07 Aug 2018

Camera-Radar Data Fusion for Target Detection via Kalman Filter and Bayesian Estimation

Zhexiang Yu, Jie Bai, Sihan Chen, Libo Huang, and Xin Bi Tongji Univ.

Citation: Yu, Z., Bai, J., Chen, S., Huang, L. et al., "Camera-Radar Data Fusion for Target Detection via Kalman Filter and Bayesian Estimation," SAE Technical Paper 2018-01-1608, 2018, doi:10.4271/2018-01-1608.

Abstract

Target detection is essential to the advanced driving assistance system (ADAS) and automatic driving. And the data fusion of millimeter wave radar and camera could provide more accurate and complete information of targets and enhance the environmental perception performance. In this paper, a method of vehicle and pedestrian detection based on the data fusion of millimeter wave radar and camera is proposed to improve the target distance estimation accuracy. The first step is the targets data acquisition. A deep learning model called Single Shot MultiBox Detector (SSD) is utilized for targets detection in consecutive video frames captured by camera and further optimized for high real-time performance and accuracy. Secondly, the coordinate system of camera and radar are unified by coordinate transformation matrix. Then, the parallel Kalman filter is used to track the targets detected by radar and camera respectively. Since targets data provided by the camera and radar are different, different Kalman filters are designed to achieve the tracking process. Finally, the targets data are fused based on Bayesian Estimation. At first, several simulation experiments were designed to test and optimize the proposed method, then the real data was used to prove further. Through experiments, it shows that the measurement noise can be considerably reduced by Kalman filter and the fusion algorithm could improve the estimation accuracy.

Introduction

To develop advanced driving assistance system (ADAS) and automatic driving, real-time and robust on-road target detection is one of the key modules of vehicle environmental perception. Because the on-road driving circumstances are complex and unpredictable, it's necessary for vehicles to equipped with different types of sensors to deal very well with the issues of environmental perception and recognition. Multi-sensor fusion can take advantage of different types of sensors like camera, radar, and laser radar to acquire exactly and completely target information.

Although the millimeter wave radar can provide relatively high range and velocity resolution in bad weather condition, it suffers from limited field of views(FOV), low lateral resolution and incapability of recognition of target type. On the contrary, camera can provide target type but low accuracy of obstacle range estimation. The fusion of camera and radar could make up the defects of the two sensors.

Due to the poor real-time performance of vision-based target detection algorithm, some camera-radar fusion algorithms [1, 2, 3, 4] work as the following steps: Firstly, through coordinate transformation, the radar targets are utilized to determine the regions of interest (ROIs) on the image. Then, the target detection algorithm is just implemented in these ROIs to judge the existence of targets to reduce the run time. Finally, refine the detected targets' boundary.

However, the performance of this kind of algorithm is limited by the capability of millimeter wave radar. Once a target is missed by radar, it can't be detected by the following detection algorithm.

This paper proposes a new fusion method of camera and radar data for target detection via Kalman filter and Bayesian Estimation. Firstly, the SSD algorithm is used to detect targets in images captured by camera, the classification and boundary of targets are obtained. Secondly, the coordinate system of camera and radar are unified by coordinate transformation matrix. Then, the parallel Kalman filter is used to track the targets detected by radar and camera respectively and reduce the noise. Finally, fusion method based on Bayesian Estimation is used to fuse the tracking results of the two sensors.

The paper is organized as follows: Section II introduces the vison-based target detection algorithm, SSD. Section III introduces the coordinate transformational matrix. Section IV introduces the fusion algorithm. Section V and VI present the results of simulation experiments and real experiments respectively. Finally, the conclusions are presented in section VII.

Vision-Based Target Detection

Traditional object detection algorithms based on machine learning utilize the handcrafted features of the target and well-trained classifier. However, these methods are usually with poor real-time performance.

As the development of deep learning method, some object detection algorithms based on convolutional neural network(CNN) such as You Only Look Once (YOLO) or SSD have competitive accuracy and real-time performance. In this paper, SSD [5] is implemented to detect target with frame rate of 11fps on TX2. The final detection results are the classification and boundary of valid targets.

Brief Introduction of SSD

The SSD approach is based on a feed-forward convolutional network that produces a set of bounding boxes of different aspect ratios and scales and scores for the presence of object class instances in those boxes.

Model Structure

The model structure of SSD is shown in Figure 1.

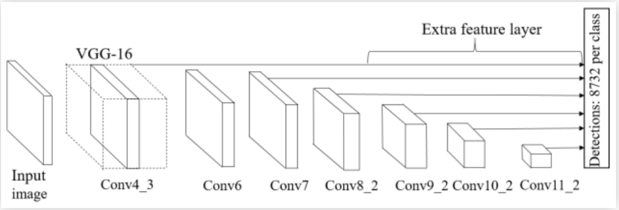
Base network. A truncated standard network is used as early network layers of SSD model for high quality image classification, which is called the base network. In this paper, VGG-16 network is used as the base network.

