




Article

Adaptive Multi-Pedestrian Tracking by Multi-Sensor: Track-to-Track Fusion Using Monocular 3D Detection and MMW Radar

Yipeng Zhu ¹, Tao Wang ^{1,2,3,4,*}  and Shiqiang Zhu ^{1,5}¹ Ocean College, Zhejiang University, Zhoushan 316000, China; pheo@zju.edu.cn (Y.Z.); sqzhu@zju.edu.cn (S.Z.)² Engineering Research Center of Oceanic Sensing Technology and Equipment, Ministry of Education, Zhoushan 316000, China³ State Key Laboratory of Fluid Power and Mechatronic Systems, Zhejiang University, Hangzhou 310027, China⁴ Key Laboratory of Ocean Observation-Imaging Testbed of Zhejiang Province, Zhoushan 316000, China⁵ Zhejiang Lab, Hangzhou 311121, China

* Correspondence: twang001@zju.edu.cn

Abstract: Accurate and reliable tracking of multi-pedestrian is of great importance for autonomous driving, human-robot interaction and video surveillance. Since different scenarios have different best-performing sensors, sensor fusion perception plans are believed to have complementary modalities and be capable of handling situations which are challenging for single sensor. In this paper, we propose a novel track-to-track fusion strategy for multi-pedestrian tracking by using a millimeter-wave (MMW) radar and a monocular camera. Pedestrians are firstly tracked by each sensor according to the sensor characteristic. Specifically, the 3D monocular pedestrian detections are obtained by a convolutional neural network (CNN). The trajectory is formed by the tracking-by-detection approach, combined with Bayesian estimation. The measurement noise of the 3D monocular detection is modeled by a detection uncertainty value obtained from the same CNN, as an approach to estimate the pedestrian state more accurately. The MMW radar utilizes the track-before-detection method due to the sparseness of the radar features. Afterwards, the pedestrian trajectories are obtained by the proposed track-to-track fusion strategy, which can work adaptively under challenging weather conditions, low-illumination conditions and clutter scenarios. A group of tests are carried out to validate our pedestrian tracking strategy. Tracking trajectories and optimal sub-pattern assignment (OSPA) metric demonstrate the accuracy and robustness of the proposed multi-sensor multi-pedestrian tracking system.

Keywords: pedestrian tracking; sensor fusion; monocular 3D detection; MMW radar; track-to-track fusion



Citation: Zhu, Y.; Wang, T.; Zhu, S. Adaptive Multi-Pedestrian Tracking by Multi-Sensor: Track-to-Track Fusion Using Monocular 3D Detection and MMW Radar. *Remote Sens.* **2022**, *14*, 1837. <https://doi.org/10.3390/rs14081837>

Academic Editor: Andrzej Stateczny

Received: 7 March 2022

Accepted: 6 April 2022

Published: 11 April 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Pedestrian detection and tracking are fundamental tasks for a number of applications including autonomous vehicle and video surveillance. Specifically, information about object classes, locations as well as velocity are required in the process of environment perception. The more accurate and detailed the tracking results are, the easier for the system to manage motion planning. For example, autonomous vehicles need reliable pedestrian tracking results in high frame rate to finish path planning and avoid collision. Robots need accurate pedestrian tracking results to interact with users. With the significant progress in artificial neural network, 2D pedestrian detection and tracking in image plane have achieved satisfying performance and are almost regarded as solved problems. However, the 2D tracks of the pedestrians are far from adequate for demanding applications in a variety of challenging scenarios. Accurate 3D space localization is a more intuitive choice for system robustness and reliability. As a result, 3D pedestrian tracking with detection attracts more attention of researchers and becomes a more and more important task [1].

For multi-pedestrian scenario, detection and tracking are regarded as two separate but correlated tasks [2,3]. In a situation where the target features are sparse or missed detections occur frequently, the track-before-detection method is widely adopted. For example, a low resolution radar can only obtain limited reflections without convincing detections. The obtained tracks can help identify the target category by offering the history location as an additional feature. The lost of tracking targets and miss matches are prone to happen when sensors perform poorly or the targets number is large. As a result, the track-before-detection method is feasible when the target identification process is challenging.

In other situations, the features from sensors results are rich enough for detection task. The tracking-by-detection method can be adopted and the tracking results are more robust, compared to the former track-before-detection method. Since the detections are of high accuracy and the features are enough for target identification, the track of each target is unlikely to have miss matches and therefore has a higher continuity. From the detection results to the tracking results, the main difficulty is about properly associating detections across frames since the detection results may contain some false positive or missed targets. Conventional data association methods like GNN [4], JPDA [5], MHT [6] are widely used for tracking. The possible association proposals grow with target number and time steps. As a result, only part of the proposals are kept according to the cost of the computation complexity. After proper association, the pedestrian states are updated according to some Bayesian-based estimation methods [7–9]. Advanced tracking theories provide novel approaches [10,11]. Famous instances based on random finite set like PHD-filter [12], CPHD-filter and GM-PHD-filter are widely used when the object number in the scene is unknown and inconstant.

In order to obtain the pedestrian detection and tracking results, a list of sensors are adopted. Among them, the most popular and widely deployed choices are light detection and ranging (LiDAR), camera and radar. LiDARs are capable of providing accurate and dense point clouds, however they may fail under challenging weather conditions like strong sunlight or fog. In addition, LiDARs suffer from issues of high deployment cost and sensitivity to shocks.

The cameras can provide rich semantic information while the depth information for 3D detections cannot be directly measured. Stereo cameras recover the depth by matching point pairs from similar perspectives with known baseline. The accuracy is much influenced by the calibration result and thus causes dependency on a mild environment. Also the computational burden is high compared to other sensors. Another solution is the RGB-D camera. The depth is obtained by other components in the camera. However, the detection range is limited and the sensor is likely to be interfered in outdoors. For monocular camera, the depth information is estimated according to strong prior on camera extrinsic parameter and object size. Recent progress in convolutional neural network (CNN) as well as its derivatives have provided a new approach for monocular depth estimation. Therefore, a list of solutions for monocular 3D detection become feasible. Specially designed neural network for this task is currently widely studied and the performance is improving on a daily basis [13,14].

Progress in MMW radar has offered a new promising object detection option. The MMW radar can provide accurate localization results up to centimeter level [15]. At the same time, the velocity of the object is inferred from the Doppler information. The radar can handle challenging weather like foggy or dusty environment. However, MMW radars are short in angular detection resolution due to the nature of radar.

Since different kinds of sensors have strengths and weaknesses for different scenarios in terms of accuracy and robustness, multi-sensor schemes are proposed to handle sensor degradation under certain condition [16]. Specifically, the multi-modal fusion methods have been getting popularity nowadays [17]. Not only the perceptual field is enlarged, but also the data quality and reliability are increased. Less noise, less uncertainty and fewer deviations from the truth can be achieved if proper sensor fusion is implemented. LiDAR and RGB-D camera have the best accuracy, while the working range is limited [18].

Monocular camera can help LiDAR in terms of detection [19,20]. LiDAR with radar can enhance the output point clouds by increasing density or introducing velocity [21]. Radar and camera can provide accurate and robust fusion results at a reasonable cost and are thoroughly developed [22–32]. Utilizing all three types of sensors is also an option, at a cost of complexity [33]. Other plans have also been proposed, with the help of acoustic sensors [34] or the introduction of tags for targets [35].

In addition to sensor combination selection, fusion strategy also influences the multi-sensor tracking performance. Sensor fusion can be categorized into low-level raw data fusion, mid-level feature fusion and high-level track fusion [33]. When entering into higher level of fusion, the sensor data is selected and compressed. Therefore, the raw data fusion retains the most information while faces the highest processing difficulty. Compared to raw data, extracted features like bounding box from images and clusters from point clouds are more representative. Zhao et al. [20] use modality from one sensor to generate region of interest (ROI) for faster detection and more accurate clustering. Some researchers combine data or feature matrix from both sensors to one unified neural network for detection [36–40]. Compared to sequential fusion approach [32], track-to-track fusion method has the best flexibility and scalability. It not only has the least requirement for data bandwidth and computation source, but also supports specialized sensor models for specific sensors. The data collected by each sensor are used to detect and track the targets through a tracker which suits the sensor best [18,33,41].

Despite the fact that some fusion strategies have been proposed, multi-pedestrian tracking task remains challenging, especially in situations where sensor failures may occur. Though the track-to-track fusion strategy can reduce the missed detection rate to some extent, the occasional absence of the sensing modality may influence the fusion process and thus should be handled separately. The confidence of the monocular camera detection and tracking results is another issue to be solved. With proper modeling of the results from the individual sensors, the accuracy of the fusion results can be improved.

In this paper, we aim at multi-pedestrian challenging scenario, and propose a novel 3D pedestrian tracking framework based on a track-to-track fusion strategy by using MMW radar and monocular camera. The MMW radar follows the track-before-detection approach and provides reliable tracks against smoky or low-light interference. The monocular camera uses a CNN to generate pedestrian locations in Bird's eye view directly from 2D images. The track-by-detection approach is adopted to form pedestrian tracks with the help of Bayesian-based filters. The proposed track-to-track fusion strategy is utilized to leverage strengths from both sensors according to their complementarity. Experiments are carried out to validate the framework. It is shown that the system can work in different scenarios, including visually degraded situations, smoky situations and pedestrian clutter situations where the pedestrians are too close for the radar to distinguish. The main contributions are as follows:

- An improved 3D monocular multi-pedestrian tracking-by-detection method is implemented, with its measurement noise modeled by the detection uncertainty from the 3D pedestrian detection neural network.
- A novel track-to-track fusion strategy is proposed to integrate the pedestrian tracks obtained by MMW radar and monocular camera. The adaptive multi-pedestrian tracking strategy is able to automatically detect the occurrence and handle challenging weather condition, low-illumination condition and clutter situation. Also, the track-to-track fusion approach enables the pedestrians to be more accurately tracked by individual sensors before fusion.
- The performance of the proposed tracking strategy is compared in both normal and challenging scenarios using the optimal sub-pattern assignment (OSPA) metric. The superiority of the fusion approach is demonstrated both intuitively and numerically.

The remainder of this paper is organized as follows. In Section 2, some representative works on monocular 3D object detection, MMW radar tracking, and radar-camera based sensor fusion method are listed and explained. In Section 3, the details of our multi-

pedestrian detection and tracking strategy are presented, including how the pedestrians are tracked by individual sensors and how the track-to-track fusion is implemented. Experimental validation for the proposed strategy and comparison to other methods are shown in Section 4, followed by a comprehensive discussion part in Section 5. Conclusions and future research interest are in Section 6.

2. Related Work

2.1. Radar Pedestrian Tracking

Radar object tracking has been studied in depth. Since the features derived from radar signals are limited, the classification task is challenging [42,43]. Based on the dependency of detection results, radar tracking methods can be roughly separated in two directions.

The first category is the tracking-by-detection approach [3]. The tracks are formed by associating detection results across time. The radar detection step is challenging due to the nature of radar signal. Firstly, most tracking theories rely on the point object hypothesis. However, radars resolution in depth is high enough to form a set of point detections for the same object. The point cloud needs to be partitioned and clustered to form a unique detection for each target. Some researchers use K-means to get the centroids of each cluster while others use object shape and boundary to match the points. To handle situations when the total number of objects is unknown, DBSCAN method is adopted. Secondly, object classification and identification can be hard for the radars. Until recently, neural networks are adopted to obtain a satisfying performance [44]. For the data association process, it is challenging to assign new detections to existing tracks when features are limited. When multiple choices are available, the association complexity increases and no features can be used to reduce the difficulty. Though conventional methods including GNN, JPDA, MHT and advanced methods like PHD are adopted, the performance cannot match those association methods by features matching. Also, the algorithm complexity and computing burden is high [45,46].

