IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received August 16, 2019, accepted September 3, 2019, date of publication September 19, 2019, date of current version October 2, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2942382

Extending Reliability of mmWave Radar Tracking and Detection via Fusion With Camera

RENYUAN ZHANG, (Student Member, IEEE), AND SIYANG CAO, (Member, IEEE)
Department of Electrical and Computer Engineering, The University of Arizona, Tucson, AZ 85721, USA

Corresponding author: Siyang Cao (caos@email.arizona.edu)

This work was supported by the University of Arizona.

ABSTRACT In this paper, a new radar-camera fusion system is presented. The fusion system takes into consideration the error bounds of the two different coordinate systems from the heterogeneous sensors, and further a new fusion-extended Kalman filter is utilized to adapt to the heterogeneous sensors. Real-world application considerations such as asynchronous sensors, multi-target tracking and association are also studied and illustrated in this paper. Experimental results demonstrated that the proposed fusion system can realize a range accuracy of 0.29m with an angular accuracy of 0.013rad in real-time. Therefore, the proposed fusion system is effective, reliable and computationally efficient for real-time kinematic fusion applications.

INDEX TERMS Millimeter wave radar, Kalman filtering, error bounds, multisensor systems, sensor fusion, fusion-EKF, homography estimation, multi-target tracking.

I. INTRODUCTION

Radar, working at W-band, is becoming an important sensor in advanced driver assistance system (ADAS) and autonomous driving fields [1]–[5]. Ongoing developments in research and industry primarily focus on safety, reliability, compact and low-cost sensor systems. ADAS systems comprise of multiple sensors, such as radar, camera and Lidar, for specific applications commensurate with the sensor's offered advantage. The latest millimeter wave (mmWave) radars at W-band are surging for automobile applications, *e.g.*, adaptive cruise control, pedestrian detection, collision avoidance, lane changing monitoring and emergency braking. mmWave radar research has gained immense popularity with the recent increasing demand of automotive radars in ADAS and autonomous driving industry, for instance, the automotive target shape (height and width) estimation using relaxation algorithm [6]; the super-resolution automotive radars [7]; and the automotive radar sensor fusion with other sensors to improve target tracking and classification [4]. To take advantage of other sensors for sensing accuracy and reliability, we are focusing on the mmWave radar sensor fusing with a camera sensor in this paper.

Sensors' reliability, especially for avoiding false detection, missing detection, blocking, blind spot, adverse weather and failure, is essential in ADAS and autonomous driving applications. One way to reduce all uncertainties and failures is to fuse outputs from multiple sensors [8]. Sensor fusion has been developing rapidly in recent years. However, there are limited studies around fusing mmWave radars with other sensors due to the fact that radars provide a limited number of detection points representing targets-of-interest [4], which make it difficult to recognize from a snapshot of radar detection. But if this challenge of the mmWave radar data can be solved, radar can be used to further increase the reliability of detecting moving targets, avoiding blockage and tracking dramatically. In this paper, we aim to increase mmWave radar's informative capability of targets and improve its versatility by fusion with monocameras.

In this paper, we present a new fusion-extended Kalman filter (fusion-EKF), which is designed to fuse data from heterogeneous sensors such as mmWave radar and monocamera with real-time fusion algorithm running for tracking. Sensor fusion and association are done within the fusion-EKF using a homography estimation method (HEM) [9], timeline alignment and region search. Reliable detection and cross-validated target tracking are realized. As we have not used a machine learning approach to achieve this, we achieve a significant reduction in computational complexity. Our experimental result shows that the proposed system can provide a reliable tracking and detecting result with low calculation costs. An embedded system like Arduino or Raspberry Pi can be utilized to process the data for real-time applications.

For the new fusion-EKF, we introduced a new concept, *i.e.* error bounds (EBs), which is defined as the sensor's region of approximation. EBs are not from the uncertainty of

The associate editor coordinating the review of this manuscript and approving it for publication was Pietro Savazzi.

sensors [10] but the sensors' resolutions from their respective perspectives. An HEM is applied in associating heterogeneous sensors via their EBs. The fusion-EKF is designed to take both radar and camera as inputs and associate the data inside the filter to obtain ideal target tracking outputs. Data association of the fusion-EKF is capable to support tracking of multiple targets.

In summary, our contributions are:

1) Building a new real-time sensor fusion system for mmWave radar and camera.
2) Designing a new fusion-EKF to support the two heterogeneous sensors.
3) Defining a new concept, EBs, for data association and gating inside the fusion-EKF.
4) An HEM is proposed for building the transformation matrix between mmWave radar and cameras.

The structure of this paper is as follows. In Section II, related work on mmWave radar sensor fusion and techniques of data fusion are studied and presented. In Section III, the proposed methodology is presented including radar and camera data preprocessing, radar-camera fusion-EKF and sensor synchronization. In Section IV, results of radar-camera fusion-EKF and root mean square errors (RMSEs) of improved EBs are shown. In Section V, we conclude the paper.

II. RELATED WORK

mmWave radars have been implemented to fuse with Lidars [11], [12] in recent years. In [11], Hajri and Rahal introduced radar and Lidar real-time sensor fusion that used the mean square error with radar, Lidar, and the fusion variance in the experiment, to yield convincing results. By default, radar and Lidar, both can provide detections in the birds-eye-view (BEV) plane which is the top view looking onto the ground [13]. However, Lidar works on the infrared band, which is (i) prone to interference, (ii) cost-intensive and (iii) bulky in size. Not many studies have been conducted with associating heterogeneous sensors like cameras or infrared sensors with radars [14], [15], primarily due to the fact that the camera's and radar's detections are unrelated and difficult to align without certain assumptions. In this case, we are implementing HEM and creating a track-oriented association and fusion algorithm for calibration and tracking.

Alternate approaches to sensor fusion have been studied using machine learning methods. The sensor fusion is typically performed via concatenating tensors inside the neural network [16]. However, it requires large neural networks for high resolution radar and cameras, which makes it difficult to implement for real-time applications. Similarly, in [17], Zhong *et al*. introduced the Texas Instruments' (TI) commercial TDA3x/TDA2x boards for camera radar fusion. It is so far the most reliable fusion system for the mmWave radar's detection compounded with the camera's vision to the best of our knowledge. In their paper, a complete fusion system provides effective tracking and detections over time. The shortcomings of this research are that, the object detections and classifications rely on machine learning techniques with online computer vision libraries. Therefore, the complete pipeline is difficult to implement for real-time applications. In the meantime, radar noise handling is not ideal for various scenarios. In our work, we strongly overcome these shortcomings.

In [18], Muresan and Nedevschi presents a two-step data association method for Lidar. It is based on spreading parameter weights from KF. In this paper, the weights are further studied to support both the heterogeneous sensors. Specifically, the EBs are introduced in the data association phase as radar provides raw Doppler velocities and camera vision generates the bounding boxes of targets. EB supports the two sensors to form a clear gating for the association of the fusion-EKF. Thus comparing to other weight-based or probability-based association methods, the proposed fusion-EKF with EBs will provide an extra dimension in association for heterogeneous sensors. Motion model selection for multi-target tracking can also be an important part of KF (Kalman Filter) tracking. In [19], the survey on different motion models is introduced and compared. In our setup, the constant acceleration (CA) model is used because we assumed targets can have the linearity of the state transition equation which allows an optimal propagation of the state probability distribution of humans and cars, say. Yaw angle and turn rate are less essential in these models. Thus it allows tracks to have more flexibility in association from heterogeneous sensors.

Some papers [14], [15], [20], [21] illustrated the alignment between radars and other heterogeneous sensors. They assumed radars and other sensors are working at identical principle. But in fact, they are heterogeneous in dimensions, processing chain and sensing wavelength. In [20], vision lateral position improvement of radar detections are introduced that worked for radar alignment onto the vision well using cross-correlation. However, the errors are not corrected during detections. In [21], Folster and Rohling introduced a lateral velocity estimation for automotive radar sensors for ADAS. The method simply refers to the location of the target on the actual detections. The disadvantages are obvious: there are not many cross-correlations between sensors, and the approximation relies only on locations. To overcome limitations in these cases, EBs proposed in this paper are essential to apply to radar-vision fusion-EKF to improve region search and tracking. The HEM further enhances the transformation and track-oriented association between heterogeneous sensors, which makes the proposed fusion system to takes both sensors' advantages, and be robust even one sensor is temporarily out of work.

The fusion-EKF is designed based on the Kalman filter (KF). KF typically uses a series of measurements observed over time, containing statistical noise and other inaccuracies, and produces estimates of unknown variables that tend to be more accurate than those based on a single measurement alone [22]. The difference between the EKF [23] and traditional KF is that EKF can handle linear

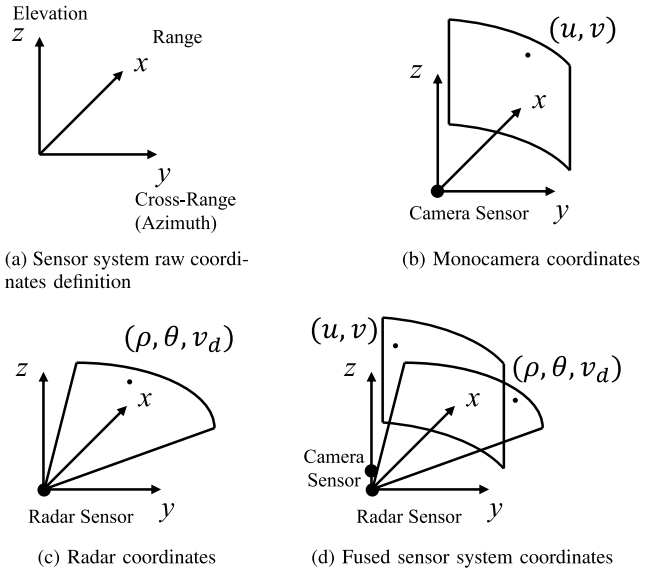
equations as well as can adapt to non-linearities. EKF uses the Jacobian matrix to linearize non-linear functions, for instance, radar's Doppler velocities (range rates) in our case. Fusion-EKF extends EKF's ability in handling data from heterogeneous sensors. In this paper, we are not considering unscented Kalman filter (UKF) because radar's non-linearity comes from Doppler, which is the first-order approximation of range. Thus, UKF can be unstable and its results will be biased when estimating radar's Doppler velocities.

III. METHODOLOGY

In this paper, a new fusion system for mmWave radar and camera is proposed. We will first introduce the EBs in Section III-A. It is a fundamental concept that is utilized in the fusion-EKF system for data association. Then, we will discuss the preprocessing of the two sensors and the HEM method for finding out the transformation matrix between the two sensors in Sections III-B and III-C. The new fusion-EKF will be discussed in Section III-D. The data association and synchronization will be introduced in Section III-E. The fusion system performance will be evaluated based on RMSE as discussed in Section III-F.