Multi-scale feature maps for detection. Some extra convolutional feature layers are added to the end of the base network to achieve detection. These layers decrease in size progressively to detect objects at multiple scales.

Convolutional predictors for detection. In each feature layer, a set of convolutional filters are used to produce a set of detection predictions. For a m × n feature layer with p channels, a 3 × 3 × p convolutional kernel is applied at each of the m × n locations to produce either a score for a category, or a shape offset relative to the default box coordinates.

Default boxes and aspect ratios. At each feature map cell, a set of default boxes of different aspect ratios at each location in several feature maps with different scales are produced to predict. For each default box out of k at a given cell in m × n feature layer, the c class scores and the 4 offsets relative to the original default box shape are computed. This results in a total of (c + 4) × k filters that are applied around each location in the feature map, yielding (c + 4) kmn outputs.

FIGURE 1 Model structure of SSD



© SAE International

Train

Default Boxes Generation

Suppose m feature maps are used for prediction. The scale of the default boxes for each feature map is computed as:

$$s_k = s_{\min} + \frac{s_{\max} - s_{\min}}{m-1}(k-1), k \in [1,m] \tag{1}$$

where s_{min} is 0.2 and s_{max} is 0.9, meaning the lowest layer has a scale of 0.2 and the highest layer has a scale of 0.9. And the different aspect ratios of default box are denoted as $a_r\{1,2,3,1/2,1/3\}$. The width and the height for each default box are computed as $w_k^a = s_k\sqrt{a_r}$ and $h_k^a = s_k/\sqrt{a_r}$ respectively. For the aspect ratio of 1, an extra default box whose scale is $s_k' = \sqrt{s_k s_{k+1}}$ is added, resulting in 6 default boxes per feature map location.

Loss Function

The overall objective loss function is a weighted sum of the localization loss (loc) and the confidence loss (conf):

$$L(x,c,l,g) = \frac{1}{N}\left(L_{conf}(x,c) + \alpha L_{loc}(x,l,g)\right) \tag{2}$$

where N is the number of matched default boxes, and the localization loss is the Smooth L1 loss between the predicted box (l) and the ground truth box(g). Confidence loss is the softmax loss over multiple classes confidences (c) and the weight term α is set to 1 by cross validation.

confidence loss:

$$L_{conf}(x,c) = -\sum_{i \in Pos}^{N} x_{ij}^{p} \log\left(\hat{c}_i^p\right) - \sum_{i \in Neg}^{N} \log\left(\hat{c}_i^0\right)$$

$$where\ \hat{c}_i^p = \frac{\exp\left(c_i^p\right)}{\sum_p \exp\left(c_i^p\right)} \tag{3}$$

localization loss (loc):

$$L_{loc}(x,l,g) = \sum_{i \in Pos}^{N} \sum_{m \in \{cx,cy,w,h\}} x_{ij}^{k} \text{smooth}_{L1}\left(l_i^m - \hat{g}_j^m\right) \tag{4}$$

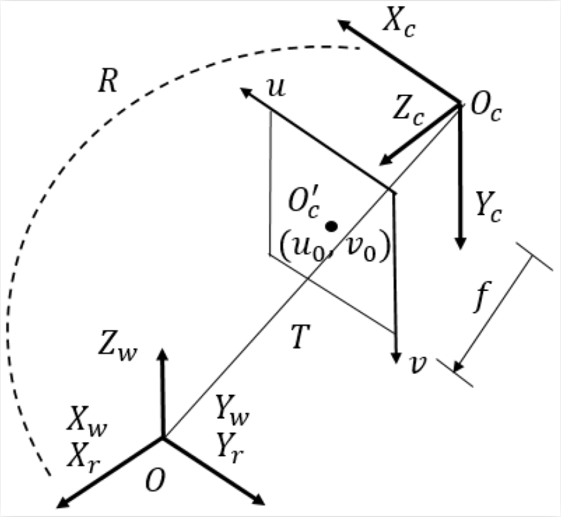
$$\hat{g}_j^{cx} = \left(g_j^{cx} - d_i^{cx}\right) / d_i^w \quad \hat{g}_j^{cy} = \left(g_j^{cy} - d_i^{cy}\right)/d_i^h \tag{5}$$

$$\hat{g}_j^w = \log\left(\frac{g_j^w}{d_i^w}\right) \quad \hat{g}_j^h = \log\left(\frac{g_j^h}{d_i^h}\right) \tag{6}$$

Training Sample

The selection of samples has a great influence on the performance of SSD model.

Vehicle detection training set consists of samples from KITTI [6], GTI Vehicle Image Database [7] and the samples made by ourselves. Pedestrian detection training set consists of samples from CVC pedestrian database [8] and the samples made by ourselves.



© SAE International

FIGURE 2 Coordinate transformational relation of coordinate systems

Conversion of Coordinates

Radar and camera are fixed in different place in vehicle. And the coordinate systems of the two sensors are different. Therefore, the unity of coordinate system of two sensors is the basis of achieving target information fusion. The related coordinate systems include world coordinate system, radar coordinate system, camera coordinate system and image pixel coordinate system. The transformational relation of these coordinate systems is shown in Figure 2.