The second category is the tracking-before-detection [47]. It is designed for extended object tracking. Multiple measurements may belong to the same object. The clustering, associating and filtering are implemented simultaneously. Different from the detection-association-filter approach, the tracking results rely much on the accuracy of the prediction step. When the motion model differs from the reality, the generated measurement boundary has a large chance to include wrong raw point cloud or miss the true point cloud. Afterwards, the trajectory can provide more information for higher identification accuracy. Some researches have been made, though the performance could be further improved [48,49].

For the aforementioned tracking methods, the pedestrian state is usually estimated by Bayesian methods. Popular options are Kalman filter, EKF, UKF and particle filter. The distribution assumption of the objects varies from simple to complex for different situations.

2.2. Monocular 3D Pedestrian Detection

Most works on pedestrian tracking for camera are limited in 2D RGB plane, for deriving depth information from single image is regarded as an ill-posed problem. The object detection task in the 2D image plane has been studied extensively. Some recent works have achieved satisfying performance and serve as a feature extractor for the 3D detection task. Based on the 2D detection results, in order to further find the 3D location or at least Bird's eye view location of the targets, geometry constraints and human height hypotheses are widely used [14]. Some previous works focus on the homography matrix from real world to image plane and assume the ground plane to be flat. Using pre-calibrated extrinsic parameters of the camera and assuming average human height as reference, the pedestrian location can be determined. However, these methods are valid only when targets stand straightly on the ground. In a walking scenario, the height varies with time, not to mention the strict calibration requirements and the corresponding errors. These prior-based methods rely on a strong assumption and are therefore less flexible in real situations.

With the development in deep learning, feature representing ability and model complexity are strong enough to form a neural network for 3D object detection directly from image. Learned paradigm has replaced hand-engineered features for proposal generating [13]. In recent works, 2D human pose estimations and 2D segmentation results from image are introduced to enhance the performance. The derived keypoints or contours are more suitable for state representing, compared to 2D bounding boxes. Another way to upgrade 2D detections to 3D space is to combine them with a pseudo depth image. Monocular depth estimation methods can estimate the depth for each pixel, so as to inversely project the 2D bounding boxes or segmented contours to 3D space. An alternative solution is to directly generate pseudo LiDAR point cloud and make use of point cloud 3D detection network [50]. By decoupling the 3D detection process into separate tasks, the sub-tasks can be tested and tuned more thoroughly to increase the interpretability. However, the total inference time would increase significantly due to the additional depth estimation step for each pixel. The training process of the neural network is also challenging since the ground truth depth for each pixel in image plane is hard to obtain.

The end-to-end monocular 3D detection methods, regarded as ultimate solutions, are widely explored in recent years. Chen et al. [41] follow the general 2D object detection pipeline by generating class-specific 3D bounding box proposals and scoring afterwards. Candidates are scored with a CNN by class semantic, instance semantic, context, shape features as well as location priors. Zhou et al. [51] propose an extrinsic free method by predicting camera parameters. The detection performance is exclusive from the camera perturbation and thus more robust. Potholed and uneven roads scenarios are therefore successfully handled. Hu et al. [52] can detect and track the target at the same time by integrating spatial feature learning and 3D state estimation. MonoGRNet [53] combines 2D object detection, instance depth estimation and 3D center localization subnetworks into one network. The unified network makes use of the results from the 4 task-specific networks to predict the depth and the poses of the targeting 3D bounding box. Though pixel-level depth estimation is avoided, the overall computational burden remains heavy. PoseCNN [54] estimates 6D object pose with 3D rotation and 3D translation. Occlusion and symmetric objects in cluttered scenes can be handled due to its new loss functions. Though most of the methods are designed for tracking of general objects or vehicles, features of pedestrians can also be learned if trained with proper ground truth. MonoLoco [55] focuses on the pedestrian detection task by extracting 2D human poses first. The 3D localization results are obtained through another fully-connected network. It also introduces confidence intervals to address the ambiguity problem in the task, which can be used as an additional feature of the detection results.

2.3. Fusion of Radar and Camera

Results from complementary sensors can be helpful in improving the pedestrian detection and tracking performance. Common sensor combinations are camera with LiDAR, camera with radar, Lidar with radar as well as camera with radar and LiDAR. Due to the high cost in the deployment of LiDAR, here we mainly discuss the radar and camera fusion method. As for the tracked targets, vehicles and pedestrians are of different sizes but high similarity. Fusion tracking strategies for both targets are collected and discussed.

Recent works by fusion of radar point clouds and camera images have improved accuracy and running speed for different tasks. Since the data from radar and camera are of different modalities, the fusion strategies can be grouped into two categories, namely the single-modality way and the multi-modality way.

The single-modality way firstly transforms the image detections and tracks into 3D space points, before fusing with the radar points. 3D monocular detections are generated based on different prior and assumptions. Tracks are formed after associating detections from radar and camera. Since the detections are of the same modality and format, the fusion process is simplified as a probabilistic problem. Otto et al. [22] present an improved data association filter to tackle the problem of pedestrian tracking across the blind region where

only radar is available. However, the adopted 3D monocular detection algorithm depends on histograms of gradients (HOG) features and radar detects all objects instead of focusing on pedestrians. Liu et al. [23] fuse information from radar and monocular vision by combining interacting multiple model with probabilistic data association. They adopt an asynchronous tracking system with low level data fusion to avoid loss of sensor information. Dimitrievski et al. [24] calculate the joint likelihood of radar and camera after the camera observations are back-projected on the ground plane. Both tracking-by-detection and tracking-before-detection are adopted in the particle update process, depending on the association results. The idea is innovative but clutter environments are not well handled. The fusion can also be applied after each object is detected and tracked. The track-to-track fusion approach has flexibility and scalability due to the sensor-specific processing pipeline. Lee et al. [33] propose a permutation matrix track association method to associate the object tracks provided by different trackers and use a sequential approach to update new observations by an unscented Kalman filter. The improvement in association is validated while the track fusion can be further developed. Zhong et al. [26] cluster the detection point cloud from radar sensors into ROI and use HOG for image classification. The fusion block tracks both in 3D space for radar and 2D image plane for camera. In addition, the 3D velocity synthesis recovers the associated 3D velocity by using partial information from each sensor. Kim et al. [27] compensate the low resolution of radar bearing angle by camera observation and adopt the integrated probabilistic data association to handle clutter environment.

The multi-modality way focuses on the association problem cross modalities. ROI is widely used to improve the detection performance. Radar points can be projected into image plane, indicating where the objects are likely to be. The image detection can also provide additional information to guide the radar detection process [28]. Cho et al. [29] make use of the complementary sensing modality by separating the task. Camera determines class of object and object size. Radar points provide object location update information. Tracking is improved by a switching mechanism of two motion models based on object distance. Bai et al. [56] also use radar for spatial position and camera for classification, while some improvements are made by adopting GM-PHD. Another way to handle multi-modality detections is using neural network. Nabati et al. [36] use the radar detections to generate radar-based feature maps to complement the image features. Features are fed into a trained network and the 3D object detection performance is improved. Wang et al. [37] follow a similar approach and generate ROI in time-frequency spectrum for neural network input. Zhang et al. [38] propose an end-to-end deep neural network for multi-modality tracking. Two neural networks are adopted as the image feature extractor and the point cloud feature extractor. The fusion module includes simple concatenating of features, linear combination as well as attention mechanism with weighted importance. The affinity estimator and the start-end estimator are used for adjacency before giving final results. All modules are combined in one framework, making it possible to be trained in an end-to-end manner for joint optimization.

A combination of the single-modality and multi-modality fusion is also possible. Wang et al. [40] compensate the vision processing by using ROI provided by the MMW radar. Afterwards, objects are tracked by each sensor in the form of single-modality 3D locations. Finally, the fusion is simplified as a verification process for false alarms since the camera uses an outdated edge symmetry detection algorithm. The components can be replaced by recent progress and the tracking system should be further customized. A summary of the mentioned fusion methods is given in Table 1.

Table 1. Summary of some tracking methods by sensor fusion with radar and camera.

Fusion Methods	Work	Highlights
Single-modality way	Otto et al. [22]	continue tracking across blind regions
	Liu et al. [23]	combine interacting multiple models
	Dimi et al. [24]	calculate joint likelihood of radar and camera
	Lee et al. [33]	update sequentially by all sensor observations
Multi-modality way	Nobis et al. [28]	use ROIs to guide detection by other sensors
	Cho et al. [29]	camera determines class, radar provides location
	Bai et al. [56]	detect target independently, track with GM-PHD
	Nabati et al. [36]	use radar feature maps to complement images
	Zhang et al. [38]	use e2e neural network with feature extractors
Combination way	Wang et al. [40]	use both ROIs and single-modality fusion

2.4. Tracking Evaluation Metrics

The tracking performance can be evaluated in different ways. For multi-pedestrian tracking problem, both the target number and the localization accuracy should be considered. Missed targets and false targets can cause the estimated number to be wrong or unmatched. The difficulty lies on the unifying of the localization error and cardinality mismatches, namely on how to quantify the performance by one index.

Classic metrics like Wasserstein metric lack a consistent physical interpretation [57]. Scoring methods proposed by Fridling and Drummond [58] can evaluate the performance of multiple target tracking algorithms fairly but the track to truth association needs to be further developed.

The OSPA metric [57] is considered to be the most popular metric for multi target tracking. OSPA optimally assigns all targets in the test set and the target set. Afterwards, it computes the localization error based on this assignment. The mismatches of the targets are modeled by a cardinality mismatch penalty. A combination value of the two parts, namely the localization error and the cardinality error, is used to evaluate the tracking performance.

Recent works advance the problem by adjusting the metric. The generalized optimal sub-pattern assignment (GOSPA) metric [59] penalizes localization errors for detected targets, missed targets and false targets in different ways. It emphasizes the usefulness of GOSPA by encouraging trackers to have few false and missed targets. However, the improvement compared to the original OSPA is limited.

When object size is also considered, the metric should be extended to bounding boxes error estimation. Conventional metrics based on CLEAR metrics [60], including MOTP, MOTA, are not able to describe the performance when false positive occurs. The metrics may also be influenced by the choice of confidence threshold. Integral metric named by AMOTA and scaled accuracy metric named by sAMOTA are proposed by [61] to standardize the evaluation of 3D multi object tracking. In this paper, pedestrian size is not considered and OSPA metric is selected for tracking evaluation because it balances the localization error, the cardinality error and the computational burden well.

3. Methods

3.1. MMW Radar Pedestrian Tracking

MMW radars are radars which transmit wavelength in millimeter range. A typical MMW radar operating at 76–81 GHz can have a centimeter-level range resolution. With multiple antennas working as transmitters and receivers, the system becomes a multiple-input multiple-output (MIMO) radar and gives a finer spatial resolution. Using frequency modulated continuous wave (FMCW) technology, the adopted radar transmits a frequency-modulated signal continuously and receives range, angle and velocity information.

For the range measurement, the calculated intermediate frequency (IF) signal is modeled as a sine wave. Afterwards, the distance from the object is derived by the IF frequency and phase using Equation (1).

$$\begin{aligned} Y_t &= A \sin(2\pi f_o t + \phi_0) \\ f_o &= \frac{2dS}{c} \\ \phi_0 &= \frac{4\pi d f_c}{c} \end{aligned} \quad (1)$$

where A is the signal amplitude, d is the distance, S is the slope of the chirp (rate of change of frequency), c is the speed of light and f_c is the start frequency.

Additionally, the range resolution is determined by the bandwidth swept by the chirp, equals to

$$d_{\text{res}} = \frac{c}{2B} \quad (2)$$

where c is the speed of light and B is the bandwidth swept by the chirp.

The velocity is measured according to Doppler effect and the velocity resolution is

$$v_{\text{res}} = \frac{\lambda}{2T_f} \quad (3)$$

where T_f is the frame time and λ is the signal wavelength.

For the angle estimation, at least two RX antennas are used to find the angle of arrival (AoA). Array length is properly designed and small θ approximation is adopted. As a result, the estimation accuracy decreases when θ increases. Similarly, the angle resolution achieves its best performance when θ equals to 90 degrees.

$$\theta_{\text{res}} = \frac{2}{N \cos(\theta)} \quad (4)$$

where N is the number of antennas in the array.