A. COORDINATES AND ERROR BOUNDS

In this sensor fusion system, sensors operating in two different coordinate systems are used, *viz.*, monocamera and mmWave radar. In Fig. 1, the coordinates in this fusion system are presented.



(a) Sensor system raw coordinates definition

(b) Monocamera coordinates

(c) Radar coordinates

(d) Fused sensor system coordinates

FIGURE 1. Coordinates definition.

With a monocamera and linear phased array mmWave radar, the fusion can improve the detection to a 3D space. For the fused sensor system, we use a left-hand coordinate system, shown in Fig. 1a. The x-axis is the range axis which is radially away from the sensor center along the line of sight. y-axis is the cross-range axis, *a.k.a.*, azimuth axis. z-axis is the elevation axis. In Fig. 1b, monocamera vision plane with monocamera sensor placed at zero is presented. (u, v) pixel variables at camera plane are used with u representing horizontal pixel and v representing vertical pixel, respectively. In Fig. 1c, The mmWave radar sensor is conducting measurements with (ρ, θ, v_d), which are radar measurement range, azimuth angle and Doppler velocity (*a.k.a.* range rate), respectively.

If both sensors are presented, with a slight vertical difference in the positions of the camera and radar, shown in Fig. 1d, a fused system reconstructs the coordinate system with planes of vision and mmWave radar detections perpendicular to each other. As vision data is not totally limited to one plane, blockage of some targets could exist on different range and azimuth locations. Thus, the vision plane is penetrating throughout range and azimuth. In this case, the fusion can be constructed with correlations of (u, v) and (ρ, θ, v_d). Additionally, Cartesian and polar conversions are made throughout the fusion process. Radar's angular θ-angle axis and azimuth axis are used in following statements.

Error bounds of heterogeneous sensors have different coordinate systems as stated above. Here we need to define how the fused system will behave compared to before and after the sensor fusion.

Definition 1: Measurement point of the next inquiry lays within a certain region of approximation, from either radar or camera sensors. That region is then the theoretical error bounds of the sensor.

Error bounds of the sensor fusion system is defined as following:

Definition 2: The actual localization of the detecting target lays within the approximate prediction region. The approximation is estimated by two or more sensors' system, and the region takes advantages from the sensors' error bounds. The prediction region is then the fused error bounds from the multi-sensor system.

With Definitions 1 and 2, the error bounds of fused sensor system can be plotted as Fig. 2. For vision error, the detecting target is difficult to solve along the range axis (depth). For radar error, the detecting target is limited as solving along θ-angle axis, which is caused by 4-by-2 multiple-input and multiple-output (MIMO) transceivers configuration (15 degrees θ resolution at center). Fusion takes advantages

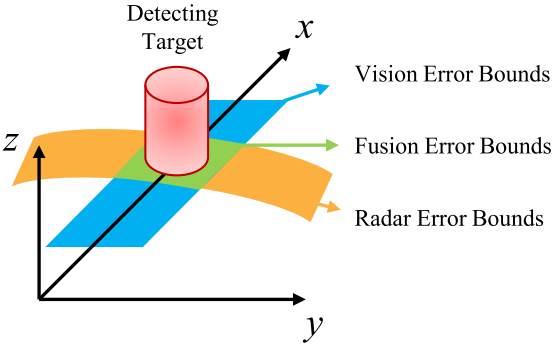


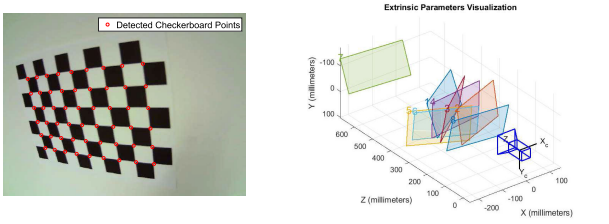
FIGURE 2. Error bounds of fused sensor system.

from both the sensors, which results in better range and angle resolutions. However, in this application, we do not solve for the elevation (height) due to the constraint of having a 1D antenna array from the mmWave radar sensor. With better mmWave radars with antenna channels along the elevation, a similar approach can be used to obtain the elevation error bounds.

B. CAMERA AND mmWave RADAR PREPROCESSING

Cameras, including stereo and monocameras, have intrinsic and extrinsic parameters which are inherent due to the physical build of the pinhole and lens [24]–[27]. The extrinsic parameters with rotation and translation matrices are intended to map the world coordinates of the objects into camera coordinates. On the other hand, intrinsic parameters deal with the camera coordinates into the image pixel coordinates (u, v).

In this paper, we generally follow the four-step procedure from [27] and calibrate the monocular camera on the robotic operating system (ROS). In this case, intrinsic, extrinsic and lens distortion parameters of the camera are prepared for warping and distortion image processing [28]. The camera calibration for this fusion project is shown in Fig. 3. Fig. 3a shows the sample of detected checkerboard points in image coordinates. With help from the extrinsic visualization in Fig. 3b, multiple samples are collected and used to estimate intrinsic parameters.



(a) Checkerboard calibration for intrinsic parameters (b) Extrinsic parameters visualization

FIGURE 3. Camera calibration.

For mmWave radars preprocessing, the radar sensor configuration determines its range resolution ($\Delta\rho$), bearing resolution ($\Delta\theta$) and Doppler resolution (Δv_d).

Range resolution $\Delta\rho$ can be calculated from radar sweeping chirp bandwidth:

$$\Delta\rho = \frac{c_0}{2B}, \tag{1}$$

where

c_0 : the speed of electromagnetic waves
B : the bandwidth of the radar chirp signal.

The bearing resolution physically depends on the number of virtual antennas MIMO radar emulated. The angular resolution is given as:

$$\Delta\theta = \frac{c_0}{f_c d N_{\mathrm{RX}} N_{\mathrm{TX}} \cos\theta_i}, \tag{2}$$

where

f_c : center frequency of linear frequency modulated signal
d : receiver element spacing, typically $\lambda/2$
λ : carrier signal wavelength
N_{RX} : number of receivers (RX)
N_{TX} : number of transmitters (TX)
θ_i : the angle of interest.

Doppler resolution provides the target range changing rate and is defined as:

$$\Delta v_d = \frac{c_0}{2f_c(\mathrm{PRI})N_d}, \tag{3}$$

where

PRI : the pulse repetition interval
N_d : number of chirps per radar frame.

PRI is commonly the total chirp time of a single frequency modulated signal. Thus Δv_d is typically restricted by number of chirps and chirp repetition interval.

An important part for mmWave radar in fusion-EKF is that mmWave radar measures targets' scatters with compound noises. These noises come from the following effects:

1) Multipath effect [29], that results in detecting targets to undergo ghost or fading where a phase-shifted reflection is produced by walls or ground.
2) mmWave radar has its inherent measurement noise ($\mathbf{R}_{\mathrm{radar}}$). In this case, we compute the noise model by calculating the variance of the point cloud and signal-to-noise ratio (SNR) of actual measurement based on the set radar configuration.
3) Measurement processing noise from stretch processing, and FFTs produces noisy point cloud points. The noise is typically dense when multiple targets are being detected. At the meantime, a higher processing rate causes the processing noise to lift further.
4) Antenna gain contributes to a varied SNR at different angles. This may cause the SNR at the center (line of sight) to be higher than SNR at the sides.

For the first case, which is the multipath effect in 77 GHz automotive radar, a typical symptom is the presence of ghost targets, that are produced with random phase but similar shape, Doppler and amplitude compared to the original target. A method to reduce when applied to tracking is to use a multilayer perceptron [30]. In this case, the ghost is largely alleviated and then the tracking becoming more reliable.

For the second case, $\mathbf{R}_{\mathrm{radar}}$ obeys the Gaussian distribution. $\mathbf{R}_{\mathrm{radar}}$ is then

$$\mathbf{R}_{\mathrm{radar}} \sim \mathcal{N}\left(\begin{pmatrix}\mu_\rho \\ \mu_\theta \\ \mu_{v_d}\end{pmatrix}, \begin{pmatrix}\sigma_\rho^2 \\ \sigma_\theta^2 \\ \sigma_{v_d}^2\end{pmatrix}\right). \tag{4}$$

The $\begin{pmatrix}\mu_\rho & \mu_\theta & \mu_{v_d}\end{pmatrix}^{\mathsf{T}}$ matrix, corresponds to the targets' detection matrix. It specifies target's range, bearing and Doppler

velocity. On the other hand, the $\left(\sigma_\rho^2\ \sigma_\theta^2\ \sigma_{v_d}^2\right)^\mathsf{T}$ matrix, is the variances matrix of the targets, which in turn correspond to the uncertainty of the target's range, bearing and Doppler velocity. So the $\left(\sigma_\rho^2\ \sigma_\theta^2\ \sigma_{v_d}^2\right)^\mathsf{T}$ matrix represents corresponds to the EB of mmWave radars. With all the resolutions discussed in Section III-A, the relations are below:

$$\begin{pmatrix} \Delta\rho^2 \\ \Delta\theta^2 \\ \Delta {v_d}^2 \end{pmatrix} = \begin{bmatrix} \frac{1}{4} & 0 & 0 \\ 0 & \frac{1}{4} & 0 \\ 0 & 0 & \frac{1}{4} \end{bmatrix} \begin{pmatrix} \sigma_\rho^2 \\ \sigma_\theta^2 \\ \sigma_{v_d}^2 \end{pmatrix}. \tag{5}$$

Here, we have the assumptions that the discrete Fourier transform (DFT) size is the same as input data size. However, as direction of arrival (DOA) estimation can be estimated by conventional DFT/FFT method and several high resolution algorithms, *e.g.*, MUSIC algorithm [31] and ESPIRIT algorithm [32], the Δ_θ^2 term is modified to $\Delta_{\theta_{\text{DOA}}}^2$ according to number of angle bins. The 0's in the matrix represents the uncorrelated from range to angle, angle to Doppler and range to Doppler, respectively. The $1/4$ term is the minimum requirement to avoid aliasing. This way, the EB of mmWave radar is defined.