In this paper, the origin position of world coordinate system is consistent with the counterpart of radar coordinate system. The transformational relation between the camera coordinate system and the world coordinate system is written as:

$$\begin{bmatrix} X_c \\ Y_c \\ Z_c \\ 1 \end{bmatrix} = \begin{bmatrix} R & T \\ 0^T & 1 \end{bmatrix} \begin{bmatrix} X_w \\ Y_w \\ Z_w \\ 1 \end{bmatrix} = M_1 \begin{bmatrix} X_w \\ Y_w \\ Z_w \\ 1 \end{bmatrix} \tag{7}$$

The R is a three order rotation matrix, which is determined by the rotation relation between the two coordinate system. And T is the three-dimension translation vector, which is determined by the origin position of the two coordinate system. The M_1 is external parameter matrix of the camera.

And the transformation relation between the camera coordinate system and the image pixel coordinate system is:

$$Z_c \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} \frac{1}{d_x} & 0 & u_0 \\ 0 & \frac{1}{d_y} & v_0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} f & 0 & 0 & 0 \\ 0 & f & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} X_c \\ Y_c \\ Z_c \\ 1 \end{bmatrix} = M_2 \begin{bmatrix} X_c \\ Y_c \\ Z_c \\ 1 \end{bmatrix} \tag{8}$$

The d_x and d_y represent the physical length of each pixel in the X axis and the Y axis in the image pixel coordinate system, the pixel coordinates (u_0, v_0) is the intersection point of the optical axis Z_c and the image plane, f is the focal length of camera. The M_2 is internal parameter matrix of the camera.

And the transformation relation between the world coordinate system and the image pixel coordinate system could be obtained by the formula (7) and formula (8):

$$Z_c \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = M_1 M_2 \begin{bmatrix} X_w \\ Y_w \\ Z_w \\ 1 \end{bmatrix} \tag{9}$$

Fusion Algorithm

In the fusion algorithm proposed in this paper, the Kalman filter is used to track the targets detected by radar and camera to reduce the measurement noise. After that, fusion weight is calculated based on Bayesian Estimation. Finally, the targets data are fused according to fusion weight and tracking results.

Bayesian Estimation

Bayesian Estimation is a data fusion algorithm to estimate the unknown state vector X by the known measurement vector Z.

Bayes' theorem is given by:

$$p\left(X = x \mid Z = z\right) = \frac{p\left(Z = z \mid X = x\right) p\left(X = x\right)}{p\left(Z = z\right)} \tag{10}$$

An estimation of X can be made by maximizing this posterior distribution, i.e., by maximizing the $p(X = x| Z = z)$, which is called the Maximum a posteriori (MAP) estimate [9]. Since the denominator in the formula (10) is a normalization factor. The problem is equal to maximize the numerator:

$$\hat{x}_{MAP} = \arg\max p\left(X = x \mid Z = z\right) \propto p\left(Z = z \mid X = x\right) p\left(X = x\right) \tag{11}$$

In the case of two-sensors model, the formula (11) could be extended as:

$$p\left(X = x \mid Z = z_1, z_2\right) = \frac{p\left(Z = z_1 \mid X = x\right) p\left(Z = z_2 \mid X = x\right) p\left(X = x\right)}{p\left(Z = z_1, z_2\right)} \tag{12}$$

Suppose the measurement uncertainties of the two sensors could be represented by Gaussian distribution:

$$p\left(Z = z_j \mid X = x\right) = \frac{1}{\sigma_j \sqrt{2\pi}} \exp\left\{ \frac{-\left(x - z_j\right)^2}{2\sigma_j^2} \right\} j = 1,2 \tag{13}$$

the fused MAP estimate is given by:

$$\hat{x}_{MAP} = \arg\max \left[p\left(Z = z_1 \mid X = x\right) p\left(Z = z_2 \mid X = x\right) \right] \tag{14}$$

$$\hat{x}_{MAP} = \arg\max\left[\frac{1}{\sigma_1\sqrt{2\pi}}\exp\left\{\frac{-\left(x-z_1\right)^2}{2\sigma_1^2}\right\} \times \frac{1}{\sigma_2\sqrt{2\pi}}\exp\left\{\frac{-\left(x-z_2\right)^2}{2\sigma_2^2}\right\}\right] \tag{15}$$

And the fusion result is given by:

$$x_f = \hat{x}_{MAP} = \frac{\sigma_2^2}{\sigma_1^2+\sigma_2^2}z_1 + \frac{\sigma_1^2}{\sigma_1^2+\sigma_2^2}z_2 \tag{16}$$

σ_1 and σ_2 are the measurement standard deviation of sensors.