In order to make use of the height dimension, the MMW radar works in 3D detection mode. The radar can generate dense measurements without introducing extra clutters, comparing to the 2D mode. As a result, the radar can provide point cloud measurements with range, velocity, azimuth and elevation information.

Since the radar resolution is much smaller than the pedestrian size, each target generates multiple measurements which are spatially structured around the objects. Under this situation, extended object tracking has to be adopted. Traditionally, extended object tracking is implemented through spatial clustering and temporal filtering. The spatial clustering generates detection candidates. The filters are applied to smooth the associated candidates. However, due to the signal-to-noise ratio (SNR) nature of radar, the raw measurement which is generated after constant false alarm rate (CFAR) detection can be either too dense or too sparse. Both sides will influence the spatial clustering performance. It is also challenging for MMW radar to detect object boundary, which makes the classification accuracy of clusters not promising. As a result, we adopt the tracking-before-detection approach. By combining the spatial and temporal process, the tracker can achieve more stable performance. The proposed radar tracking algorithm includes (i) prediction, (ii) association with clustering, and (iii) update. It is shown in Figure 1 and explained in detail.

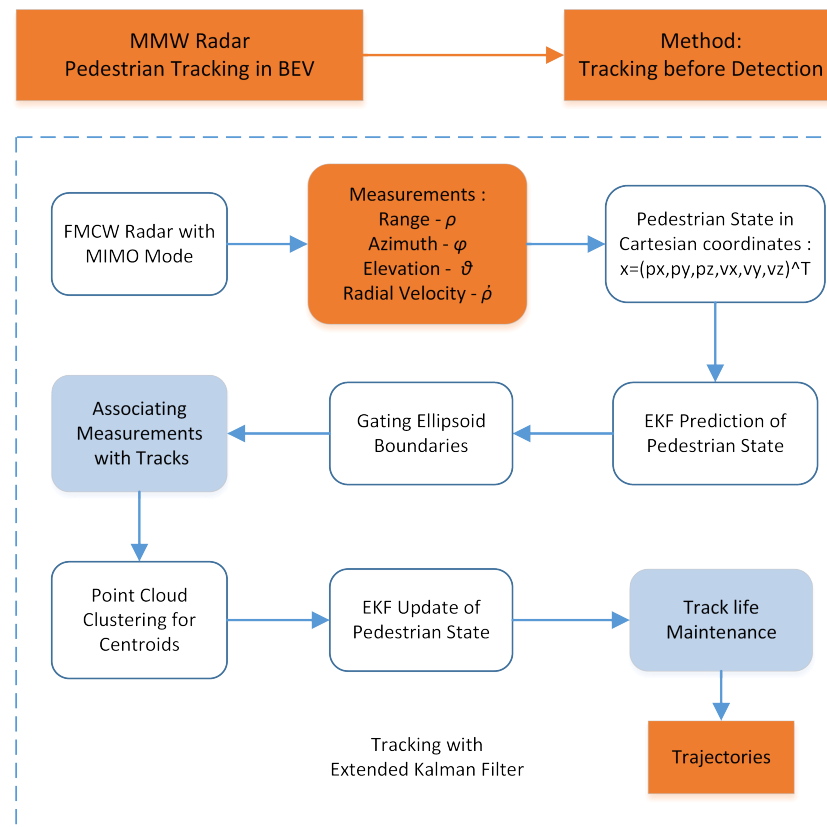


Figure 1. MMW radar pedestrian tracking algorithm.

The state of the pedestrian is modeled by the position and velocity in the OXYZ coordinate system, with p and u representing location and velocity, respectively.

$$\mathbf{x} = (p_x, p_y, p_z, u_x, u_y, u_z)^T \tag{5}$$

Based on the Chapman-Kolmogorov equation and Bayes' rule, the probability density function of the pedestrians can be described as

$$P(\mathbf{x}_k | \mathbf{z}_{1:k-1}) = \int P(\mathbf{x}_k | \mathbf{x}_{k-1})P(\mathbf{x}_{k-1} | \mathbf{z}_{1:k-1})\delta\mathbf{x}_{k-1}$$

$$P(\mathbf{x}_k | \mathbf{z}_{1:k}) = \frac{P(\mathbf{z}_k|\mathbf{x}_k)P(\mathbf{x}_k|\mathbf{z}_{1:k-1})}{\int P(\mathbf{z}_k|\mathbf{x}'_k)P(\mathbf{x}'_k|\mathbf{z}_{1:k-1})\delta\mathbf{x}'_k} \tag{6}$$

where subscript k represents the time step.

When the state variables are assumed to be Gaussian, the extended Kalman filter is implemented for the prediction and update step. The pedestrians are modeled by the constant velocity model, which assumes the target velocity to be constant during a measurement interval. The measurement \mathbf{z} contains distance ρ , azimuth ϕ , elevation θ , and radial velocity $\dot{\rho}$,

$$\mathbf{z} = (\rho, \phi, \theta, \dot{\rho})^T \tag{7}$$

Since in our coordinate system the azimuth ϕ of Y axis is considered as 0, the azimuth ϕ is calculated as the arc tangent of p_x over p_y . The state and measurement are modeled as,

$$\mathbf{x}_k = \mathbf{F}\mathbf{x}_{k-1} + \mathbf{w}_k \tag{8}$$

$$\mathbf{z}_k = \mathbf{H}(\mathbf{x}_k) + \mathbf{v}_k \tag{9}$$

$$F = \begin{bmatrix} 1 & 0 & 0 & \Delta T & 0 & 0 \\ 0 & 1 & 0 & 0 & \Delta T & 0 \\ 0 & 0 & 1 & 0 & 0 & \Delta T \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \tag{10}$$

$$H(x_k) = \begin{bmatrix} \sqrt{p_x^2 + p_y^2 + p_z^2} \\ \tan^{-1}\left(\frac{p_x}{p_y}\right) \\ \tan^{-1}\left(\frac{p_z}{\sqrt{p_x^2 + p_y^2}}\right) \\ \frac{p_x v_x + p_y v_y + p_z v_z}{\sqrt{p_x^2 + p_y^2 + p_z^2}} \end{bmatrix} \tag{11}$$

where F is the state transition matrix, H is the state observation matrix and P is the uncertainty matrix. Process noise w and measurement noise v follow

$$w_k \sim \mathcal{N}(0, Q), \quad v_k \sim \mathcal{N}(0, R) \tag{12}$$

The prediction step estimates the state and the covariance in the next time step,

$$\hat{x}_{k,k-1} = F \hat{x}_{k-1,k-1} \tag{13}$$

$$P_{k,k-1} = F_{k-1} P_{k-1,k-1} F_{k-1}^T + Q_{k-1} \tag{14}$$

Afterwards, each measurement in the next frame is associated with one pedestrian track. The association process requires a distance metric, which is computed by distance between the measurements and the predicted track centroids. The measurements are assigned to the track with the closest distance. For each track, the centroid of all measurements is calculated as the pedestrian location. Since MMW radar has limited angle resolution for boundary detection, the shape of pedestrian is modeled as a point.

For measurements not associated with any track, a new track is initialized for them if the SNR of the measurements is dense and strong enough. For tracks not associated with any measurements, track deletion is implemented.

Finally, the extended Kalman filter update step is used to estimate the state and covariance of the pedestrians. Due to the nonlinearity of the measurement model, the relation between state and measurement has to be approximated by Taylor series expansion.

$$H_k = \begin{bmatrix} \frac{p_x}{r} & \frac{p_y}{r} & 0 & 0 & 0 \\ \frac{p_x}{\sqrt{p_x^2 + p_y^2}} & -\frac{p_y}{\sqrt{p_x^2 + p_y^2}} & 0 & 0 & 0 \\ -\frac{p_x}{r^2} \frac{p_z}{\sqrt{p_x^2 + p_y^2}} & -\frac{p_y}{r^2} \frac{p_z}{\sqrt{p_x^2 + p_y^2}} & 0 & 0 & 0 \\ \frac{p_y(v_x p_y - v_y p_x) + p_z(v_x p_z - v_z p_x)}{r^3} & \frac{p_x(v_y p_x - v_x p_y) + p_z(v_y p_z - v_z p_y)}{r^3} & \frac{p_z}{r} & 0 & 0 \\ & & 0 & 0 & 0 \\ & & \frac{\sqrt{p_x^2 + p_y^2}}{r^2} & 0 & 0 \\ & & \frac{p_x(v_z p_x - v_x p_z) + p_y(v_z p_y - v_y p_z)}{r^3} & \frac{p_x}{r} & \frac{p_y}{r} & \frac{p_z}{r} \end{bmatrix} \tag{15}$$

where r equals to $\sqrt{p_x^2 + p_y^2 + p_z^2}$.

$$S_k = H_k P_k H_k^T + R_k \tag{16}$$

$$K_k = P_k H_k^T S_k^{-1} \tag{17}$$

$$\hat{\mathbf{x}}_{k,k} = \hat{\mathbf{x}}_{k,k-1} + \mathbf{K}_k(\mathbf{z}_k - \mathbf{H}(\hat{\mathbf{x}}_{k,k-1})) \quad (18)$$

$$\mathbf{P}_{k,k} = (\mathbf{I} - \mathbf{K}_k\mathbf{H}_k)\mathbf{P}_{k,k-1} \quad (19)$$

In order to evaluate the tracking performance, the 3D tracking results are translated to Bird's-eye view by omitting the p_z axis. By introducing the extra height dimension in the tracking process, the number of features for detection is increased. Therefore, the tracking performance is improved when compared to 2D working mode.

3.2. Monocular Vision Pedestrian Detection and Tracking

3.2.1. Bird's-Eye View Monocular Detection

Following the idea of MonoLoco [55], 3D pedestrian detection approach with confidence interval is adopted. The 3D detection process consists of two submodules, namely the 2D pose estimation neural network and the 3D location estimation neural network. The derived keypoints in image plane are used to represent human poses and are served as the input of the 3D localization neural network. To be specific, the 2D human pose keypoints are firstly estimated by pose detector PifPaf [62] and Mask R-CNN [63]. Afterwards, a fully-connected network with six linear layers can output the 3D center locations for each pedestrian. The estimated height of pedestrian is of less interest in reality and the result is translated to Bird's-eye view.

Apart from the pedestrian location centroids, MonoLoco is also able to evaluate the location uncertainty. Since the human height has a variation, the consequent localization error is inevitable. The ambiguity of the 3D detection task and the uncertainty due to noisy observation are both modeled as a probability distribution. Results on datasets show that around 84% of the pedestrians lie inside the predicted confidence intervals [55].

3.2.2. Tracking by Detection

Different from the MMW radar tracking, we adopt the tracking-by-detection fashion to find the tracks for monocular vision since the 3D detections are highly reliable. The algorithm is shown in Figure 2. For monocular Bird's-eye view detection, each pedestrian generates at most one single detection at each time step. It satisfies the point object hypothesis and the tracking can be realized through two steps, namely the data association step and the filter step.

The association step is implemented first. To assign newly obtained detections in new frame to previous tracks, features need to be designed, extracted and compared. Apart from location information, camera also has access to RGB data to form more discriminative features. As a result, associating detections from consecutive frames is finished by calculating the color and location distances between pedestrian detections.