For the third case, we assume that the mmWave radar is a linear system with an output range $\boldsymbol{\rho_o}$ which is the response of an input signal $\mathbf{s}$, the measurement noise $\mathbf{e}$ can then be obtained by:

$$\boldsymbol{\rho_o} = \boldsymbol{\rho} + \mathbf{e} = \mathbf{hs} + \mathbf{e}\,, \tag{6}$$

where $\boldsymbol{\rho}$ is the raw measurement from sensor and $\mathbf{h}$ is the transform from input signal to output signal. Using least-squares estimation from orthogonal principles [33], and choosing appropriate $\mathbf{h}$ to minimize the Euclidean distance, the minimized projection error $\mathbf{e}$ is orthogonal to every column of $\mathbf{s}$ if and only if [34]:

$$\mathbf{s}^\mathsf{T}\mathbf{e} = 0\,, \tag{7}$$

Therefore, by orthogonality,

$$\mathbf{h} = (\mathbf{s}^\mathsf{T}\mathbf{s})^{-1}\mathbf{s}^\mathsf{T}\boldsymbol{\rho_o}. \tag{8}$$

In this fusion-EKF, the measurement processing noise is minimized with the least sum of $\|\mathbf{e}\|^2$ to reduce effects onto sensors' fusion. In the later fusion-EKF stages, this measurement noise only depends on the time variance of the receiving signal from equations. Therefore, the inherent measurement noise $\mathbf{R}_{\text{radar}}$ is used to produce the EKF estimations.

In this application, the processing noise is then the variance from the processing time from the radar board. The processing covariance noise from radar, $\mathbf{Q}_{\text{radar}}$, is related to the motion variance of acceleration of the object and the processing time variance Δt from radar. Similarly, we have camera's $\mathbf{Q}_{\text{cam}}$ which is associated with the camera's processing time variance.

In the final scenario, the antenna gain results in SNR difference, which results in the center detections to carry less noise than the surroundings. When it comes to sensor fusion, the radial detections around corners are noisy and could lead to missed detections. This potential failure should be compensated by other sensors like cameras. Note that the antenna gain in this application only reflects the SNR at different detecting angles. Measurements in the anechoic chamber are conducted and the antenna gain at 0 degrees is 10 dB compared to 0 dB at ± 60 degrees for the TI mmWave radar that we used in this study. The reference can be seen in TI's document [35].

C. HEM

The transformation between heterogeneous sensors assumes that there exists a relationship between the targets detected in different sensors' domain. An HEM is used to associate radar's detection plane and camera's image plane:

$$c\begin{pmatrix} u \\ v \\ 1 \end{pmatrix} = \mathbf{T}\begin{pmatrix} x \\ y \\ 1 \end{pmatrix}, \tag{9}$$

where c is a non-zero constant and x, y are the coordinates of the warped image in world coordinates. $\mathbf{T}$ is the transformation matrix which we are interested in, in order to warp the image plane to world plane. It is a 3-by-3 matrix and can be decomposed as:

$$\mathbf{T} = \begin{bmatrix} \mathbf{T}_1 & \mathbf{T}_2 & \mathbf{T}_3 \end{bmatrix}^\mathsf{T}. \tag{10}$$

We collect vision data with a calibrated camera, and try to associate the radar's localization of a specific point target like corner reflector. With Moore-Penrose pseudo inverse [36], the computing of the Euclidean norm solution can be obtained by linking several measured pairs. By applying least squares to the HEM, the transformation matrix $\mathbf{T}$ can be obtained as:

$$\begin{cases} \mathbf{T}_1 = (\mathbf{J}^\mathsf{T}\mathbf{J})^{-1}\mathbf{J}^\mathsf{T}\mathbf{U} \\ \mathbf{T}_2 = (\mathbf{J}^\mathsf{T}\mathbf{J})^{-1}\mathbf{J}^\mathsf{T}\mathbf{V} \\ \mathbf{T}_3 = (\mathbf{J}^\mathsf{T}\mathbf{J})^{-1}\mathbf{J}^\mathsf{T}\mathbf{I}, \end{cases} \tag{11}$$

where

$$\begin{aligned} \mathbf{J} &= \begin{bmatrix} x_1 & y_1 & 1 \\ x_2 & y_2 & 1 \\ \vdots & \vdots & \vdots \\ x_n & y_n & 1 \end{bmatrix} \\ \mathbf{U} &= \begin{bmatrix} cu_1 & cu_2 & \cdots & cu_n \end{bmatrix} \\ \mathbf{V} &= \begin{bmatrix} cv_1 & cv_2 & \cdots & cv_n \end{bmatrix} \\ \mathbf{I} &: \text{identity matrix} \\ n &: \text{number of measurements.} \end{aligned}$$

When collecting n samples of pairs from world coordinates to the camera image plane, the warped BEV can be best estimated by the transformation matrix $\mathbf{T}$.

After warped view from image towards BEV is applied, the BEV view of warped image can be used to associate with the radar's detection point cloud map. The warped BEV image with bounding box (BBox) is consistently used

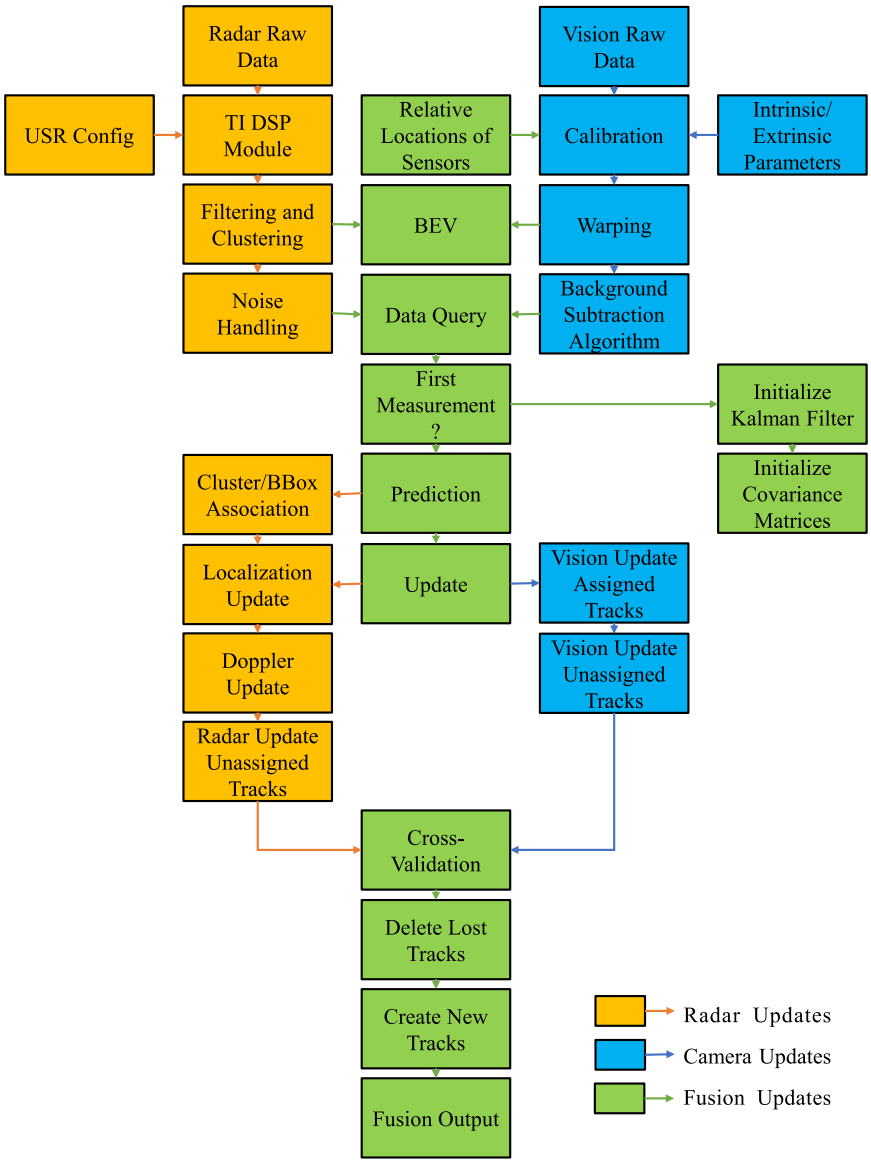


FIGURE 4. Radar-camera fusion-EKF workflow.

in fusion-EKF for further processing. With camera's (x, y) coordinates for objects, the radar's detections with range, azimuth angle and Doppler velocities are then fused and provide tracking from fusion-EKF.

The BBox generation is performed using a background subtraction algorithm (BSA) [37]. In future processing within the EKF, the BBoxes and radar clusters are associated according to the algorithm introduced in Section III-E.

D. FUSION-EKF

The full workflow of this radar-camera fusion-EKF is shown in Fig. 4.

The radar, camera and fusion updates are shown in yellow, blue and green, respectively. Radar raw data is fed with an ultra short range (USR) config on the digital signal processing (DSP) chip module on TI radar. With output through a USB port to a host computer, the point cloud with Doppler information is extracted. Additional filtering like moving target indicator (MTI) is applied to collaborate with the camera's BSA. A DBSCAN [38] or REDBSCAN [39] clustering algorithm is then applied to obtain clusters and potential target shape from radar scans. These are important parts to identify targets and similarly create "bounding box" for radar targets before the radar-camera fusion-EKF. After the first stage, the noise deductions and the query data is sent to radar-camera fusion-EKF for further association and tracking.

Similarly, camera raw data is fed with intrinsic and extrinsic parameters with the exact locations and directions measured during the experiment. The calibration stage provides the BEV from warping the image onto the perspective view. The BSA is then used for image plane BBox generation.

The clustering in radar preprocessing and the BBox generation in vision preprocessing have an inherent association. But without applied machine learning or segmentation, the data association needs to be done within the radar-camera fusion-EKF. Further details are discussed in Section III-E.

The fusion-EKF uses the CA maneuvering model in our setup. The CA model is a simplified model of Singer model [40] but without the acceleration factor, thus lowering orders in time deviation and further improve the processing speed. For tracking multiple targets with different velocities and positions, the CA model can handle different preset acceleration noises.

The key difference in our setup compared to other relevant work in this area is that (i) EBs are updated by both sensors, and (ii) radar-camera fusion-EKF updates both vision BBox plane and radar localization BEV continuously. In this case, the time stamps of the prediction and update are different based on each sensor's frames-per-second (fps). This will be discussed in Section III-E. In the next set of relations, we denote the current state with k and the previous state with $k-1$.

1) PREDICTION

The prediction part of this radar-camera fusion-EKF is no different from a typical single sensor maneuvering tracking, except for the CA model. If we have a state vector of a target $\mathbf{x}$ with position p and velocity v, we can define the state vector as a four-element vector with position and velocity projections onto x-axis and y-axis. That is:

$$\mathbf{x} = (p, v)^\mathsf{T} = (p_x, p_y, v_x, v_y)^\mathsf{T}. \tag{12}$$

State prediction does not depend on whether it is from the radar or camera. However, it will predict the updated state from whichever sensor it gets the current state from. Here, we need a state transition matrix which predicts the fusion-EKF output with the state transition matrix:

$$\mathbf{x}_k = \mathbf{F}_k \mathbf{x}_{k-1} + \mathbf{B}_k \mathbf{u}_k + \mathbf{w}, \tag{13}$$

where

$\mathbf{F}_k$: state transition matrix
$\mathbf{x}_{k-1}$: previous state $k-1$ state vector
$\mathbf{B}_k$: control input matrix
$\mathbf{u}_k$: control input vector
$\mathbf{w}$: compound noise from system (motion noise and processing noise).