Kalman Filter

The target dynamic model is given by:

$$\hat{x}_k = \phi_{k-1}\hat{x}_{k-1} + \beta_{k-1}u_{k-1} + w_{k-1} \quad w_k = N\left(0,Q_k\right) \tag{17}$$

x_k is the target state vector, $\hat{x}_k = \left[x \;\; v_x \;\; y \;\; v_y\right]^T$, the x, v_x, y, v_y are the target's longitudinal distance, longitudinal velocity, lateral distance and lateral velocity respectively. ϕ_{k-1} is the state transition matrix, u_{k-1} is the controlled input, w_{k-1} is the process noise and Q_k is process noise covariance.

The measurement model is given by:

$$z_k = H_k x_k + v_k \quad v_k = N\left(0, R_k\right) \tag{18}$$

z_k is the target measurement vector, H_k is the measurement matrix,v_k is the measurement noise and R_k is the measurement noise covariance. Since the radar could measure target distance and velocity while camera could only measure target distance, the H_k and v_k for radar and camera are different. Suppose sensor 1 represents radar and sensor 2 represents camera.

$$z_{k,1} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} x_k + v_{k,1} \; v_{k,1} = N\left(0,R_{k,1}\right) \tag{19}$$

$$z_{k,2} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} x_k + v_{k,2} \; v_{k,2} = N\left(0,R_{k,2}\right) \tag{20}$$

Since the measurement models of radar and camera are different, the parallel Kalman filter is used to track the targets detected by radar and camera respectively. The Kalman filter involves five basic equations.

For j = 1,2 represent radar and camera respectively.

Predictor Equation:

$$\hat{x}_{k|k-1,j} = \phi\hat{x}_{k-1|k-1,j} + \beta_{k-1}u_{k-1} \tag{21}$$

Which predicts state vector $\hat{x}_{k|k-1}$ at time k based on estimated target state $\hat{x}_{k-1|k-1}$ and state transition matrix ϕ.

Error covariance Equation:

$$P_{k|k-1,j} = \phi P_{k-1|k-1,j}\phi^T + Q \tag{22}$$

Which calculates the covariance matrix of the state vector $\hat{x}_{k-1|k-1}$.

Weight Equation:

$$K_{k,j} = P_{k|k-1,j}H_j^T / \left(H_j P_{k|k-1,j}H_j^T + R_j\right) \tag{23}$$

The weight is calculated based on covariance matrix of the state vector $P_{k|k-1}$ and measurement noise covariance R.

Filtering Equation:

$$\hat{x}_{k|k,j} = \hat{x}_{k|k-1,j} + K_{k,j}\left(\hat{z}_{k,j} - H_j\hat{x}_{k|k-1,j}\right) \tag{24}$$

the state vector $\hat{x}_{k|k}$ at time k is calculated by a weighted sum the predicted state vector $\hat{x}_{k,k-1}$ and measurement vector $\hat{z}_k$. The weight is calculated by formula (23)

Update error covariance Equation:

$$P_{k|k,j} = P_{k|k-1,j}\left(I - K_{k,j}H_j\right) \tag{25}$$

According to formula (16), the final fusion result is given by:

$$\hat{x}_{k|k,f} = \frac{\sigma_2^2}{\sigma_1^2+\sigma_2^2}\hat{x}_{k|k,1} + \frac{\sigma_1^2}{\sigma_1^2+\sigma_2^2}\hat{x}_{k|k,2} \tag{26}$$

σ_1 and σ_2 are obtained from the error covariance matrix of the state vector $\hat{x}_{k|k}$.

Simulation Experiment

In the simulation experiment, the type of target is pedestrian and vehicle. The target's motion model is CV model. Since the radar could measure target distance and velocity while camera could only measure target distance, the target's distances in longitudinal and lateral direction are the fused data object.

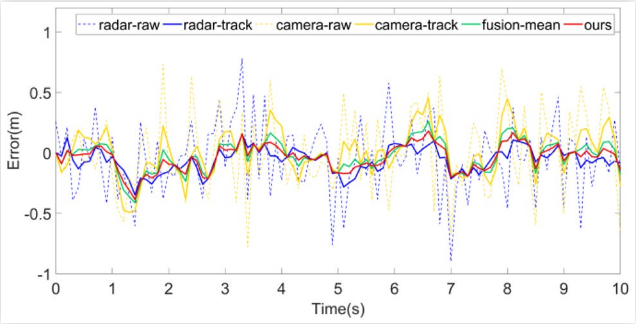
In simulation experiment, radar model is a 77GHz radar, FOV is 18°, measurement distance is 150 m, sampling period is 100 ms. The camera model is a monocular camera, FOV is 40°, measurement distance is 120 m, sampling period is 100 ms. The measurement noise of sensors meet Gaussian distribution.