After successful association, each track has a new measurement. A Kalman filter is utilized to merge the information from present and the past. The state vector of the pedestrian \mathbf{x} contains its location and velocity in Bird's-eye view. The measurement vector \mathbf{z} contains location alone.

$$\mathbf{x} = (p_x, p_y, u_x, u_y)^T \quad (20)$$

$$\mathbf{z} = (p_x, p_y)^T \quad (21)$$

Similar to the extended Kalman filter built for MMW radar tracking, a constant velocity motion model is adopted. The prediction and the update steps can be represented by Equations (13), (14) and (16)–(19), in which \mathbf{P} represents the state variance of camera detection, with different models of state transition matrix \mathbf{F} and state observation matrix \mathbf{H} ,

$$\mathbf{F} = \begin{bmatrix} 1 & 0 & \Delta T & 0 \\ 0 & 1 & 0 & \Delta T \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad (22)$$

$$\mathbf{H} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \quad (23)$$

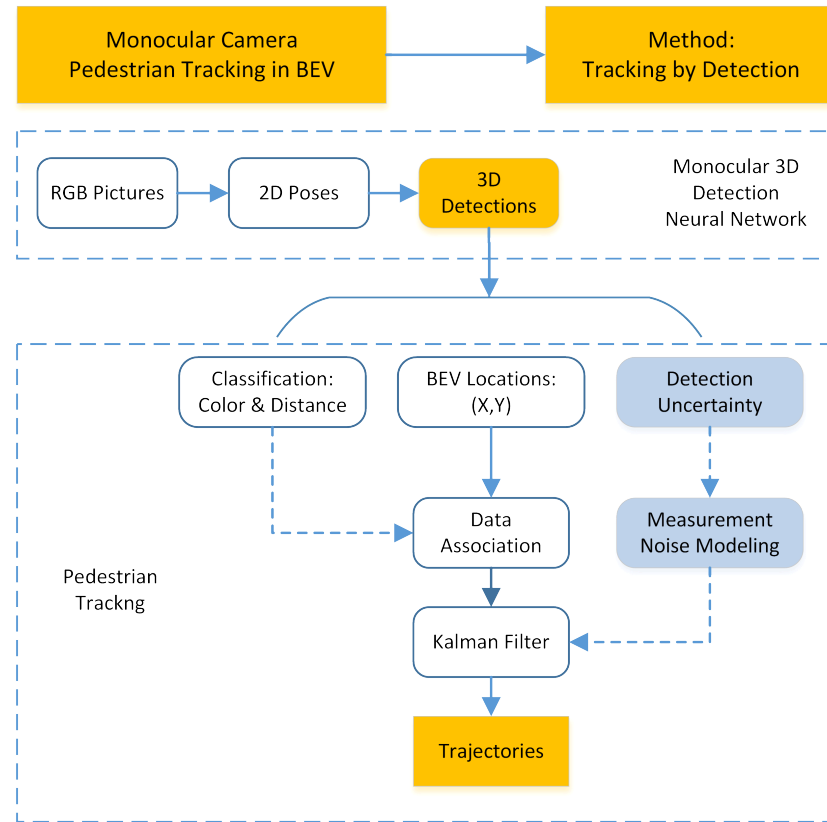


Figure 2. Monocular camera pedestrian tracking algorithm.

Here we propose a novel measurement noise modeling method by utilizing the obtained detection uncertainty. Different from the traditional way to assume that the measurement uncertainty is constant, we adopt the detection uncertainty to simulate the measurement uncertainty fluctuation. When the predicted pedestrian confidence interval is large, the detection uncertainty is believed to be high. The actual state has a larger probability to be away from the estimated centroid. As a result, the measurements should have a smaller weight during the Bayesian update step. To avoid unnecessary complexity, the measurement noise is still modeled as zero mean Gaussian distribution, while the variances are set to be linear to the detection uncertainty d^2 .

$$\mathbf{v}_k \sim \mathcal{N}(0, \mathbf{R}_k) \quad (24)$$

$$\mathbf{R}_k = \begin{bmatrix} \sigma_{p_x}^2 & 0 \\ 0 & \sigma_{p_y}^2 \end{bmatrix} = \begin{bmatrix} m_x d^2 + c_x & 0 \\ 0 & m_y d^2 + c_y \end{bmatrix} \quad (25)$$

The scale factor and the intercept (m_x , m_y , c_x and c_y) are determined by a calibration process. The noise variance is calculated by the detection result and the ground truth. By minimizing the distance between $(\sigma_{p_x}^2, \sigma_{p_y}^2)$ and $(m_x d^2 + c_x, m_y d^2 + c_y)$, the parameters are calibrated after proper fitting.

3.3. Track-to-Track Fusion Strategy

The track-to-track fusion strategy can be divided into two steps. Firstly, the transformation matrix between the two sensors is calibrated. The tracks from both sensors are therefore successfully aligned and associated. Secondly, the statuses of the sensors are examined according to track association results. When one sensor is believed to fail

or degrade, the other sensor will handle the situation according to the proposed fusion strategy. The overall fusion strategy is shown in Figure 3.

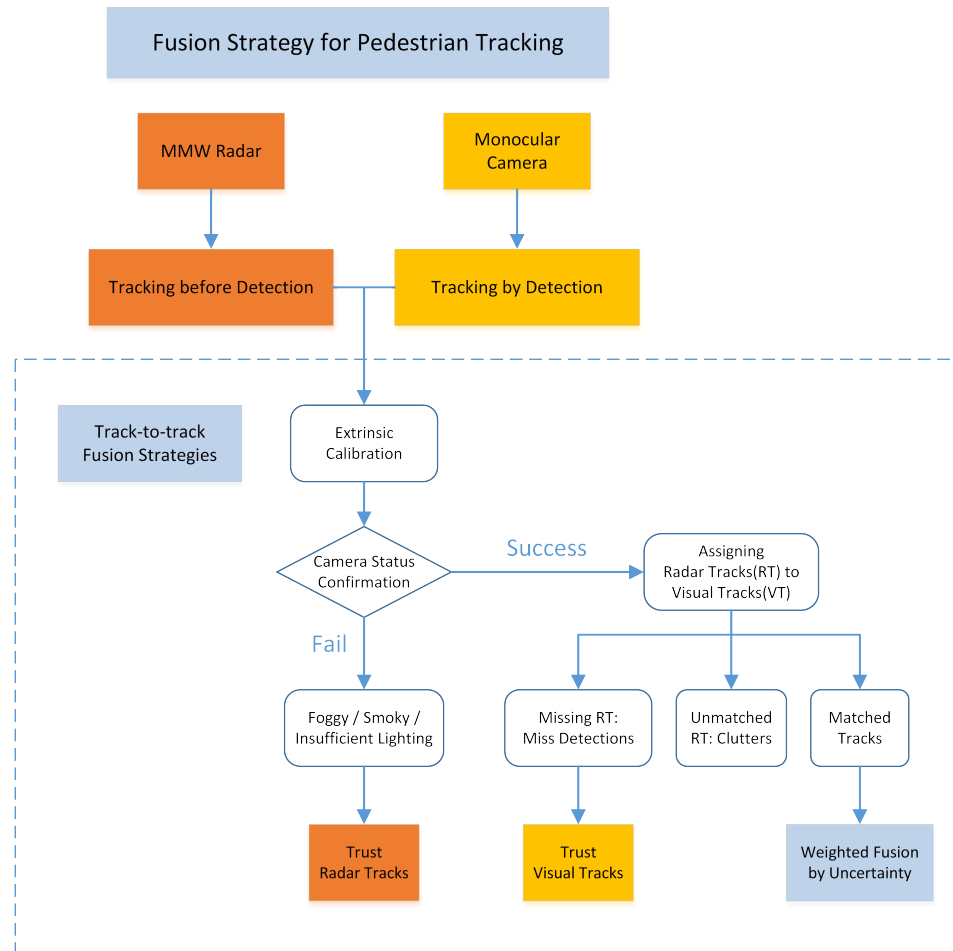


Figure 3. The proposed track-to-track fusion strategy.

3.3.1. Extrinsic Calibration

The previous results from MMW radar and monocular camera are from different coordinate system, so an extrinsic calibration for the two sensors is needed before fusion. The proposed monocular vision tracking method and the MMW radar tracking method are able to generate results in the bird's-eye view plane. As a result, the spatial transformation matrix is relatively simple compared to other visual tracking methods which are restricted in image plane. The centers of camera and radar are aligned and their headings are adjusted to the same direction. In this way, the transformation matrix consists of only three translation terms.

$$A = [R \quad T] = \begin{bmatrix} 1 & 0 & 0 & t_1 \\ 0 & 1 & 0 & t_2 \\ 0 & 0 & 1 & t_3 \end{bmatrix} \quad (26)$$

where R is the rotation matrix, T is the translation matrix, t_1, t_2 and t_3 represent the translations in three directions.

3.3.2. Fusion Strategy

The main idea of the fusion strategy is to complement the tracks according to the sensor characteristic. The monocular tracking result is trusted to provide the number of pedestrians due to its dominant detection rate. On the other hand, the MMW radar tracking result can provide more precise and accurate location information. Since the track-to-track fusion is based on tracking results from individual sensors and the trajectories before

fusion have already been smoothed by filters, the detections from sensors are more robust and accurate.

The fusion process firstly determines if the camera works well or not. When visual detections are confirmed, radar detections are assigned to vision detections by a distance gating and bidding method. All the unassociated radar detections are regarded as clutter. For all the vision detections that have been associated with one radar detection, a weighted fusion method is implemented. Missed detection of radar is confirmed when no radar detections are associated with the vision detection. In that situation, the vision tracking results are passed to the final results. The implementation details are shown in Algorithm 1.

Algorithm 1 Track-to-track fusion strategy for pedestrian tracking

Input: A_{ext} : extrinsic transformation matrix, T_{all} : time length of the input, $N_{radar,t}, N_{camera,t}$: number of tracks by radar or camera at time t , $X_{radar,i,t}$: pedestrian state (x, y) of the i th radar track at time t , $X_{camera,i,t}$: pedestrian state (x, y) of the i th camera track at time t , $P_{radar,i,t}$: state variance (P_x, P_y) of the i th radar track at time t , $P_{camera,i,t}$: state variance (P_x, P_y) of the i th camera track at time t , $Img_{threshold}$: the threshold to judge whether in fog, smoke or dark, $Dis_{threshold}$: the threshold to judge whether two tracks match.

Output: $X_{fusion,i,t}$: pedestrian state (x, y) of the i th track after fusion algorithm.

```

Initialize;
Coordinates alignment using  $A_{ext}$ ;
if variance of camera image <  $Img_{threshold}$  then
  Return  $X_{radar}$  ; // Trust radar if in fog, smoke or dark
else
  while  $t < T_{all}$  do
    for  $i \leftarrow 1$  to  $N_{camera,t}$  do
      for  $j \leftarrow 1$  to  $N_{radar,t}$  do
        find the closest  $X_{radar,j,t}$  to  $X_{camera,i,t}$ ;
      end
      if  $X_{radar,t} = \emptyset$  or distance >  $Dis_{threshold}$  then
         $X_{fusion,i,t} = X_{camera,i,t}$ ;
        ; // Trust camera for radar missed detections
      else
        remove  $X_{radar,j,t}$  from  $X_{radar,t}$ ;
        ; // Remove matched tracks and ignore clutter tracks
         $k_c = P_{radar,j,t} / (P_{camera,i,t} + P_{radar,j,t})$ ;
         $k_r = P_{camera,i,t} / (P_{camera,i,t} + P_{radar,j,t})$ ;
         $X_{fusion,i,t} = X_{camera,i,t} * k_c + X_{radar,j,t} * k_r$ ;
        ; // Try weighted fusion by state uncertainty
      end
    end
     $t \leftarrow t + \Delta t$ ;
  end
  Return  $X_{fusion}$  ; // Output fusion tracking results
end

```

3.4. Evaluation Metric of the Tracking Performance

The proposed strategy aims for multi-pedestrian tracking. Since the number of targets in the scene is not constant and not known, the detection correctness and localization accuracy should be both modeled meaningfully. A commonly used metric called OSPA is adopted. It combines cardinality error and state error to describe the difference between

two sets of vectors. It also solves the consistency and compatibility problem from previous metrics.

$$\begin{aligned} \bar{d}_p^{(c)}(X, Y) \\ = \left(\frac{1}{n} \left(\min_{\pi \in \Pi_n} \sum_{i=1}^m d^{(c)}(x_i, y_{\pi(i)})^p + c^p(n - m) \right) \right)^{1/p} \end{aligned} \quad (27)$$

where p is the order parameter and c is the cut-off parameter.