The compound noise $\mathbf{w}$ obeys zero mean Gaussian distribution

$$\mathbf{w} \sim \mathcal{N}(0, \mathbf{Q}), \tag{14}$$

where $\mathbf{Q}$ is the processing noise introduced before as $\mathbf{Q}_{\text{radar}}$ or $\mathbf{Q}_{\text{cam}}$. As time stamps are changing, the time stamp difference Δt should also be updated throughout the process.

From the normal distribution of noise, an uncertainty covariance can be obtained for linearity. The uncertainty covariance matrix, also often referred to priori error covariance, for the prediction step is:

$$\mathbf{P}_k = \mathbf{F}_k \mathbf{P}_{k-1} \mathbf{F}_k^\mathsf{T} + \mathbf{Q}_k. \tag{15}$$

2) UPDATE

Similar to the prediction state vector definition, we need an update measurement vector $\mathbf{z}$ which consists of radar's and camera's measurement. Both measurements follow the same procedures. For radar data,

$$\mathbf{z}_{\text{radar}} = (\rho, \theta, v_d)^\mathsf{T}, \tag{16}$$

and for camera data,

$$\mathbf{z}_{\text{cam}} = (u, v)^\mathsf{T}. \tag{17}$$

The update consists of a residual $\mathbf{y}$ term, which is the error of the measurement from the prediction

$$\mathbf{z}_k = \mathbf{H}_k \mathbf{x}_k + \mathbf{y}_k, \tag{18}$$

where

$\mathbf{H}_k$: measurement function matrix.

The $\mathbf{y}$ part obeys

$$\mathbf{y} \sim \mathcal{N}(0, \mathbf{R}). \tag{19}$$

where the $\mathbf{R}$ is the inherent measurement noise introduced before as $\mathbf{R}_{\text{radar}}$ or $\mathbf{R}_{\text{cam}}$. The random variable also has a normal distribution. Thus the measurement error covariance matrix becomes:

$$\mathbf{S}_k = \mathbf{H}_k \mathbf{P}_k \mathbf{H}_k^\mathsf{T} + \mathbf{R}_k. \tag{20}$$

The optimal KF gain $\mathbf{G}$ can be obtained as:

$$\mathbf{G}_k = \mathbf{P}_k \mathbf{H}_k^\mathsf{T} \mathbf{S}_k^{-1}. \tag{21}$$

The state vector update $\mathbf{x}_{k(\text{update})}$ is based on prediction:

$$\mathbf{x}_{k(\text{update})} = \mathbf{x}_k + \mathbf{G}_k \mathbf{y}_k. \tag{22}$$

$\mathbf{x}_{k(\text{update})}$ is then used for the next state prediction step as $\mathbf{x}_{k-1}$. Similarly, it is also applied to the uncertainty covariance matrix $\mathbf{P}_{k(\text{update})}$:

$$\mathbf{P}_{k(\text{update})} = (\mathbf{I} - \mathbf{G}_k \mathbf{H}_k) \mathbf{P}_k, \tag{23}$$

The uncertainty covariance, is often used to scale the response of the KF. Commonly, in terms of representing the response, the KF gain $\mathbf{G}$ is used. It represents the relative weight of the measurements compared to the current state estimate, and is typically changing over time as the noise keeps changing throughout time. Rewriting the Equation (21) with Equation (20):

$$\mathbf{G}_k = \frac{\mathbf{P}_k \mathbf{H}_k^\mathsf{T}}{\mathbf{H}_k \mathbf{P}_k \mathbf{H}_k^\mathsf{T} + \mathbf{R}_k}. \tag{24}$$

Thus we can conclude that $\mathbf{G}$ as depends on $\mathbf{H}_k\mathbf{P}_k\mathbf{H}_k^\intercal$ and $\mathbf{R}_k$ terms, with $\mathbf{R}_k$ is fixed in a single measurement and not changing throughout the k steps. When $\mathbf{H}_k\mathbf{P}_k\mathbf{H}_k^\intercal$ is much higher than the $\mathbf{R}_k$ term, $\mathbf{G}$ is smaller. The low $\mathbf{G}$ will result in the model relying more on predictions. Hence, it is less responsive to measurements. On the other hand, when $\mathbf{H}_k\mathbf{P}_k\mathbf{H}_k^\intercal$ is much smaller than the $\mathbf{R}_k$ term, $\mathbf{G}$ is larger, and the fusion-EKF puts more weights on the current measurement. Thus, it gets more responsive to measurements. Fusion-EKF has the ability to adjust camera's or radar's weights to obtain better EBs on different dimensions, *e.g.*, camera's angular resolution and radar's range resolution. The $\mathbf{G}$ term also affects the region query and therefore the EBs of both sensors, which will be discussed in Section III-E.

3) NON-LINEARITY

One of our contributions in this paper is to exploit the non-linear property in EKF to deal with the radar's Doppler data, because Doppler is the range rate - the partial derivative of the range with time.

Because radar has detections in polar coordinates and BEV is in Cartesian coordinates, the conversion is needed for every incoming measurement. Additionally, from Equation (18), the $\mathbf{H}$ matrix should have a convertible formation to map radar detections from the state vector $\mathbf{x}$ to analyze $\mathbf{y}$. If we change Equation (18) for radar observations:

$$\mathbf{z}_k = h(\mathbf{x}_k) + \mathbf{y}_k. \tag{25}$$

The $h(\mathbf{x}_k)$ function represents a way to convert position and velocity of state vector to radar coordinates. Hence the $h(\mathbf{x}_k)$ function can be further updated as:

$$h(\mathbf{x}_k) = \begin{pmatrix} \rho \\ \theta \\ v_d \end{pmatrix} = \begin{pmatrix} \rho \\ \theta \\ \frac{\partial \rho}{\partial t} \end{pmatrix} = \begin{pmatrix} \sqrt{p_x^2 + p_y^2} \\ \arctan(p_y/p_x) \\ \frac{p_x v_x + p_y v_y}{\sqrt{p_x^2 + p_y^2}} \end{pmatrix}. \tag{26}$$

Heading back to the $\mathbf{H}$ matrix, the 3-by-4 matrix has higher dimensional components as the $h(\mathbf{x}_k)$ function indicates. Thus it is a first-order non-linear function in terms of radar-camera fusion-EKF. A way to deal with this is to use Jacobian matrix on $\mathbf{H}$:

$$\begin{aligned} \mathbf{H} &= \left[\frac{\partial h(\mathbf{x}_k)}{\partial \mathbf{x}_k}\right]_{\mathbf{x}_{k-1}} \\ &= \begin{bmatrix} \frac{\partial \rho}{\partial p_x} & \frac{\partial \rho}{\partial p_y} & \frac{\partial \rho}{\partial v_x} & \frac{\partial \rho}{\partial v_y} \\ \frac{\partial \theta}{\partial p_x} & \frac{\partial \theta}{\partial p_y} & \frac{\partial \theta}{\partial v_x} & \frac{\partial \theta}{\partial v_y} \\ \frac{\partial v_d}{\partial p_x} & \frac{\partial v_d}{\partial p_y} & \frac{\partial v_d}{\partial v_x} & \frac{\partial v_d}{\partial v_y} \end{bmatrix}. \end{aligned} \tag{27}$$

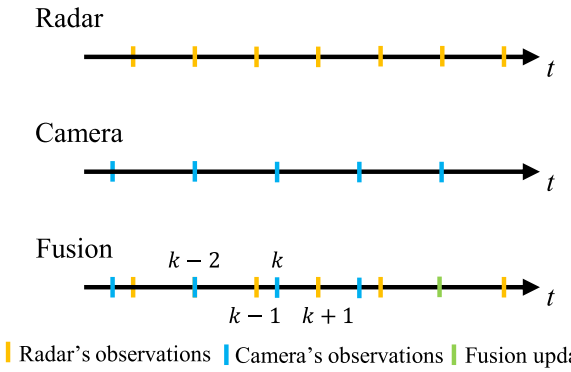
Combining Equations (26) and (28), the resulting $\mathbf{H}$ for radar is shown at the top of the next page.

E. DATA ASSOCIATION AND SENSOR SYNCHRONIZATION

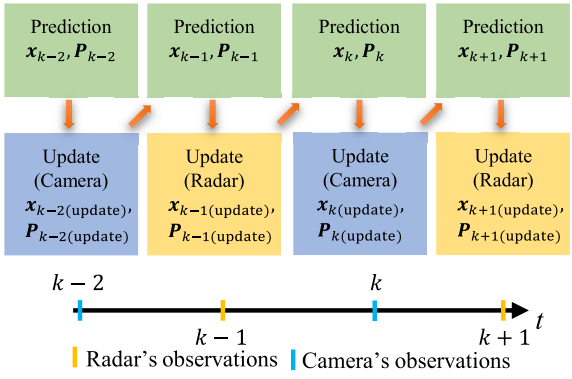
The track-oriented data association is implemented using timeline alignment of sensors to (i) get updates from heterogeneous sensor's perspective, and (ii) synchronization over certain regions of fusion-EKF output to obtain the tracked target localization. This process is overlying on the radar-camera fusion-EKF in Fig. 4. With HEM to transform between coordinates, EBs are recalled to optimize the association of radar clusters and camera BBoxes. The fusion-EKF track is created or deleted according to the existence from both sensors.

1) TIMELINE ALIGNMENT

The timeline alignment for radar-camera fusion system can be seen in Fig. 5. With different fps from both sensors, observations of different sensor type on different time stamps are collected on the processing unit. And fusion association is shown with different Δt with different state steps k in Fig. 5a. The green-colored step on the fusion timeline actually means both sensors output their observations at the same time ($\Delta t \leq 0.005s$). In this case, predictions and updates follow the same rule as the radar-camera fusion-EKF.



(a) Radar-camera association timeline. A sample state steps $k-2, k-1, k, k+1$ are shown in the fusion timeline.



(b) Association actions based on timeline alignment of state steps $k-2, k-1, k, k+1$.

FIGURE 5. Radar-camera association timeline alignment.