The measurement standard deviation of sensors is designed according to the sensor characteristics of radar and camera. Since the lateral resolution of radar is relatively low, the measurement of longitudinal distance is more accurate than lateral distance. The standard deviation of longitudinal distance is smaller than lateral distance. For camera, the measurement accuracy is lower than radar but in the same level. The camera measurement accuracy in longitudinal and lateral direction are almost equal. The relationship between the sensors measurement standard deviation in simulation experiments is set as:

$$\sigma_{camera\_longitudinal} = \sigma_{camera\_lateral} > \sigma_{radar\_lateral} > \sigma_{radar\_longitudinal} \tag{27}$$

According to the type of target, the measurement distance accuracy for vehicle target is higher than the counterpart of pedestrian. Consequently, the standard deviation for vehicle targets is smaller than pedestrian target. Three experiments were designed to test and evaluate the fusion algorithm.

FIGURE 3 The error of raw measurement results, tracking results and fusion results in the longitudinal direction.



Experiment 1: Pedestrian Moves in Longitudinal Direction

The target type was pedestrian. And the target's initial longitudinal distance, lateral distance, longitudinal velocity and lateral velocity were set to 10 meter, 0 meters, 5 km/h and 0 km/h respectively. The duration time of experiment was 10 seconds. The experiment result is shown in Figure 3.

The Figure 3 shows that Kalman filter could reduce the measurement noise. Consequently, the tracked sensor data were used to fuse.

The pedestrian's longitudinal velocity increased from 5 km/h to 10 km/h progressively. The residual sum of squares(RSS), which represents the sum of error squares between ideal and estimated value in every sampling point, is used for evaluation. The formula of RSS is given by:

$$RSS = \sum_{i=1}^{n} \left(x_{ideal,i} - x_{estimated,i} \right)^2 \tag{28}$$

According to Figure 4, Figure 5 and Table 1:

The RSS mean value and standard deviation of two fusion algorithms are better than solo-sensor tracking results in both longitudinal and lateral direction. The data fusion algorithm could improve the estimation accuracy effectively. Since the radar measurement accuracy in longitudinal is better, the RSS mean value of radar tracking result is far less than the camera's. Thus, the proposed algorithm takes the advantage of radar and assigns more fusion weight to radar. Its RSS mean value is 12.7% lower than the equal weight fusion algorithm, and their RSS standard deviations are almost equal. Consequently, it outperforms the equal weight fusion algorithm in longitudinal direction. In the lateral direction, the RSS mean value of radar tracking result increased but is still better than the camera's. Thus, compared with the longitudinal direction, radar is assigned less fusion weight. And the RSS mean value of proposed algorithm is almost equal to equal weight fusion algorithm. Although its RSS standard deviation is 8.6% more than the equal weight fusion algorithm, the base is small and the indicator is secondary indicator compared to the mean value indicator. Consequently, the performance of proposed algorithm is similar to equal weight fusion algorithm in lateral direction.

FIGURE 4 The RSS in the longitudinal direction in different velocity.

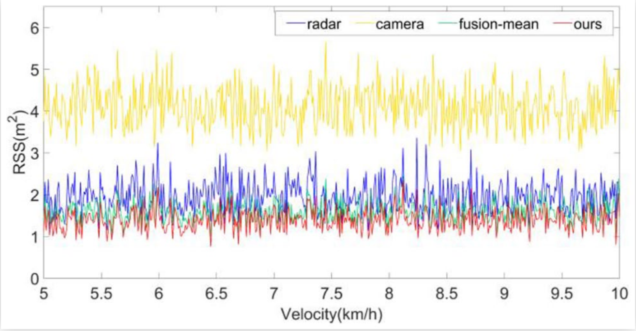


FIGURE 5 The RSS in the lateral direction in different velocity.

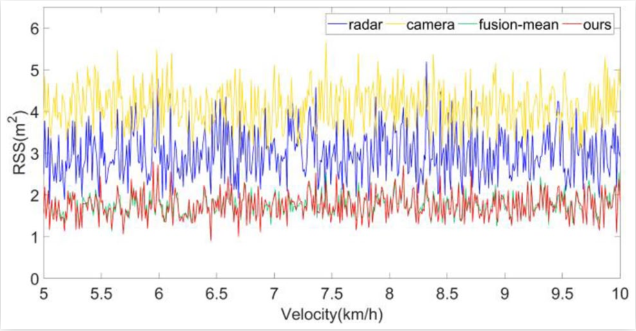


TABLE 1 The mean value and standard deviation of the RSS in experiment 1. The x- represents the longitudinal direction; the y- represents the lateral direction.

	radar	camera	fusion-mean	ours
x-mean(m^2)	1.9540	4.1123	1.5894	1.3877
x-standard deviation(m^2)	0.3841	0.4981	0.2476	0.2522
y-mean(m^2)	2.9877	4.1123	1.7619	1.7439
y-standard deviation(m^2)	0.5785	0.4970	0.2784	0.3045

Experiment 2: Vehicle Moves in Longitudinal Direction

The target type was vehicle. And the target's initial longitudinal distance, lateral distance, longitudinal velocity and lateral velocity were set to 10 meter, 0 meters, 20 km/h and 0 km/h respectively. The vehicle's longitudinal velocity increased from 20 km/h to 45 km/h progressively.