The value of p represents the sensitivity of the metric to outlier estimates. Following the advice in [57], we set a practical p value of 2 since it yields smooth distance curves. The cut-off c determines the penalty of cardinality error as opposed to localization error and is set constant during the test.

4. Experimental Results

4.1. Test Setup

Tests are taken out in an indoor scenario, with single or multiple pedestrians walking around. Sensor data are collected by our radar-camera fusion system and a Jetson NX Xavier device. The system consists of a Texas Instruments IWR 1843 MMW radar and a monocular camera. The camera has a 90 degree range of view with 1280×720 pixel resolution. To align the sampling speed, the camera and the MMW radar are both set to work at 15 Hz. The relative position of the two sensors is fixed and pre-calibrated. The specification of the MMW radar is given in Table 2.

Table 2. Detail specifications of the millimeter wave radar.

MMW Radar	Parameters
Working frequency	76–81 GHz
Max working range	15 m
Range resolution	0.09 m
Field of view	$\pm 50^\circ$
Azimuth resolution	15°
Update rate	15 Hz

The ground truth of the pedestrian trajectory is obtained by a ultra-wide band (UWB) locating system. The system consists of four pre-calibrated anchor stations and multiple tags. Pedestrians wear the tags while walking to transmit signals. Distances from tags to stations are measured and the exact locations are calculated according to triangulation theory. The adopted commercial UWB system works from 3.5 GHz to 6.5 GHz with bandwidth at 500 MHz. The frame rate is 15 Hz and the maximum detection range can reach 50 m. The test scene and the UWB system layout can be seen in Figure 4.

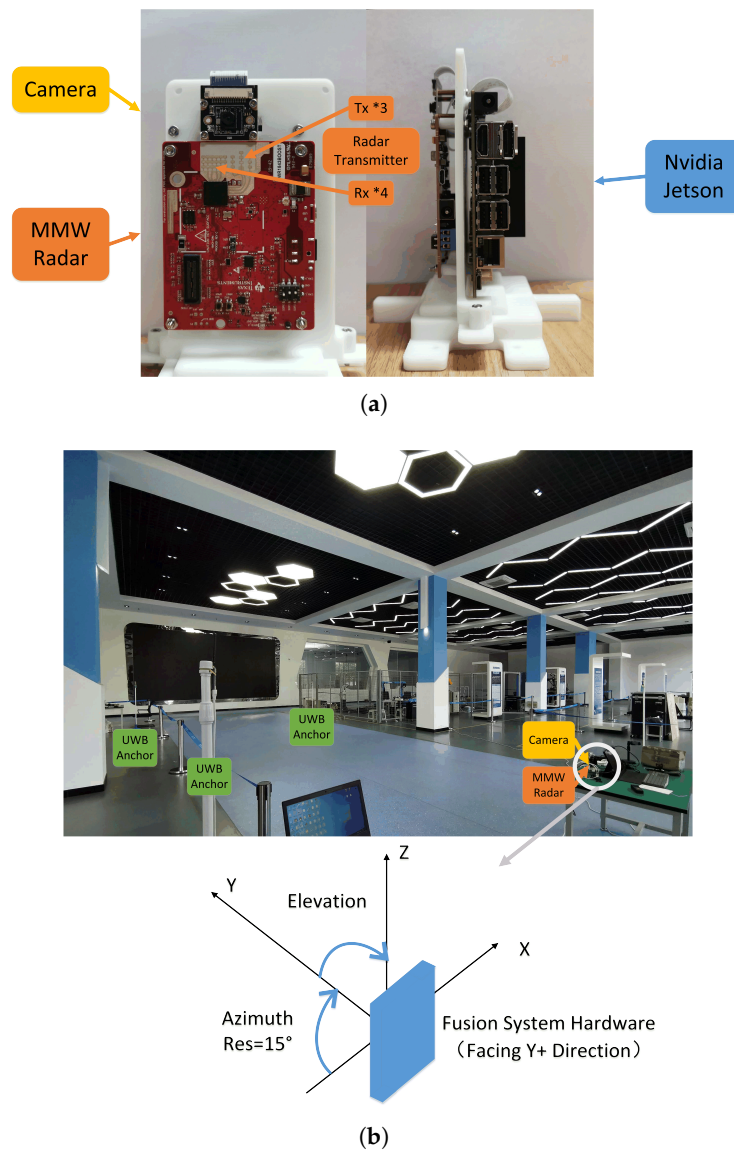


Figure 4. The test setup with camera, MMW radar and UWB system. (a) Fusion system layout. (b) UWB system layout and the test scene.

4.2. Monocular 3D Localization Noise Model Validation

The proposed monocular 3D pedestrian detection method can provide location and uncertainty information of the targets. The detection results are shown in Figure 5.

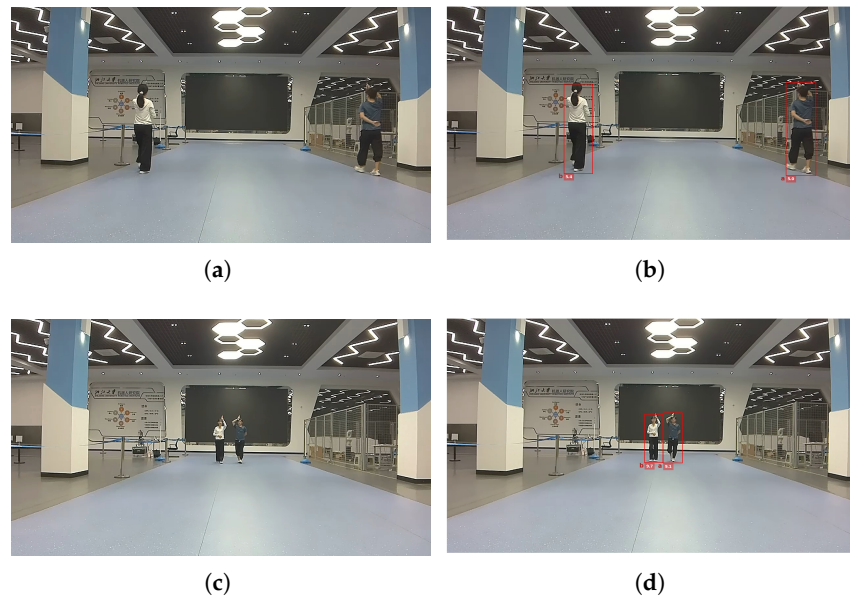


Figure 5. Monocular 3D pedestrian detection results. (a) Raw RGB capture example 1. (b) 3D detection example 1. (c) Raw RGB capture example 2. (d) 3D detection example 2.

We have assumed that the variances of the monocular location measurements are linear to the neural network detection uncertainty. To validate the proposed hypothesis, the differences between localization results from monocular detections and ground truth are collected. The localization variance is the fitting objective and the neural network detection uncertainties are the variables. The fitting results are shown in Figure 6. The m_x , m_y , c_x and c_y are therefore successfully calibrated.

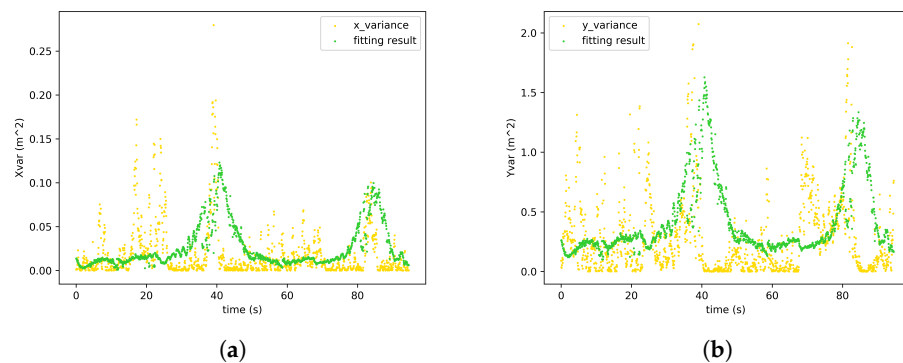


Figure 6. Calibration results for the measurement noise model for monocular 3D localization. (a) X variance. (b) Y variance.

4.3. Tracking Performance with the Proposed Fusion Strategy

Tracking accuracy and robustness tests are conducted in different situations where the number of pedestrians varies. The tests can be separated as single-pedestrian scenario and multiple-pedestrian scenario. To evaluate the performance analytically, the OSPA metrics of all the tracking methods are also given. Figure 7 illustrates an example result with single target in view. Pedestrian tracking trajectory from MMW radar, from monocular camera, from the proposed fusion strategy as well as from a state-of-the-art comparison by [32] are all shown. In addition, the change of OSPA metrics over time are given and compared. Similarly, tracking performance of the multiple-pedestrian scenario is shown in Figures 8 and 9. Each scenario contains three rounds of tests to improve representativeness. The mean OSPA values of different strategies in different scenarios are listed in Table 3.

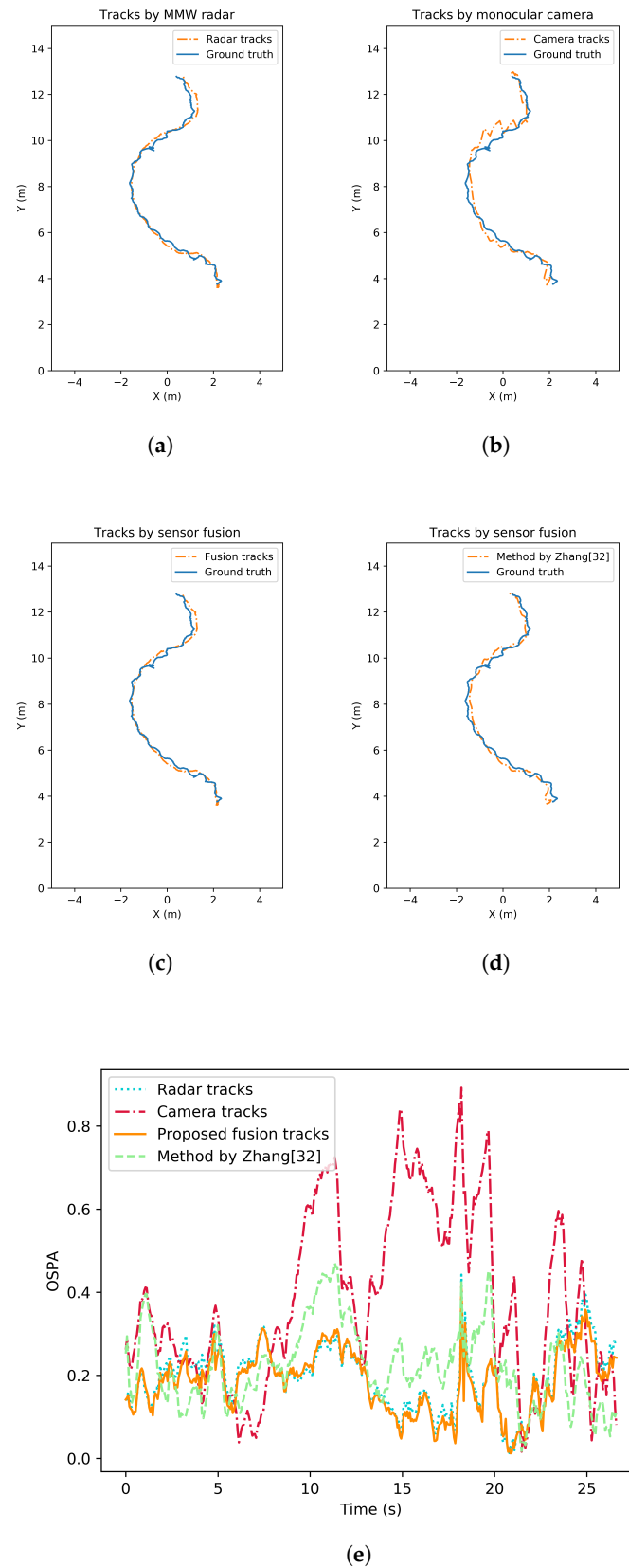


Figure 7. Pedestrian tracking performance comparison under single target situation. (a) Tracking trajectories by MMW radar. (b) Tracking trajectories by monocular camera. (c) Tracking trajectories by the proposed fusion strategy. (d) Tracking trajectories by the method from [32]. (e) Tracking performance comparison in OSPA.