We set the model with a random acceleration vector $\mathbf{a} = (a_x, a_y)^\intercal$ with variances of $\sigma_a x^2$ and $\sigma_a y^2$ on x-axis and y-axis, respectively. In this case, the $\mathbf{w}$ noise matrix, which is

$$\mathbf{H} = \begin{bmatrix} \frac{p_x}{\sqrt{p_x^2+p_y^2}} & \frac{p_y}{\sqrt{p_x^2+p_y^2}} & 0 & 0 \\ -\frac{p_y}{p_x^2+p_y^2} & \frac{p_x}{p_x^2+p_y^2} & 0 & 0 \\ \frac{p_y(v_x p_y - v_y p_x)}{(p_x^2+p_y^2)^{3/2}} & \frac{p_x(v_y p_x - v_x p_y)}{(p_x^2+p_y^2)^{3/2}} & \frac{p_x}{\sqrt{p_x^2+p_y^2}} & \frac{p_y}{\sqrt{p_x^2+p_y^2}} \end{bmatrix}. \tag{28}$$

the motion EB for the tracked model, can be expressed as:

$$\mathbf{w} = \begin{pmatrix} w_{px} \\ w_{py} \\ w_{vx} \\ w_{vy} \end{pmatrix} = \begin{pmatrix} \frac{a_x \Delta t^2}{2} \\ \frac{a_y \Delta t^2}{2} \\ a_x \Delta t \\ a_y \Delta t \end{pmatrix}. \tag{29}$$

From Equation (14), the noise is described by a zero mean and a covariance matrix $\mathbf{Q}$. When referring to Equations (15) and (29), with variances of $\sigma_a x^2$ and $\sigma_a y^2$ applied, the $\mathbf{Q}$ matrix can be estimated by taking expectations of the $\mathbf{w}$ and $\mathbf{w}^\intercal$ matrix as:

$$\begin{aligned} \mathbf{Q} &= \mathbb{E}[\mathbf{w}w^\intercal] \\ &= \begin{bmatrix} \frac{\sigma_{ax}^2 \Delta t^4}{4} & 0 & \frac{\sigma_{ax}^2 \Delta t^3}{2} & 0 \\ 0 & \frac{\sigma_{ay}^2 \Delta t^4}{4} & 0 & \frac{\sigma_{ay}^2 \Delta t^3}{2} \\ \frac{\sigma_{ax}^2 \Delta t^3}{2} & 0 & \sigma_{ax}^2 \Delta t^2 & 0 \\ 0 & \frac{\sigma_{ax}^2 \Delta t^3}{2} & 0 & \sigma_{ax}^2 \Delta t^2 \end{bmatrix}. \end{aligned} \tag{30}$$

Hence $\mathbf{Q}$ depends on Δt and the random acceleration variances. Δt changes from time to time because of different fps of camera and radar. With high Δt, the uncertainty and the large EB are generated, and vice versa.

Similarly, we apply this to measurement noise matrix $\mathbf{R}_{\text{radar}}$. Combining Equations 1 to 5. The result $\mathbf{R}_{\text{radar}}$ is:

$$\begin{aligned} \mathbf{R}_{\text{radar}} &= \mathbb{E}[\mathbf{y}\mathbf{y}^\intercal] \\ &= \begin{bmatrix} \sigma_\rho^2 & 0 & 0 \\ 0 & \sigma_\theta^2 & 0 \\ 0 & 0 & \sigma_{v_d}^2 \end{bmatrix}. \end{aligned} \tag{31}$$

In Fig. 5b, as we have two tuned updates for each step depending on which sensor data is obtained from, the prediction and update for each observation from radar or camera are shown. As stated in the previous section, the updated $\mathbf{x}_{k(\text{update})}$ and $\mathbf{P}_{k(\text{update})}$ matrices are used for the next prediction step. In other words, the measurement update of $\mathbf{x}_{k(\text{update})}$ and $\mathbf{P}_{k(\text{update})}$ is only based on the prediction within the same step. And the prediction relies on the previous update step.

2) REGION SEARCH WITHIN EBs

From Section III-D, we have an idea of the KF gain $\mathbf{G}$. The association of radar clusters and camera BBoxes based on KF gain $\mathbf{G}$ using HEM is shown in Fig. 6. The fusion-EKF trajectory in blue dash line is shown with steps $k-1, k, k+1$. The different predictions after update measurements are shown with positions and velocities. The yellow region query for the upcoming measurement is set in Fig. 6. Here are the EBs resulting from Definitions 1 and 2:

1) EB along x-axis of radar is

$$\text{EB}|x_{\text{radar}} = \sigma_\rho \cos\theta_i\,, \tag{32}$$

where θ_i is the angle of interest from target, which is the target angular location. EB along azimuth (y-axis) of radar is

$$\text{EB}|y_{\text{radar}} = \sigma_\rho \sin\theta_i. \tag{33}$$

The assumption is that the radar has even spacing MIMO antenna with even angular resolution. Thus we can conclude that at the center, radar has better range EB and less accurate when it is on the side. On the other hand, radar has relative bad azimuth EB and needs other sensors to compensate for the angular EB.

2) EB along range of camera is

$$\text{EB}|x_{\text{cam}} = c(\mathbf{T}^\intercal\mathbf{T})^{-1}\mathbf{T}^\intercal\mathbf{U}. \tag{34}$$

And EB along azimuth of camera is

$$\text{EB}|y_{\text{cam}} = c(\mathbf{T}^\intercal\mathbf{T})^{-1}\mathbf{T}^\intercal\mathbf{V}. \tag{35}$$

For camera pixels using HEM, warping along the vertical axis to x-axis is typically limited compared to the horizontal axis to y-axis. Thus the range resolution is restricted especially for long ranges. The azimuth resolution is consistent when looking at horizontal pixels to angle pair. So camera's EB along the range is worse compared to radar's. And EB along the azimuth is better with adequacy horizontal pixels.

3) Combining radar and camera EBs. There exist two standalone query regions to search for cluster or BBox from heterogeneous sensors.
4) Fusion EBs use EKF output of the new prediction to get the estimated localization of the cluster/BBox. This results in missing clusters/BBoxes or multiple clusters/BBoxes.

Even with the fusion EBs, there still exists multiple clusters/BBoxes or missing clusters/BBoxes. Reasons include:

1) Concealed by front object using the camera.
2) Sensor failure by radar or camera.

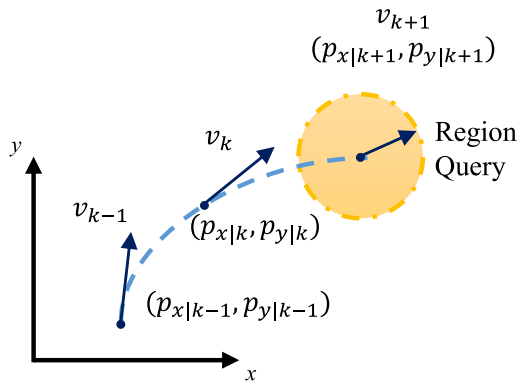


FIGURE 6. Radar-camera association region search. Blue dash line: EKF projected trajectory, yellow region: Search region for next state for radar clusters and camera bounding box BEV projection.

3) Radar's multipath effect.
4) Lighting condition (over-exposed or no lighting) by the camera.
5) Others.

Algorithm 1 shows the region search and track updates for tracking multiple targets. Different tracks share the incoming clusters/BBoxes measurement updates from sensors. Fusion-EKF deals with the region search based on EBs estimations. The fusion-EKF provides invisible counts for heterogeneous sensors. If either sensor fails temporary, the tracking will continue using the other sensor's update. The tracking will still be valid even if misses several clusters/BBoxes. However, a track will be terminated after a certain duration of missing detections from both sensors after cross-validation. With updates for fusion localization, the improved EBs can provide a better result for the system and thus greatly decreases uncertainty errors from either sensor. In addition, new tracks are initialized based on cross-validation from different sensors with a certain threshold.

Therefore, the association is done with timeline alignment together with EBs region search within tracks. With help from different advantages from heterogeneous sensors, fusion EBs are greatly reduced: along range using the radar as it provides better localization; and along azimuth using the camera as it provides better localization. The experiment result will further approve of this association and EBs' deduction.

F. EB EVALUATION

From Fig. 6 and Definition 2, the EB of the next query is the region of the prediction and association. With RMSE, the EB evaluation can be estimated as:

$$\begin{aligned}\mathrm{RMSE}(\hat{o}) &= \sqrt{\mathbb{E}((\hat{o}-o)^2)} \\ &= \sqrt{\frac{1}{N}\sum_{n=1}^{N}(\hat{o}-o)^2}\,, \end{aligned} \tag{36}$$

where o is the operator of the estimation variable.

Algorithm 1 Pseudo-Code of the Tracking and Association for Multi-Targets

Require: EKF tracks: *tracks* at current time stamp
Initialize: $i = 0$
After radar-camer fusion-EKF, get number of tracks: N
while $i < N$ **do**
  if Track valid **then**
    Find EBs for next step from prediction
    Find radar clusters and BBoxes within EBs
    if *radarclusters* not equal to 1 **then**
      $radarInvisibleCount++$
    else
      Update radar's localization
    end if
    if *BBoxes* not equal to 1 **then**
      $camInvisibleCount++$
    else
      Update camera's localization
    end if
  end if
  if $radarInvisibleCount > threshold_{\mathrm{radar}}$ **then**
    Cross-validation failed.
    Track invalid, valid until camera updates
  end if
  if $camInvisibleCount > threshold_{\mathrm{cam}}$ **then**
    Cross-validation failed.
    Track invalid, valid until radar updates
  end if
  $i++$
end while
Return: EBs with localization for each track

With a window function to estimate the EB, Equation (36) is modified with

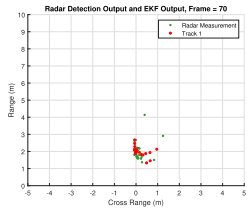
$$\mathrm{RMSE}(\hat{o}) = \sqrt{\frac{1}{w}\sum_{n=k-(w-1)/2}^{k+(w-1)/2}(\hat{o}-o)^2}\,, \tag{37}$$

where w is the window size of the estimation on the k^{th} step. As introduced before, the EBs of radar are the radar range, angle and Doppler variances, $(\sigma_\rho^2\ \sigma_\theta^2\ \sigma_{v_d}^2)^{\mathsf{T}}$. Here the sliding window applied RMSE is used to measure the EB on where the prediction region is. In this case, the EB evaluates radar's fusion performance with camera's vision.

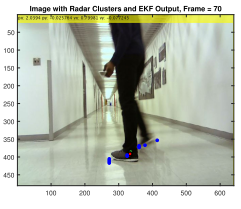
In the experiment, the EB of radar is measured and evaluated with RMSEs in range, Doppler and angle domains, respectively.