According to Figure 6, Figure 7 and Table 2:

The overall RSS mean value and standard deviation decrease compared to experiment 1, since the measurement accuracy for vehicle target is higher than the pedestrian target. In details, in the longitudinal direction, the RSS mean value of proposed algorithm is 11.0% less than the equal weight fusion algorithm, which approves its better performance. In the lateral direction, their performances are similar.

FIGURE 6 The RSS in the longitudinal direction in different velocity.

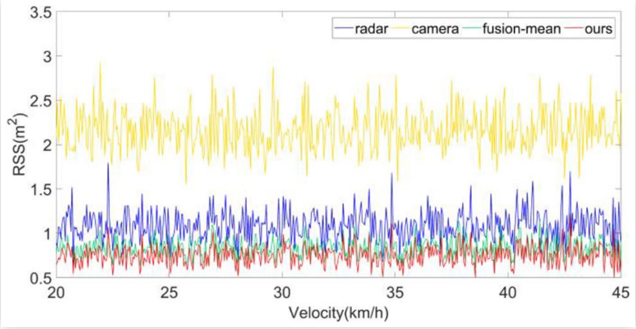


FIGURE 7 The RSS in the lateral direction in different velocity.

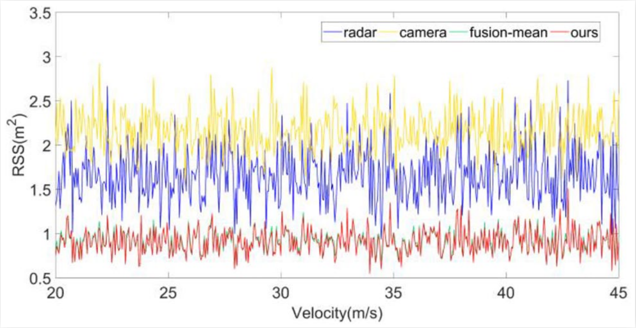


TABLE 2 The mean value and standard deviation of the RSS in experiment 2. The x- represents the longitudinal direction; the y- represents the lateral direction.

	radar	camera	fusion-mean	ours
x-mean(m^2)	1.0842	2.1630	0.8372	0.7461
x-standard deviation(m^2)	0.1674	0.2299	0.1048	0.1080
y-mean(m^2)	1.6645	2.1630	0.9263	0.9129
y-standard deviation(m^2)	0.2962	0.2299	0.1287	0.1395

Experiment 3: Pedestrian Moves in Lateral Direction

The target type was pedestrian. And the target's initial longitudinal distance, lateral distance, longitudinal velocity and lateral velocity were set to 10 meter, −5 meters, 0 km/h and 5 km/h respectively. Then the pedestrian's lateral velocity increased from 5 km/h to10 km/h progressively.

According to Figure 8, Figure 9 and Table 3, in the longitudinal direction, the RSS mean value of proposed algorithm is 12.8% less than the equal weight fusion algorithm; in the lateral direction, their performances are similar. The conclusions about proposed fusion algorithm are consistent with previous experiments.

FIGURE 8 The RSS in the longitudinal direction in different velocity.

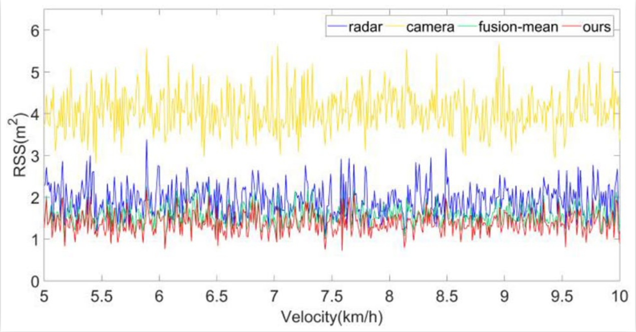


FIGURE 9 The RSS in the lateral direction in different velocity.

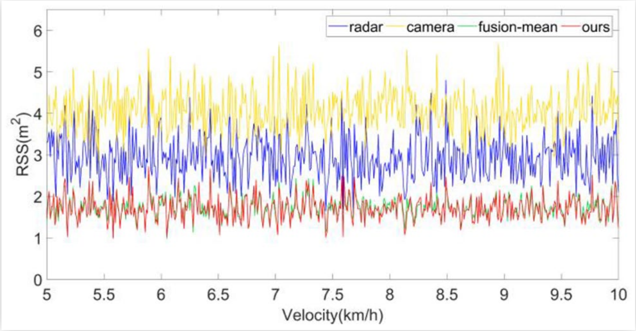


TABLE 3 The mean value and standard deviation of the RSS in experiment 3. The x- represents the longitudinal direction; the y- represents the lateral direction.

	radar	camera	fusion-mean	ours
x-mean(m^2)	1.9381	4.0730	1.5757	1.3762
x-standard deviation(m^2)	0.3803	0.5005	0.2382	0.2446
y-mean(m^2)	2.9428	4.0730	1.7344	1.7144
y-standard deviation(m^2)	0.5370	0.5005	0.2633	0.2873

In summary, the proposed algorithm is the best estimation algorithm and tolerate to the velocity variation.