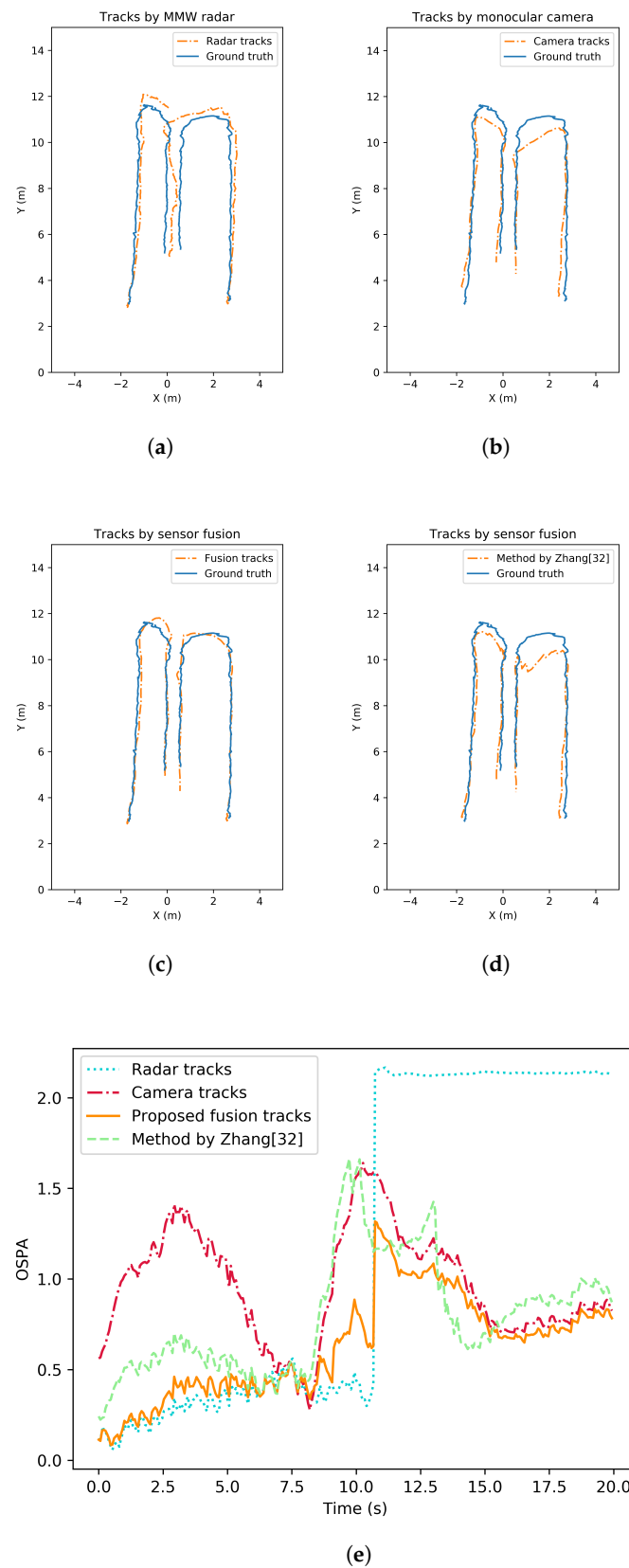


Figure 8. Pedestrian tracking performance comparison under multi targets situation. (a) Tracking trajectories by MMW radar. (b) Tracking trajectories by monocular camera. (c) Tracking trajectories by the proposed fusion strategy. (d) Tracking trajectories by the method from [32]. (e) Tracking performance comparison in OSPA.

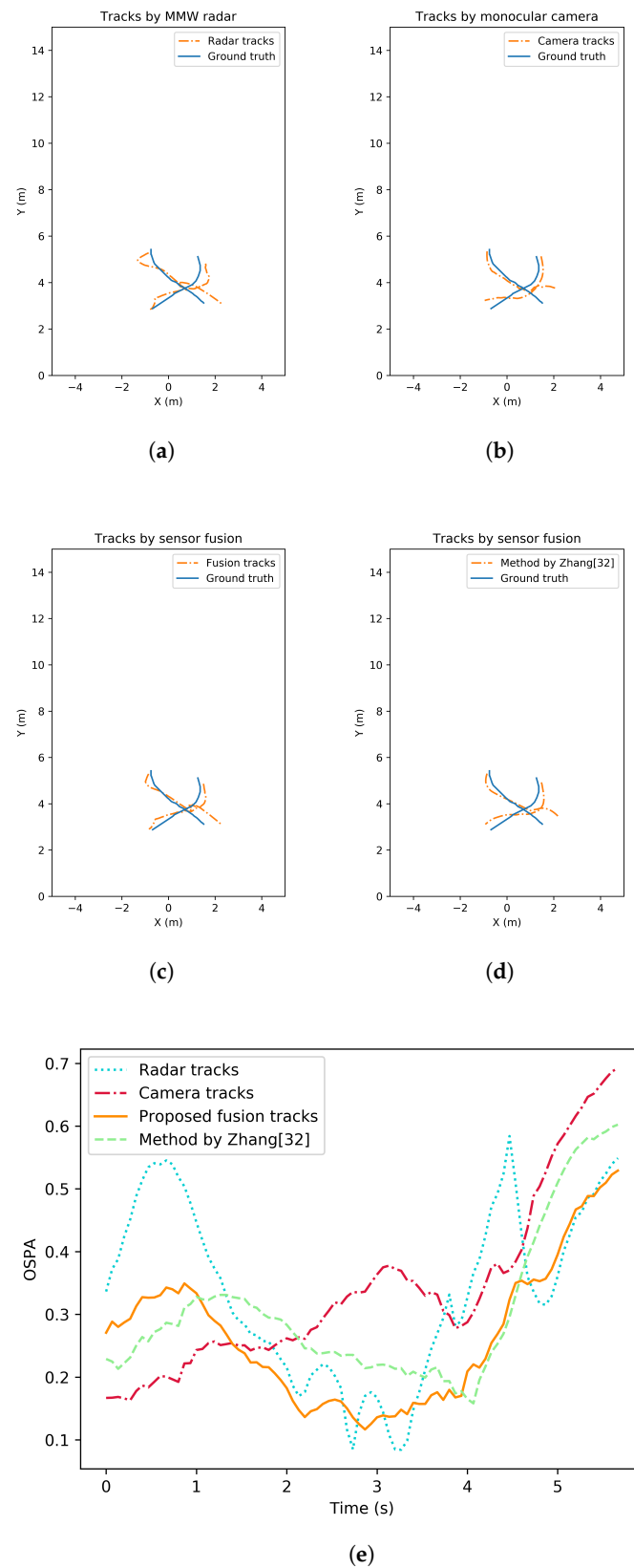


Figure 9. Pedestrian tracking performance comparison under crossover targets situation. (a) Tracking trajectories by MMW radar. (b) Tracking trajectories by monocular camera. (c) Tracking trajectories by the proposed fusion strategy. (d) Tracking trajectories by the method from [32]. (e) Tracking performance comparison in OSPA.

Table 3. The OSPA tracking performance of the proposed tracking method in different test scenarios.

Test Scenario and Fusion Strategy	OSPA
Single pedestrian	
MMW radar tracking with single pedestrian	0.197
Monocular camera 3D tracking with single pedestrian	0.383
Fusion by [32] with single pedestrian	0.222
Proposed fusion strategy with single pedestrian	0.185
Multiple pedestrians	
MMW radar tracking with multiple pedestrians	1.164
Monocular camera 3D tracking with multiple pedestrians	0.970
Fusion by [32] with multiple pedestrians	0.793
Proposed fusion strategy with multiple pedestrians	0.624

Extra tests are carried out to validate the adaptive fusion tracking strategy in camera-challenging scenarios. For example, the noise increases when the lighting condition gets worse. The view is blocked when the environment is smoky or foggy. The tracking results by our fusion strategy and the corresponding test situations are shown in Figure 10. The average OSPA metrics of the test in bad illumination and foggy environment are 0.269 and 0.223, respectively.

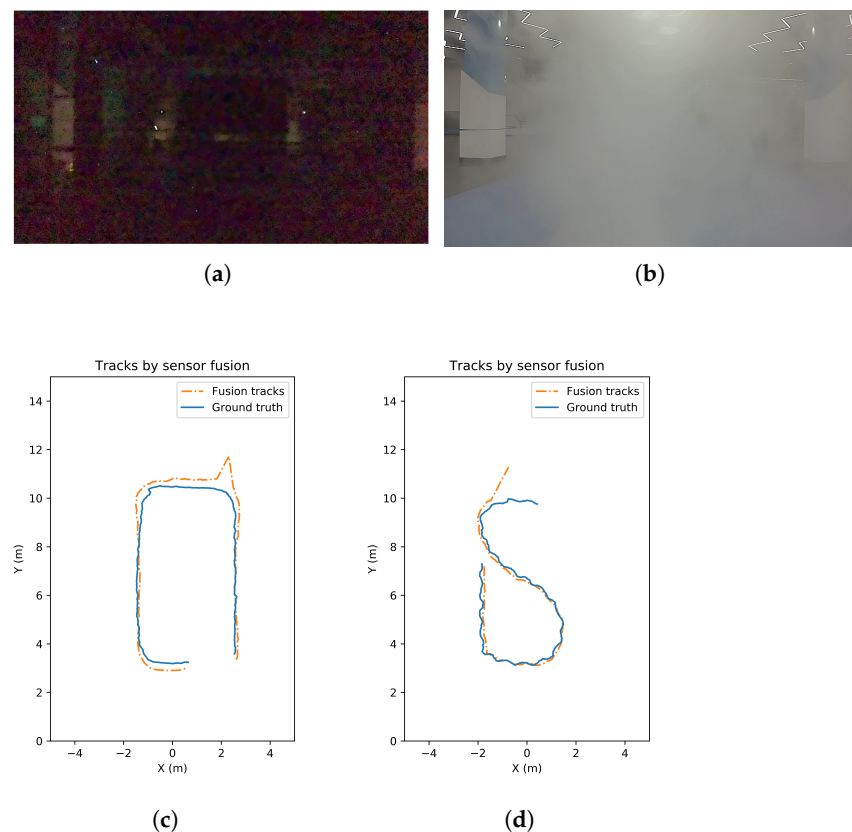
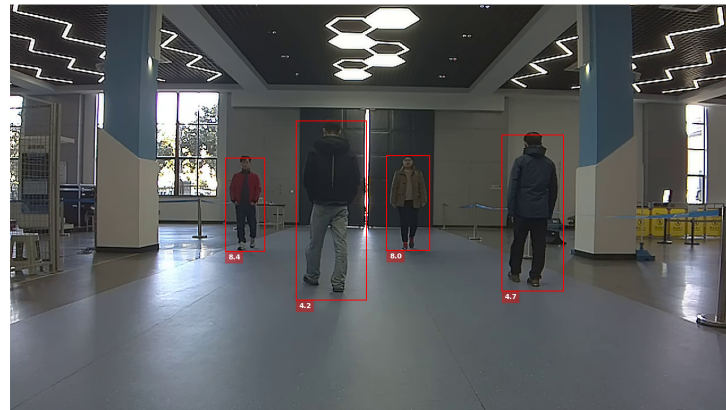


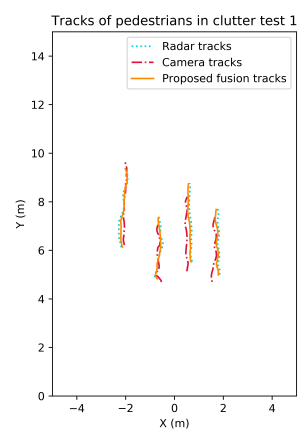
Figure 10. Camera-challenging scenarios and fusion tracking results. (a) Camera captures under bad illumination. (b) Camera captures in foggy environment. (c) Tracking results under bad illumination. (d) Tracking results under foggy environment.