IV. EXPERIMENTAL RESULTS

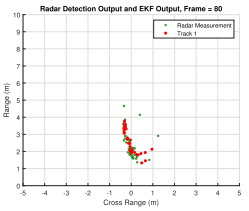
The experiment is implemented by two humans' walking inside an indoor environment. The application implemented sensor fusion in this experiment used ELP USB8MP02G-L75 monocamera with 640 × 480 YUY2 video picture encoding format [41], transmitting at 30 fps, and a TI



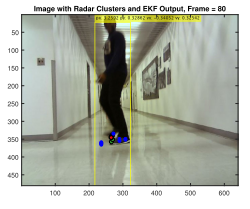
(a) Frame number 70 radar EKF output



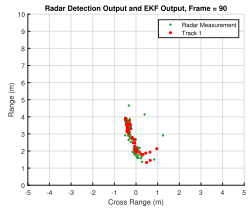
(b) Frame number 70 camera EKF output



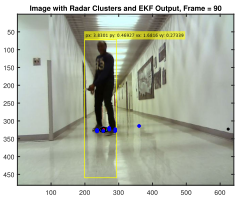
(c) Frame number 80 radar EKF output



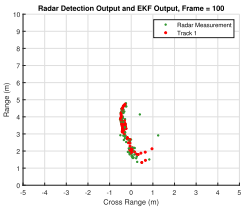
(d) Frame number 80 camera EKF output



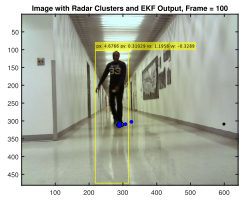
(e) Frame number 90 radar EKF output



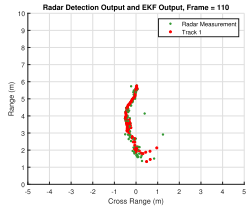
(f) Frame number 90 camera EKF output



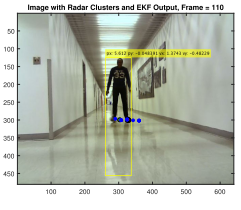
(g) Frame number 100 radar EKF output



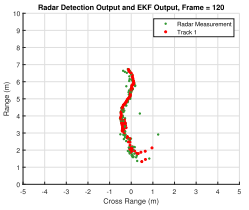
(h) Frame number 100 camera EKF output



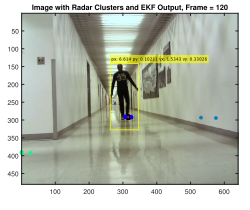
(i) Frame number 110 radar EKF output



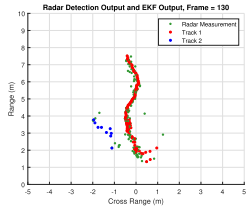
(j) Frame number 110 camera EKF output



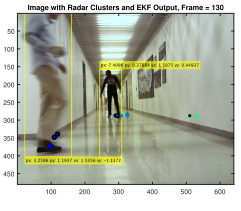
(k) Frame number 120 radar EKF output



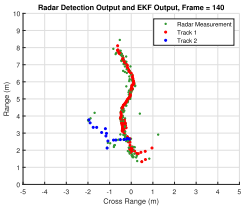
(l) Frame number 120 camera EKF output



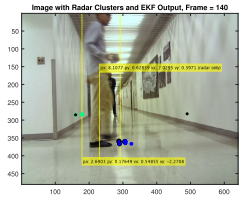
(m) Frame number 130 radar EKF output



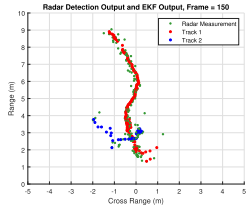
(n) Frame number 130 camera EKF output



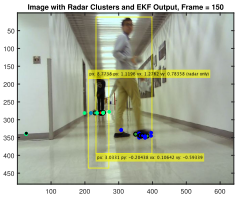
(o) Frame number 140 radar EKF output



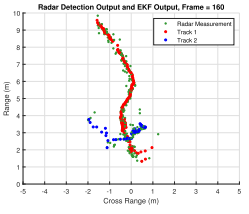
(p) Frame number 140 camera EKF output



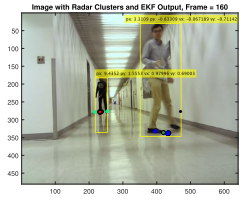
(q) Frame number 150 radar EKF output



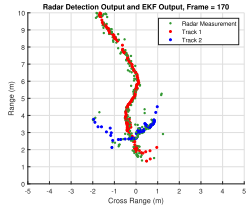
(r) Frame number 150 camera EKF output



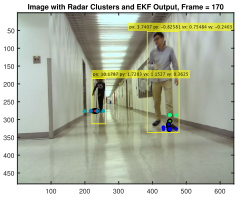
(s) Frame number 160 radar EKF output



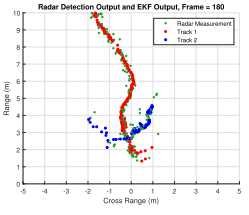
(t) Frame number 160 camera EKF output



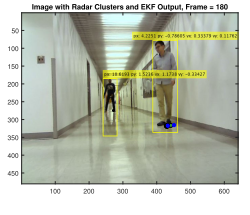
(u) Frame number 170 radar EKF output



(v) Frame number 170 camera EKF output

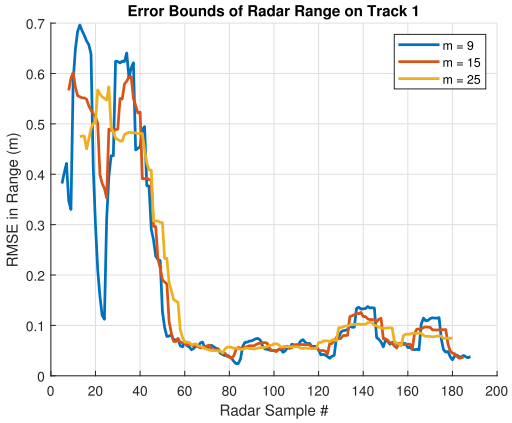


(w) Frame number 180 radar EKF output

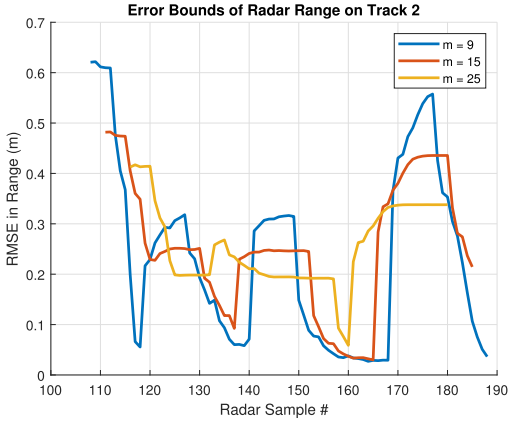


(x) Frame number 180 camera EKF output

FIGURE 7. EKF output.

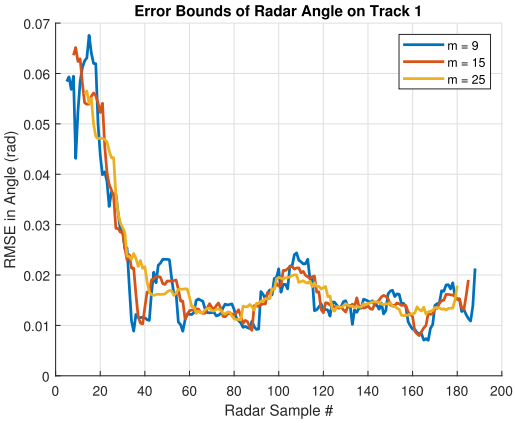


(a) Radar's range RMSE of EKF estimates with memory of 9, 15 and 25 measurements for track 1

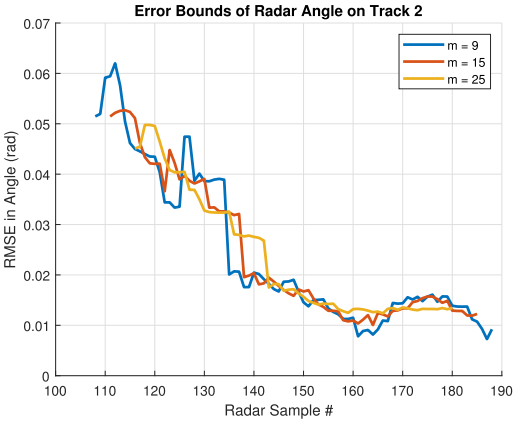


(b) Radar's range RMSE of EKF estimates with memory of 9, 15 and 25 measurements for track 2

FIGURE 8. EKF RMSE of radar's range.



(a) Radar's angle RMSE of EKF estimates with memory of 9, 15 and 25 measurements for track 1



(b) Radar's angle RMSE of EKF estimates with memory of 9, 15 and 25 measurements for track 2

FIGURE 9. EKF RMSE of radar's angle.

AWR1642 mmWave radar working at 77 GHz at 30 fps, which is a 1D 4 RX by 2 TX phased array antenna. AWR1642 is only capable of measuring target in range and azimuth plane.

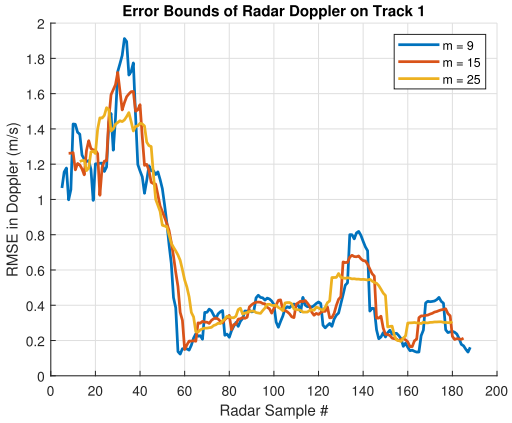
The vision output and radar output from EKF can be seen from Fig. 7. In the vision's output, bounding boxes with positions and velocities on x-axis and y-axis are shown, along with the radar's point cloud clusters. Different colors stand for different clusters with the searching radius of a human. The clusters provide median filtered centroids with a green marker and black marker edge. Red and blue markers with black marker edge are the EKF output of the tracking. In the radar's output, radar point clusters' centroid, as well as fusion-EKF prediction localization for different tracks are presented. These figures are combinations of tracks thus showing multiple targets' multiple sensors' tracking results.

As we stated in the Section III-D, we only have a 2D radar, *i.e.*, the elevation information is ignored, and thus only points on the ground are shown. This approach is also applicable when applying to the elevation information obtained from a compatible sensor to have tracking more reliable.