Real Experiment

In the real experiment, the used radar was 77GHz continental ARS 408 radar with 18° FOV, 250 m measuring range, 0.2 m range resolution and 80 ms sampling period. And the used camera's is a monocular camera with 40° FOV and 1280x720p resolution ratio. The object detection algorithm, SSD is operated in TX2 with about 11fps. Since the radar's sampling period is different from camera. The spline-based fitting interpolation method is used for time synchronization.

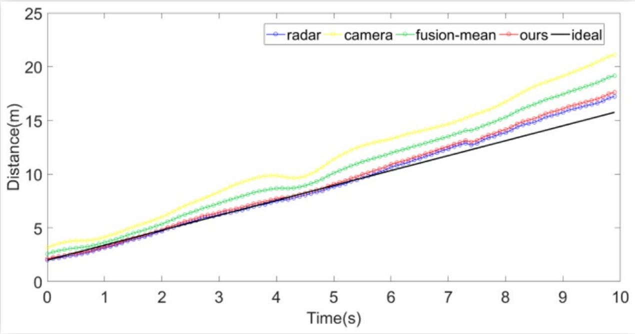
Experiment 4: Pedestrian Moves in Longitudinal Direction

The pedestrian's initial longitudinal distance and lateral distance were 2 meters and 0 meter. And the pedestrian moved away with about 5 km/h longitudinal velocity.

According to Figure 10, Figure 11 and Table 4:

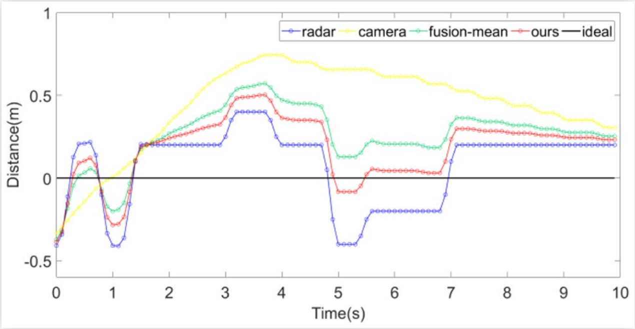
In longitudinal direction, the RSS of proposed fusion algorithm is worse than the radar tracking results, which is different from the conclusions in simulation experiment. The proposed fusion algorithm works better when the measurement accuracies of different sensors are in the same level. However, in the real experiment, camera measurement accuracy is not good as the radar does in longitudinal direction. Consequently, the addition of camera tracking results with large error would reduce the fusion algorithm performance on the contrary. But the proposed algorithm assigns more fusion weight to radar and the performance degradation is acceptable.

FIGURE 10 The tracking results and fusion results in the longitudinal direction. Suppose target moved at constant velocity in the experiment, the black line represents the ideal target longitudinal distance.



© SAE International

FIGURE 11 The tracking results and fusion results in the lateral direction



© SAE International

TABLE 4 The RSS in experiment 4. The x- represents the longitudinal direction; the y- represents the lateral direction.

	radar	camera	fusion-mean	ours
x-RSS(m^2)	36.5372	843.6917	289.2163	62.9294
y-RSS(m^2)	5.9289	8.4778	4.8071	4.6807

© SAE International

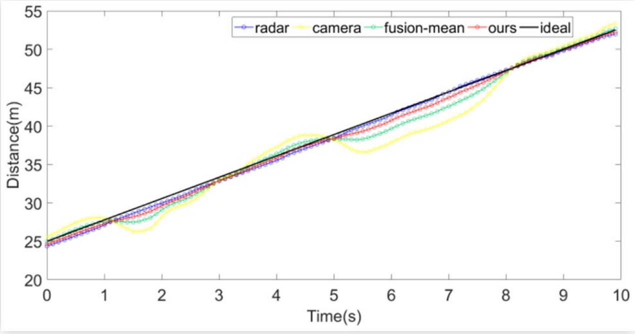
In lateral direction, the radar measurement accuracy is similar to camera. Consequently, the conclusion about proposed fusion algorithm are consistent with simulation experiments. Two fusion algorithms improve the accuracy and are better than solo-sensor tracking results. The proposed algorithm is slightly better than equal weight fusion algorithm.

Experiment 5: Vehicle Moves in Longitudinal Direction

The vehicle's initial longitudinal distance and lateral distance were 25 meters and 0 meter. And vehicle moved away with about 10 km/h longitudinal velocity.