In addition, a group of four tests are carried out to test the tracking feasibility in more clutter environment with more complex pedestrian characteristics. In this test, four pedestrians wander in the scene, each with different clothes colors. Also, the height of pedestrians varies from 1.60 m to 1.80 m for physical characteristic variety. Since the UWB

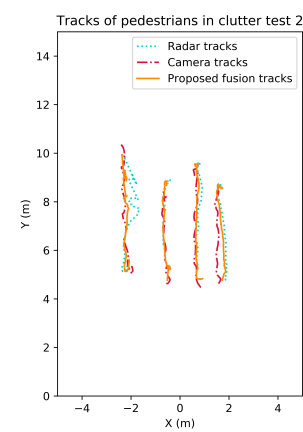
groundtruth collectors have a huge reduction of frame rate when work with numerous targets, the groundtruth is not collected here. The tracking trajectories by monocular camera, by MMW radar and by the proposed fusion strategy are given in Figure 11.



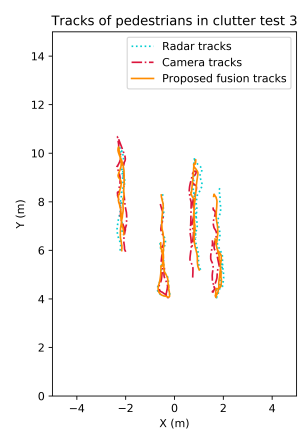
(a)



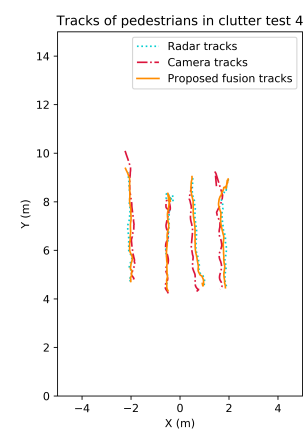
(b)



(c)



(d)



(e)

Figure 11. Pedestrian tracking performance of 4 tests by 4 tracking targets in clutter with different characteristics. (a) Camera captures of the test scene. (b) Tracking trajectories for test 1. (c) Tracking trajectories for test 2. (d) Tracking trajectories for test 3. (e) Tracking trajectories for test 4.

5. Discussion

The noise model validation tests show that the linear fitting model is consistent with the actual noise distribution derived from the observations. The noise variance is higher when the pedestrian has a larger uncertainty in neural network localization inference process. This is attributed to the representativeness of the modeled uncertainty. Described by a confidence interval, the uncertainty has the ability to quantitatively measure the detection quality. Therefore, the localization noise can be modeled as a Gaussian distribution with non-constant variances. And the variances can be modeled by a linear fit of the proposed confidence interval. The proposed noise modeling method is valid according to the results.

In the single-target pedestrian tracking test scenario, missed detection has a small probability of occurrence. OSPA of the proposed tracking strategy remains the lowest at most of the time. The mean value of OSPA across the test time also suggests the improvement of the proposed track-to-track fusion strategy compared to the sequential fusion approach by [32]. One possible reason is that the track-to-track fusion can implement sensor-specific tracking method with great flexibility and scalability. Tracks formed by each sensor are of higher reliability than a centralized sequential fusion method.

In the multi-target pedestrian tracking test, the performance of all tracking strategies drops due to the increased tracking difficulty. It can be seen that the radar performance degrades when pedestrians are relatively close or even intersect. Though the radar tracking method has reduced the number of clutter and missed detections to some extent, pedestrian cardinality is underestimated during the test. This may be caused by the nature of radar. When the gating SNR or radius is set too low, multiple clusters and detections can be generated by the same object. Clutter can also have a higher probability of occurrence. When the gating parameters are set too high, the strict confirmation process may lead to missed detections. This is shown in Figure 8 and the radar OSPA increased dramatically when cardinality is wrong. In contrast, the clutter and missed detections are less likely to happen in the monocular vision detection due to the high recall and precision of neural network. Though all targets are successfully tracked by the proposed monocular 3D tracking method, the localization error is high compared to radar. When complementing the tracking results from the two sensors, the fusion results reach lower OSPA. The sequential fusion strategy is intuitive and shows some improvements. However, the tracks filtered by each sensor can track targets more accurately. When missed detections and false alarms are handled by the adaptive track-to-track fusion strategy, the resulting tracks have a high accuracy in both location and cardinality estimation. According to the test results, the proposed track-to-track fusion strategy achieves around 46%, 36% and 21% OSPA reduction compared to the radar only, the camera only and the sequential fusion strategy.

For the camera-challenging scenarios, camera failure is detected successfully and confirmed robustly. Afterwards, detected camera failures are adaptively handled by the proposed fusion strategy during the test. The results show that our multi-sensor multi-pedestrian tracking system can work in the smoky, foggy or low-illumination situations and keep on providing reliable tracking results.

In the four pedestrian scenario, the tracking may fail when severe occlusion lasts. From the results, some deviations of the monocular tracks are observed. This may also be caused by detection cardinality error resulting from occlusion. The proposed method is able to deal with short-term occlusion but not frequent and long-term occlusion. For different pedestrian characteristics, the monocular detection accuracy is still reliable. The localization accuracy is influenced by the target height, but at an acceptable level.

To conclude, the overall performance of the proposed radar pedestrian tracking method shows its advantage in accuracy as well as its drawback in robustness. For the monocular tracking method, the OSPA remains at a reasonable while not satisfying level. The sequential fusion strategy by [32] has an improvement in terms of OSPA compared to single sensor. But the proposed track-to-track fusion strategy reaches the best performance and is validated in terms of both robustness and accuracy. It can also adaptively

handle scenarios of challenging lighting conditions, smoky conditions and pedestrians clutter situations.

6. Conclusions

In order to adaptively provide accurate and robust multi-pedestrian tracking performance, a track-to-track fusion strategy is proposed in our work. The MMW radar measurements are passed through a tracking-before-detection algorithm for pedestrian tracking. Monocular vision images are used to form pedestrian trajectory with the help of a 3D detection neural network and a Kalman filter whose noise modeling is improved. The radar tracking method has advantages in accuracy but drawbacks in robustness. The monocular 3D tracking method has reliable detections but less accurate localizations. Compared to radar, depth information cannot be directly measured by monocular camera. Though the adopted neural network method is able to detect pedestrian locations, its error is hard to explain and reduce. Distortion of the camera can also introduce some extra error with regard to the localization accuracy since the error in camera intrinsic parameter is hard to calibrate and changes with time. By complementing the two sensors, pedestrian tracking performance is improved compared to single sensor tracking or a state-of-the-art sequential fusion tracking. The proposed track-to-track fusion strategy is tested under various scenarios, including visual degradation situations with smoke and low-light environment. Results show that our camera-radar tracking system can work against these challenging interference with the proposed adaptive fusion strategy. Future work will be its implementation on an embedded system with real-time performance. A balance between high accuracy and high frame rate is to be sought.

Author Contributions: Conceptualization, Y.Z. and T.W.; methodology, Y.Z.; software, Y.Z.; validation, Y.Z., T.W.; formal analysis, Y.Z., T.W.; investigation, Y.Z.; resources, T.W.; data curation, Y.Z.; writing—original draft preparation, Y.Z.; writing—review and editing, T.W.; visualization, Y.Z.; supervision, S.Z., T.W.; project administration, T.W.; funding acquisition, T.W. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the research project of Robotics Institute, Zhejiang University under Grant K11804.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

LiDAR	Light Detection and Ranging
GNN	Global Nearest Neighbour
JPDA	Joint Probabilistic Data Association
MHT	Multiple Hypothesis Tracking
MMW	Millimeter Wave
PHD	Probability Hypothesis Density
GM-PHD	Gaussian Mixture Probability Hypothesis Density
CNN	Convolutional Neural Network
ROI	Region of Interest
EKF	Extended Kalman filter
UKF	Unscented Kalman filter
HOG	Histograms of Gradients
OSPA	Optimal Sub-Pattern Assignment
GOSPA	Generalized Optimal Sub-Pattern Assignment
MOTP	Multiple Object Tracking Precision
MOTA	Multiple Object Tracking Accuracy

MIMO	Multiple-input multiple-output
FMCW	Frequency Modulated Continuous Wave
IF	Intermediate Frequency
SNR	Signal-to-noise ratio
CFAR	Constant False Alarm Rate
BEV	Bird's Eye View