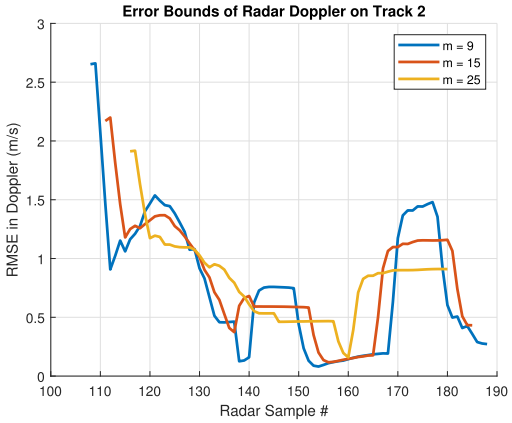
Additionally, in case of blockage between targets, the output bounding box shows "radar only" in Fig. 7p and Fig. 7r. As invisible from the camera's BSA. The "radar only" label still keeps tracking of different tracks until a camera's vision validates the track. Otherwise, after certain invisible counts of the track, the track is deleted because of lack of cross-validation.

The radar's tracking, shown in Figures 7m, 7o, 7q, 7s, 7u and 7w, provides multiple targets' tracking. If camera vision data was unavailable, the output may be unreliable until the fusion EBs' update. The region search and association avoids the missed assignment of tracks and improves tracking results.

To evaluate the performance of the radar-camera fusion-EKF system, we use the RMSE metrics, and also to further validate if the fusion-EKF takes advantages from both sensors. From Equation (37), the range, angle and Doppler RMSE plots for different tracks 1 and 2 are shown in Fig. 8, 9 and 10, respectively. The memory used for RMSE is 9, 15 and 25, which are fast, median and slow windows, respectively. With fps of 30 Hz, the window is estimating a state in about 0.3 s, 0.5 s and 0.83 s, respectively.



(a) Radar's Doppler RMSE of EKF estimates with memory of 9, 15 and 25 measurements for track 1



(b) Radar's Doppler RMSE of EKF estimates with memory of 9, 15 and 25 measurements for track 2

FIGURE 10. EKF RMSE of radar's Doppler.

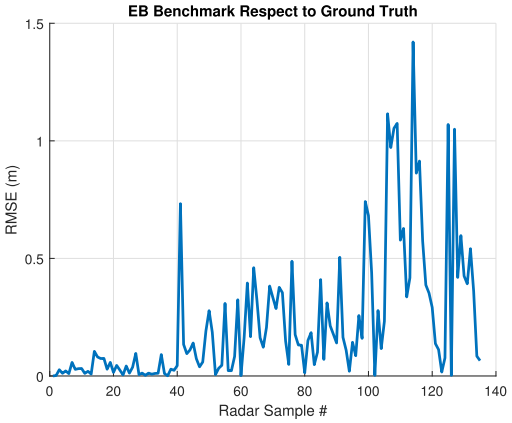


FIGURE 11. Fusion-EKF benchmark on EB respect to ground truth.

TABLE 1. Average localization error of radar-camera fusion-EKF.

	Average localization error
Range	0.2902 m
Azimuth angle	0.0134 rad
Velocity	0.7864 m/s

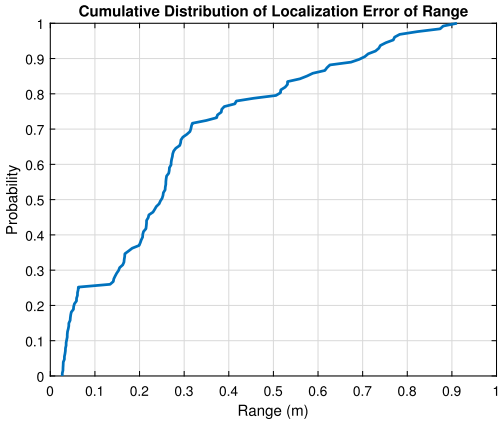
As from Fig. 8a, after initializing the track, the position RMSE drops rapidly to around 0.05 m. From Equation (1), the range resolution of the fusion system is 0.0436 m, which is the theoretical range resolution. After the fusion-EKF, the practical range resolution achieved is around 0.05 m after stabilizing from Fig. 8a. The weakness of radar's detecting cross-range objects is overcome from Fig. 8b. As for radar's cross-range objects in a typical radar scan, the cross-range velocities are close to zero because Doppler velocities are essentially range rates. This problem is solved by adding the camera's vision in the proposed fusion-EKF.

As from Fig. 9a, after initializing the track, the angular RMSE drops rapidly to around 0.015 rad, which is 0.859 degrees. From Equation (2), the angular resolution of radar at center FOV is 0.25 rad, which is the theoretical angular resolution at center FOV. Outside, that is $1/\cos\theta_i$. After the fusion-EKF, the angular resolution achieved is around 0.015 rad after stabilizing from Fig. 9a. This track is along about the center FOV so that it verifies that the camera improves fusion's track significantly. In Fig. 9b, a similar result is achieved. This confirms the improvements on radar's EB on the azimuth axis by using the camera data.

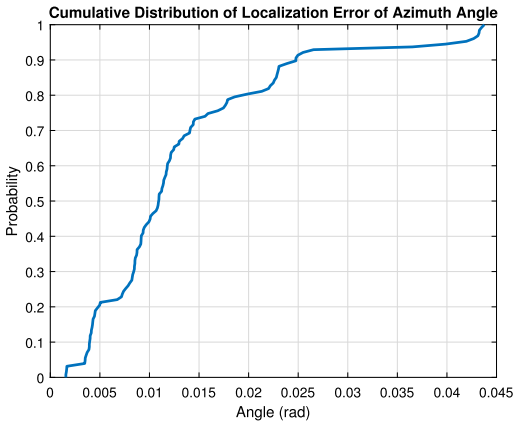
As from Fig. 10a, after initializing the track, the Doppler RMSE drops rapidly to around 0.3 m/s. From Equation (3), the theoretical Doppler resolution of radar is 0.616 m/s. After the fusion-EKF, the practical Doppler resolution achieved is around 0.3 m/s after stabilizing from Fig. 10b. The fusion-EKF improves Doppler estimations on both range rate and cross-range rate. In Fig. 10b, a resolution of only 0.5 m/s is achieved. Fusion-EKF greatly improves radar's cross-range rate detections, and thus allows for reliable cross-range objects detections.

From Figures 8 to 10, two motion maneuvering models are handled with different tracks and different CA models from fusion-EKF. The result shows different CA models can handle tracking for different targets with different maneuvering behaviors.

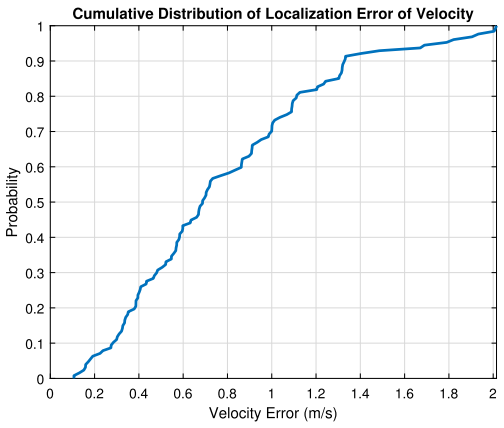
Additional experiment analysis on this EB-based sensor fusion is conducted that uses the sensor fusion result of radar and camera data and compares the localization result with ground truth. The benchmark of radar-camera fusion-EKF tracking result respect to ground truth can be seen in Fig. 11. The RMSE of fusion localization deviation from ground truth is 0.3664 m. This is a tolerable deviation that is good enough to separate human targets, and is a reasonable region search EB from Definitions 1 and 2. The average localization error comes from the track association from heterogeneous sensors and the performance of the proposed fusion-EKF can be seen in Table 1. With EBs reduced in range, angle and Doppler dimensions, the radar-camera fusion-EKF provides a solution to heterogeneous sensor fusion. Additionally, the cumulative distributions of the localization errors in range, azimuth angle and velocity are presented in Fig. 12.



(a) Cumulative distribution of localization error of range



(b) Cumulative distribution of localization error of azimuth angle



(c) Cumulative distribution of localization error of velocity

FIGURE 12. **Cumulative distribution of localization errors.**

Some comparison of errors or EBs between the proposed radar-camera fusion-EKF and other radar tracking algorithms are shown in Table 2. Because mmWave radar has better range resolution than traditional radars, range EB is much superior than the other lower frequency radars. Meanwhile, azimuth angle EB is also improved greatly due to the fusion approach with the camera. Therefore, the proposed fusion-EKF provides extra information from the camera to improve the detection and tracking of the mmWave radar sensor.

TABLE 2. **Comparison of error/EBs using radar-camera fusion EKF with other radar tracking algorithms.**

Method	Error or EBs
CA model of radar-camera fusion-EKF	0.2902 m, 0.0134 rad (Short range configuration with camera fusion)
CCA model from [19]	2.49 m (Euclidean distance without fusion)
600 m airport surveillance system from [42]	5 m, 0.15°
New track association algorithm from [43]	0.5°

V. CONCLUSION

In summary, the radar-camera fusion-EKF consists of fusing radar and camera raw data with EKF for tracking multiple objects at the same time. The association is carried out with coordinates transformation and error bounds estimations. The noisy outputs from the sensors are correspondingly processed by utilizing covariance matrices to ensure the reliability of detections. The fusion system takes both sensors' advantages and thus ensures error bounds minimization and better resolution on different perspectives from heterogeneous sensors. The RMSEs of range, angle and Doppler obtained from the experiment results are presented. Cross-validating camera's vision and radar's point clouds from raw data, the radar-camera fusion-EKF system provides better and more reliable detections of targets in real-time. The system provides a solution to heterogeneous sensors' fusion, association and synchronization, as well as cross-validating multiple targets and tracks with less computation load. The fusion-EKF provides a novel way to improve tracking and detecting reliability for radars in ADAS and autonomous driving applications. Future work includes using machine learning method to replace BSA and further improve the vision extractions of targets in the moving environments, as well as implementing classification algorithms like micro-Doppler signatures from radar [44] to classify targets. The dynamic reliability of sensors and sensor fusion should be further improved and evaluated.

ACKNOWLEDGMENT

The authors would like to thank Arindam Sengupta for his dedicated help on editing this paper. They would also like to show gratitude to Jean de Dieu Mutangana, Yiming Zhang, Kemeng Chen, Rahul Salvatore Bhadani and Feng Jin for their kind help.