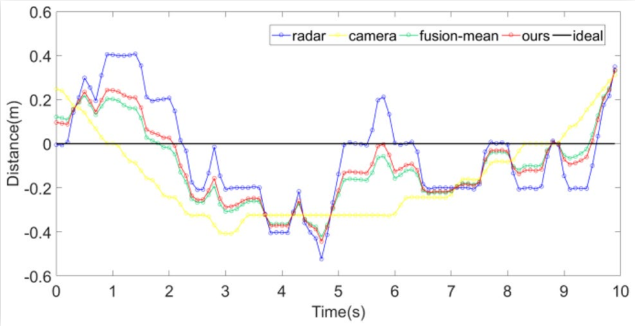
According to Figure 12, Figure 13 and Table 5, the overall RSS value decrease compared to experiment 4, since the measurement accuracy for vehicle target is higher than the pedestrian target. In longitudinal direction, the proposed algorithm takes the advantage of radar and performs satisfactory accuracy. In the lateral direction, the proposed algorithm improves the estimation accuracy and is better than equal weight fusion algorithm.

FIGURE 12 The tracking results and fusion results in the longitudinal direction Suppose target moved at constant velocity in the experiment, the black line represents the ideal target longitudinal distance.



© SAE International

FIGURE 13 The tracking results and fusion results in the lateral direction



© SAE International

TABLE 5 The RSS in experiment 5. The x- represents the longitudinal direction; the y- represents the lateral direction.

	radar	camera	fusion-mean	ours
x-RSS(m^2)	17.7746	443.0017	123.6233	34.6460
y-RSS(m^2)	5.1947	6.0827	4.1266	4.0983

© SAE International

Conclusions

In this paper, a method of vehicle and pedestrian detection based on the data fusion of millimeter wave radar and camera is proposed. A deep learning model called Single Shot MultiBox Detector (SSD) is utilized for vison-based targets detection. The parallel Kalman filter is used to track the targets detected by radar and camera. Finally, the targets data after tracking are fused based on Bayesian Estimation.

Through simulation and real experiments, it shows that the proposed algorithm could improve the estimation accuracy effectively and outperforms single sensor tracking result and equal weight fusion algorithm.

References

1. Wang, X., Xu, L., Sun, H. et al., “Bionic Vision Inspired On-Road Obstacle Detection and Tracking Using Radar and Visual Information,” *IEEE International Conference on Intelligent Transportation Systems*, 2014, 39-44. IEEE.
2. Han, S, Wang, X, Xu, L. et al., “Frontal Object Perception for Intelligent Vehicles Based on Radar and Camera Fusion,” *Control Conference*, 2016, 4003-4008. *IEEE*.
3. Alencar, F.A.R., Rosero, L.A., Massera Filho, C. et al., “Fast Metric Tracking by Detection System: Radar Blob and Camera Fusion,” *Robotics Symposium (LARS) and 2015 3rd Brazilian Symposium on Robotics (LARS-SBR), 2015 12th Latin American*, 2015,120-125. *IEEE*.
4. Bi, X., Tan, B., Xu, Z. et al., “A New Method of Target Detection Based on Autonomous Radar and Camera Data Fusion,” *Intelligent and Connected Vehicles Symposium*, 2017.
5. Liu, W., Anguelov, D., Erhan, D. et al., “SSD: Single Shot MultiBox Detector,” *European Conference on Computer Vision*, 2016, 21-37. Springer, Cham.
6. Geiger, A., Lenz, P., Urtasun, R., “Are We Ready for Autonomous Driving the Kitti Vision Benchmark Suite,” *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, 2012, 3354-3361. IEEE.
7. Arróspide, J., Salgado, L., and Nieto, M., “Video Analysis Based Vehicle Detection and Tracking Using an MCMC Sampling Framework,” *EURASIP Journal on Advances in Signal Processing*, 2012, Article ID 2012:2, Jan. 2012, doi:10.1186/1687-6180-2012-2.
8. Gonzalez, A., Fang, Z., Socarras, Y., Serrat, J., Vazquez, D., Xu, J., and Lopez, A., “Pedestrian Detection at Day/Night Time with Visible and FIR Cameras: A Comparison,” *In Sensors Journal (Sensors)*, in press, 2016.
9. Kumar, M., Garg, D.P., Zachery, R., “A Generalized Approach for Inconsistency Detection in Data Fusion from Multiple Sensors, *American Control Conference*, 2006, 6pp. IEEE Xplore.

Contact Information

Zhexiang Yu
Tongji University No.4800 Cao’An Road, Jiading District, Shanghai
Republic of China
yuzhexiang@tongji.edu.cn

Acknowledgments

Subsidized by the project of standardization and new model for intelligent manufacture: Research and Test Platform of System and Communication Standardization for Intelligent and Connected Vehicle, project ID: 2016ZXFB06002.

Definitions/Abbreviations

ADAS - Advanced Driving Assistance System
SSD - Single Shot MultiBox Detector
FOV - field of views
ROI - Region Of Interest
CNN - Convolutional Neural Network
YOLO - You Only Look Once
MAP - Maximum A Posteriori
RSS - Residual Sum of Squares

Positions and opinions advanced in this paper are those of the author(s) and not necessarily those of SAE International. The author is solely responsible for the content of the paper.

ISSN 0148-7191