REFERENCES

[1] S. Clark and H. Durrant-Whyte, "Autonomous land vehicle navigation using millimeter wave radar," in *Proc. IEEE Int. Conf. Robot. Automat.*, vol. 4, May 1998, pp. 3697–3702.
[2] T. Motomura, K. Uchiyama, and A. Kajiwara, "Measurement results of vehicular RCS characteristics for 79 GHz millimeter band," in *Proc. IEEE Top. Conf. Wireless Sensors Sensor Netw. (WiSNet)*, Jan. 2018, pp. 103–106.
[3] Q. J. O. Tan and R. A. Romero, "Ground vehicle target signature identification with cognitive automotive radar using 24–25 and 76–77 GHz bands," *IET Radar, Sonar Navigat.*, vol. 12, no. 12, pp. 1448–1465, 2018.

[4] J. Wei, J. M. Snider, J. Kim, J. M. Dolan, R. Rajkumar, and B. Litkouhi, "Towards a viable autonomous driving research platform," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Jun. 2013, pp. 763–770.

[5] N. Shima, M. Babasaki, Y. Akidzuki, K. F. Honda, T. Higuchi, H. Higashida, and R. Nakamura, "Fusion sensor for driving assistance system," *Fujitsu Ten Tech. J.*, no. 17, pp. 35–44, 2001. [Online]. Available: https://www.denso-ten.com/business/technicaljournal/10_19.html

[6] J. Li, D. Zheng, and P. Stoica, "Angle and waveform estimation via RELAX," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 33, no. 3, pp. 1077–1087, Jul. 1997.

[7] Y. Gürcan and A. Yarovoy, "Super-resolution algorithm for joint range-azimuth-Doppler estimation in automotive radars," in *Proc. IEEE Eur. Radar Conf. (EURAD)*, Oct. 2017, pp. 73–76.

[8] J. K. Wu and Y. Wong, "Bayesian approach for data fusion in sensor networks," in *Proc. IEEE Int. Conf. Inf. Fusion*, Jul. 2006, pp. 1–5.

[9] E. Dubrofsky, "Homography estimation," M.S. thesis, Fac. Graduate Stud. Comput. Sci., Univ. Brit. Columbia, Vancouver, BC, Canada, 2009. Accessed: Sep. 22, 2019. [Online]. Available: https://www.cs.ubc.ca/grads/resources/thesis/May09/Dubrofsky_Elan.pdf

[10] Y. Song, X. Wang, J. Zhu, and L. Lei, "Sensor dynamic reliability evaluation based on evidence theory and intuitionistic fuzzy sets," *Appl. Intell.*, vol. 48, no. 11, pp. 3950–3962, 2018.

[11] H. Hajri and M.-C. Rahal, "Real time lidar and radar high-level fusion for obstacle detection and tracking with evaluation on a ground truth," 2018, *arXiv:1807.11264*. [Online]. Available: https://arxiv.org/abs/1807.11264

[12] S. K. Kwon, E. Hyun, J.-H. Lee, J. Lee, and S. H. Son, "A low-complexity scheme for partially occluded pedestrian detection using LiDAR-radar sensor fusion," in *Proc. IEEE 22nd Int. Conf. Embedded Real-Time Comput. Syst. Appl. (RTCSA)*, Aug. 2016, p. 104.

[13] C. Debes, A. Merentitis, R. Heremans, J. Hahn, N. Frangiadakis, T. van Kasteren, W. Liao, R. Bellens, A. Pižurica, S. Gautama, W. Philips, S. Prasad, Q. Du, and F. Pacifici, "Hyperspectral and LiDAR data fusion: Outcome of the 2013 GRSS data fusion contest," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 6, pp. 2405–2418, Jun. 2014.

[14] M. Ulrich, F. Maile, A. Löcklin, B. Yang, B. Kleiner, and N. Ziegenspeck, "A model for improved association of radar and camera objects in an indoor environment," in *Proc. IEEE Sensor Data Fusion, Trends, Solutions, Appl. (SDF)*, Oct. 2017, pp. 1–6.

[15] A. Sikdar, S. Cao, Y. F. Zheng, and R. L. Ewing, "Radar depth association with vision detected vehicles on a highway," in *Proc. IEEE Radar Conf.*, May 2014, pp. 1159–1164.

[16] F. Bunyak, K. Palaniappan, S. K. Nath, and G. Seetharaman, "Flux tensor constrained geodesic active contours with sensor fusion for persistent object tracking," *J. Multimedia*, vol. 2, no. 4, p. 20, 2007.

[17] Z. Zhong, S. Liu, M. Mathew, and A. Dubey, "Camera radar fusion for increased reliability in adas applications," *Electron. Imag.*, vol. 2018, no. 17, pp. 258-1–258-4, 2018.

[18] M. P. Muresan and S. Nedevschi, "Multimodal sparse LiDAR object tracking in clutter," in *Proc. IEEE Int. Conf. Intell. Comput. Commun. Process. (ICCP)*, Sep. 2018, pp. 215–221.

[19] R. Schubert, E. Richter, and G. Wanielik, "Comparison and evaluation of advanced motion models for vehicle tracking," in *Proc. IEEE Int. Conf. Inf. Fusion*, Jun. 2008, pp. 1–6.

[20] M. Nishigaki, S. Rebhan, and N. Einecke, "Vision-based lateral position improvement of radar detections," in *Proc. IEEE Int. Conf. Intell. Transp. Syst. (ITSC)*, Sep. 2012, pp. 90–97.

[21] F. Folster and H. Rohling, "Lateral velocity estimation based on automotive radar sensors," in *Proc. IEEE Int. Conf. Radar*, Oct. 2006, pp. 1–4.

[22] P. Aditya, E. Apriliani, D. K. Arif, and K. Baihaqi, "Estimation of three-dimensional radar tracking using modified extended Kalman filter," *J. Phys., Conf. Ser.*, vol. 974, no. 1, 2018, Art. no. 012071.

[23] L. Ljung, "Asymptotic behavior of the extended Kalman filter as a parameter estimator for linear systems," *IEEE Trans. Autom. Control*, vol. AC-24, no. 1, pp. 36–50, Feb. 1979.

[24] F. M. Mirzaei, D. G. Kottas, and S. I. Roumeliotis, "3D LIDAR–camera intrinsic and extrinsic calibration: Identifiability and analytical least-squares-based initialization," *Int. J. Robot. Res.*, vol. 31, no. 4, pp. 452–467, Apr. 2012.

[25] Q. Zhang and R. Pless, "Extrinsic calibration of a camera and laser range finder (improves camera calibration)," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, vol. 3, Sep./Oct. 2004, pp. 2301–2306.

[26] Z. Zhang, "A flexible new technique for camera calibration," *IEEE Trans. Pattern Anal. Mach. Intell.*, to be published.

[27] J. Heikkila and O. Silven, "A four-step camera calibration procedure with implicit image correction," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 1997, pp. 1106–1112.

[28] Z. Wang and A. C. Bovik, "A universal image quality index," *IEEE Signal Process. Lett.*, vol. 9, no. 3, pp. 81–84, Mar. 2002.

[29] S. L. Wilson and B. D. Carlson, "Radar detection in multipath," *IEE Proc.-Radar, Sonar Navigat.*, vol. 146, no. 1, pp. 45–54, Feb. 1999.

[30] I.-H. Ryu, I. Won, and J. Kwon, "Detecting ghost targets using multilayer perceptron in multiple-target tracking," *Symmetry*, vol. 10, no. 1, p. 16, 2018.

[31] R. O. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propag.*, vol. 34, no. 3, pp. 276–280, Mar. 1986.

[32] R. Roy and T. Kailath, "Esprit-estimation of signal parameters via rotational invariance techniques," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 37, no. 7, pp. 984–995, Jul. 1989.

[33] T. Kailath, A. H. Sayed, and B. Hassibi, *Linear Estimation*. Upper Saddle River, NJ, USA: Prentice-Hall, 2000.

[34] J. O. Smith, *Mathematics of the Discrete Fourier Transform (DFT): With Audio Applications*. Charleston, SC, USA: BookSurge, 2007. [Online]. Available: https://books.google.com/books?id=fTOxS9huzHoC

[35] *AWR1642 Evaluation Module User Guide*. [Online]. Available: http://www.ti.com/lit/ug/swru508b/swru508b.pdf

[36] E. H. Moore, "On the reciprocal of the general algebraic matrix," *Bull. Amer. Math. Soc.*, vol. 26, no. 9, pp. 394–395, 1920.

[37] O. Barnich and M. Van Droogenbroeck, "ViBe: A universal background subtraction algorithm for video sequences," *IEEE Trans. Image Process.*, vol. 20, no. 6, pp. 1709–1724, Jun. 2011.

[38] D. Birant and A. Kut, "ST-DBSCAN: An algorithm for clustering spatial–temporal data," *Data Knowl. Eng.*, vol. 60, no. 1, pp. 208–221, Jan. 2007.

[39] R. Zhang and S. Cao, "Robust and adaptive radar elliptical density-based spatial clustering and labelling for mmWave radar point cloud," in *Proc. Asilomar Conf. Signals, Syst., Comput.*, Nov. 2019.

[40] X. R. Li and V. P. Jilkov, "Survey of maneuvering target tracking: Dynamic models," *Proc. SPIE*, vol. 4048, pp. 212–236, Jul. 2000.

[41] *YUV Pixel Formats*. Accessed: Sep. 22, 2019. [Online]. Available: http://fourcc.org/yuv.php

[42] J. Garcia, J. M. Molina, A. Berlanga, and G. de Miguel, "Data fusion alternatives for the integration of millimetre radar in airport surveillance systems," in *Proc. IEEE Int. Radar Conf.*, May 2005, pp. 796–801.

[43] B. Peng and X. Guan, "A new track association algorithm of radar and esm," in *Proc. CIE Int. Conf. Radar (RADAR)*, Oct. 2016, pp. 1–5.

[44] R. Zhang and S. Cao, "Real-time human motion behavior detection via CNN using mmWave radar," *IEEE Sensors Lett.*, vol. 3, no. 2, Feb. 2019, Art. no. 3500104.

RENYUAN ZHANG received the B.S. degree from Chongqing University, Chongqing, China, in 2009, and the M.S. degree from The University of Arizona, Tucson, AZ, USA, in 2015, where he is currently pursuing the Ph.D. degree with the Department of Electrical and Computer Engineering. His research interests include mmWave radar signal processing, micro-Doppler signatures, sensor fusion, non-coherent integration, and SAR.

SIYANG CAO (S'11–M'15) received the B.S. degree in electronic and information engineering from Xidian University, Shanxi, China, in 2007, the M.S. degree in circuits and systems from the South China University of Technology, Guangdong, China, in 2010, and the Ph.D. degree in electrical engineering from The Ohio State University, Columbus, OH, USA, in 2014. Since August 2015, he has been an Assistant Professor with the Electrical and Computer Engineering Department, The University of Arizona, Tucson, AZ, USA. His research interests include radar waveform design, synthetic aperture radar, commercial radar, and signal processing with an emphasis on radar signal.