References

- Vo, B.N.; Mallick, M.; Bar-Shalom, Y.; Coraluppi, S.; Osborne, R.; Mahler, R.; Vo, B.T. Multitarget tracking. In *Wiley Encyclopedia of Electrical and Electronics Engineering*; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 2015.
- Feichtenhofer, C.; Pinz, A.; Zisserman, A. Detect to Track and Track to Detect. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 3057–3065. [\[CrossRef\]](#)
- Andriluka, M.; Roth, S.; Schiele, B. People-tracking-by-detection and people-detection-by-tracking. In Proceedings of the 26th IEEE Conference on Computer Vision and Pattern Recognition, CVPR, Anchorage, AK, USA, 24–26 June 2008. [\[CrossRef\]](#)
- Radosavljevic, Z. A study of a target tracking method using Global Nearest Neighbor algorithm. *Vojnoteh. Glas.* **2006**, *54*, 160–167. [\[CrossRef\]](#)
- Fortmann, T.E.; Bar-Shalom, Y.; Scheffe, M. Sonar Tracking of Multiple Targets Using Joint Probabilistic Data Association. *IEEE J. Ocean. Eng.* **1983**, *8*, 173–184. [\[CrossRef\]](#)
- Reid, D.B. An Algorithm for Tracking Multiple Targets. *IEEE Trans. Autom. Control* **1979**, *24*, 843–854. [\[CrossRef\]](#)
- Kalman, R.E. A new approach to linear filtering and prediction problems. *J. Fluids Eng. Trans. ASME* **1960**, *82*, 35–45. [\[CrossRef\]](#)
- Julier, S.J.; Uhlmann, J.K. New extension of the Kalman filter to nonlinear systems. In *Signal Processing, Sensor Fusion, and Target Recognition VI*; International Society for Optics and Photonics: Bellingham, WA, USA, 1997; Volume 3068, pp. 182–193. [\[CrossRef\]](#)
- Welch, G.; Bishop, G. *An Introduction to the Kalman Filter*; Department of Computer Science, University of North Carolina: Chapel Hill, NC, USA, 2006.
- Blackman, S.S.; Popoli, R. *Design and Analysis of Modern Tracking Systems*; Artech House: Norwood, MA, USA, 1999.
- Blackman, S.S. Multiple hypothesis tracking for multiple target tracking. *IEEE Aerosp. Electron. Syst. Mag.* **2004**, *19*, 5–18. [\[CrossRef\]](#)
- Mahler, R.P. Multitarget Bayes Filtering via First-Order Multitarget Moments. *IEEE Trans. Aerosp. Electron. Syst.* **2003**, *39*, 1152–1178. [\[CrossRef\]](#)
- Arnold, E.; Al-Jarrah, O.Y.; Dianati, M.; Fallah, S.; Oxtoby, D.; Mouzakitis, A. A Survey on 3D Object Detection Methods for Autonomous Driving Applications. *IEEE Trans. Intell. Transp. Syst.* **2019**, *20*, 3782–3795. [\[CrossRef\]](#)
- Qian, R.; Lai, X.; Li, X. 3D Object Detection for Autonomous Driving: A Survey. *arXiv* **2021**, arXiv:2106.10823.
- Van Berlo, B.; Elkelany, A.; Ozcelebi, T.; Meratnia, N. Millimeter Wave Sensing: A Review of Application Pipelines and Building Blocks. *IEEE Sens. J.* **2021**, *21*, 10332–10368. [\[CrossRef\]](#)
- Wang, Z.; Wu, Y.; Niu, Q. Multi-Sensor Fusion in Automated Driving: A Survey. *IEEE Access* **2020**, *8*, 2847–2868. [\[CrossRef\]](#)
- Mahler, R.P. *Statistical Multisource-Multitarget Information Fusion*; Artech House, Inc.: Norwood, MA, USA, 2007.
- Song, H.; Choi, W.; Kim, H. Robust Vision-Based Relative-Localization Approach Using an RGB-Depth Camera and LiDAR Sensor Fusion. *IEEE Trans. Ind. Electron.* **2016**, *63*, 3725–3736. [\[CrossRef\]](#)
- Samal, K.; Kumawat, H.; Saha, P.; Wolf, M.; Mukhopadhyay, S. Task-driven RGB-Lidar Fusion for Object Tracking in Resource-Efficient Autonomous System. *IEEE Trans. Intell. Veh.* **2021**, *8858*, 1–11. [\[CrossRef\]](#)
- Zhao, X.; Sun, P.; Xu, Z.; Min, H.; Yu, H. Fusion of 3D LIDAR and Camera Data for Object Detection in Autonomous Vehicle Applications. *IEEE Sens. J.* **2020**, *20*, 4901–4913. [\[CrossRef\]](#)
- Yang, B.; Guo, R.; Liang, M.; Casas, S.; Urtasun, R. RadarNet: Exploiting Radar for Robust Perception of Dynamic Objects. In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*; Springer: Cham, Switzerland, 2020; Volume 12363 LNCS, pp. 496–512. [\[CrossRef\]](#)
- Otto, C.; Gerber, W.; León, F.P.; Wirmitzer, J. A joint integrated probabilistic data association filter for pedestrian tracking across blind regions using monocular camera and radar. In Proceedings of the IEEE Intelligent Vehicles Symposium, Madrid, Spain, 3–7 June 2012; pp. 636–641. [\[CrossRef\]](#)
- Liu, F.; Sparbert, J.; Stiller, C. IMMPDA vehicle tracking system using asynchronous sensor fusion of radar and vision. In Proceedings of the IEEE Intelligent Vehicles Symposium, Eindhoven, The Netherlands, 4–6 June 2008. [\[CrossRef\]](#)
- Dimitrievski, M.; Jacobs, L.; Veelaert, P.; Philips, W. People Tracking by Cooperative Fusion of RADAR and Camera Sensors. In Proceedings of the 2019 IEEE Intelligent Transportation Systems Conference, ITSC, Auckland, New Zealand, 27–30 October 2019; Volume 2019, pp. 509–514. [\[CrossRef\]](#)
- Xu, D.; Anguelov, D.; Jain, A. PointFusion: Deep Sensor Fusion for 3D Bounding Box Estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 244–253.
- Zhong, Z.; Liu, S.; Mathew, M.; Dubey, A. Camera radar fusion for increased reliability in ADAS applications. In Proceedings of the IS and T International Symposium on Electronic Imaging Science and Technology, Burlingame, CA, USA, 28 January–1 February 2018; pp. 1–4. [\[CrossRef\]](#)

27. Kim, D.Y.; Jeon, M. Data fusion of radar and image measurements for multi-object tracking via Kalman filtering. *Inf. Sci.* **2014**, *278*, 641–652. [[CrossRef](#)]
28. Nobis, F.; Geisslinger, M.; Weber, M.; Betz, J.; Lienkamp, M. A Deep Learning-based Radar and Camera Sensor Fusion Architecture for Object Detection. In Proceedings of the 2019 Symposium on Sensor Data Fusion: Trends, Solutions, Applications, SDF 2019, Bonn, Germany, 15–17 October 2019. [[CrossRef](#)]
29. Cho, H.; Seo, Y.W.; Kumar, B.V.; Rajkumar, R.R. A multi-sensor fusion system for moving object detection and tracking in urban driving environments. In Proceedings of the IEEE International Conference on Robotics and Automation, Hong Kong, China, 31 May–7 June 2014; pp. 1836–1843. [[CrossRef](#)]
30. Chen, B.; Pei, X.; Chen, Z. Research on target detection based on distributed track fusion for intelligent vehicles. *Sensors* **2020**, *20*, 56. [[CrossRef](#)]
31. Liu, Z.; Cai, Y.; Wang, H.; Chen, L.; Gao, H.; Jia, Y.; Li, Y. Robust Target Recognition and Tracking of Self-Driving Cars With Radar and Camera Information Fusion Under Severe Weather Conditions. *IEEE Trans. Intell. Transp. Syst.* **2021**, 1–14. [[CrossRef](#)]
32. Zhang, R.; Cao, S. Extending reliability of mmwave radar tracking and detection via fusion with camera. *IEEE Access* **2019**, *7*, 137065–137079. [[CrossRef](#)]
33. Lee, K.H.; Kanzawa, Y.; Derry, M.; James, M.R. Multi-Target Track-to-Track Fusion Based on Permutation Matrix Track Association. In Proceedings of the IEEE Intelligent Vehicles Symposium, Changshu, China, 26–30 June 2018; pp. 465–470. [[CrossRef](#)]
34. Remmas, W.; Chemori, A.; Kruusmaa, M. Diver tracking in open waters: A low-cost approach based on visual and acoustic sensor fusion. *J. Field Robot.* **2021**, *38*, 494–508. [[CrossRef](#)]
35. Tian, Q.; Wang, K.I.; Salcic, Z. An INS and UWB fusion-based gyroscope drift correction approach for indoor Pedestrian tracking. *Sensors* **2020**, *20*, 4476. [[CrossRef](#)]
36. Nabati, R.; Qi, H. CenterFusion: Center-based radar and camera fusion for 3d object detection. In Proceedings of the 2021 IEEE Winter Conference on Applications of Computer Vision, WACV 2021, Waikoloa, HI, USA, 3–8 January 2021; pp. 1526–1535. [[CrossRef](#)]
37. Wang, Z.; Miao, X.; Huang, Z.; Luo, H. Research of target detection and classification techniques using millimeter-wave radar and vision sensors. *Remote Sens.* **2021**, *13*, 1064. [[CrossRef](#)]
38. Zhang, W.; Zhou, H.; Sun, S.; Wang, Z.; Shi, J.; Loy, C.C. Robust multi-modality multi-object tracking. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Korea, 27–28 October 2019. [[CrossRef](#)]
39. Chang, S.; Zhang, Y.; Zhang, F.; Zhao, X.; Huang, S.; Feng, Z.; Wei, Z. Spatial attention fusion for obstacle detection using mmwave radar and vision sensor. *Sensors* **2020**, *20*, 956. [[CrossRef](#)] [[PubMed](#)]
40. Wang, X.; Xu, L.; Sun, H.; Xin, J.; Zheng, N. On-Road Vehicle Detection and Tracking Using MMW Radar and Monovision Fusion. *IEEE Trans. Intell. Transp. Syst.* **2016**, *17*, 2075–2084. [[CrossRef](#)]
41. Chen, X.; Kundu, K.; Zhang, Z.; Ma, H.; Fidler, S.; Urtasun, R. Monocular 3D Object Detection for Autonomous Driving. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 26 June–1 July 2016; pp. 2147–2156. [[CrossRef](#)]
42. John, V.; Mita, S. RVNet: Deep Sensor Fusion of Monocular Camera and Radar for Image-Based Obstacle Detection in Challenging Environments. In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*; Springer: Cham, Switzerland, 2019; Volume 11854 LNCS; pp. 351–364. [[CrossRef](#)]
43. Cai, X.; Giallorenzo, M.; Sarabandi, K. Machine Learning-Based Target Classification for MMW Radar in Autonomous Driving. *IEEE Trans. Intell. Veh.* **2021**, *6*, 678–689. [[CrossRef](#)]
44. Pegoraro, J.; Rossi, M. Real-Time People Tracking and Identification from Sparse mm-Wave Radar Point-Clouds. *IEEE Access* **2021**, *9*, 78504–78520. [[CrossRef](#)]
45. Lim, H.S.; Park, H.M.; Lee, J.E.; Kim, Y.H.; Lee, S. Lane-by-lane traffic monitoring using 24.1 ghz fmcw radar system. *IEEE Access* **2021**, *9*, 14677–14687. [[CrossRef](#)]
46. Held, P.; Steinhauser, D.; Koch, A.; Brandmeier, T.; Schwarz, U.T. A Novel Approach for Model-Based Pedestrian Tracking Using Automotive Radar. *IEEE Trans. Intell. Transp. Syst.* **2021**, 1–14. [[CrossRef](#)]
47. Davey, S.J.; Rutten, M.G.; Cheung, B. A Comparison of Detection Performance for Several Track-before-Detect Algorithms. *EURASIP J. Adv. Signal Process.* **2007**, *2008*, 1–10. [[CrossRef](#)]
48. Zhao, P.; Lu, C.X.; Wang, J.; Chen, C.; Wang, W.; Trigoni, N.; Markham, A. MID: Tracking and identifying people with millimeter wave radar. In Proceedings of the 15th Annual International Conference on Distributed Computing in Sensor Systems, DCOSS 2019, Santorini Island, Greece, 29–31 May 2019; pp. 33–40. [[CrossRef](#)]
49. Fiscante, N.; Addabbo, P.; Clemente, C.; Biondi, F.; Giunta, G.; Orlando, D. A track-before-detect strategy based on sparse data processing for air surveillance radar applications. *Remote Sens.* **2021**, *13*, 662. [[CrossRef](#)]
50. Weng, X.; Kitani, K. Monocular 3D object detection with pseudo-LiDAR point cloud. In Proceedings of the 2019 International Conference on Computer Vision Workshop, ICCVW 2019, Seoul, Korea, 27–28 October 2019; pp. 857–866. [[CrossRef](#)]
51. Zhou, Y.; He, Y.; Zhu, H.; Wang, C.; Li, H.; Jiang, Q. Monocular 3D Object Detection: An Extrinsic Parameter Free Approach. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 7556–7566.

52. Hu, H.N.; Cai, Q.Z.; Wang, D.; Lin, J.; Sun, M.; Kraehenbuehl, P.; Darrell, T.; Yu, F. Joint monocular 3D vehicle detection and tracking. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Korea, 27–28 October 2019; pp. 5389–5398. [[CrossRef](#)]
53. Qin, Z.; Wang, J.; Lu, Y. Monogrnet: A geometric reasoning network for monocular 3D object localization. In Proceedings of the 33rd AAAI Conference on Artificial Intelligence, AAAI 2019, 31st Innovative Applications of Artificial Intelligence Conference, IAAI 2019 and the 9th AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Sanur, Bali, Indonesia, 8–12 December 2019; pp. 8851–8858. [[CrossRef](#)]
54. Xiang, Y.; Schmidt, T.; Narayanan, V.; Fox, D. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. *arXiv* **2017**, arXiv:1711.00199.
55. Bertoni, L.; Kreiss, S.; Alahi, A. MonoLoco: Monocular 3D pedestrian localization and uncertainty estimation. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Korea, 27–28 October 2019; pp. 6860–6870. [[CrossRef](#)]
56. Bai, J.; Li, S.; Huang, L.; Chen, H. Robust Detection and Tracking Method for Moving Object Based on Radar and Camera Data Fusion. *IEEE Sens. J.* **2021**, *21*, 10761–10774. [[CrossRef](#)]
57. Schuhmacher, D.; Vo, B.T.; Vo, B.N. A consistent metric for performance evaluation of multi-object filters. *IEEE Trans. Signal Process.* **2008**, *56*, 3447–3457. [[CrossRef](#)]
58. Fridling, B.E.; Drummond, O.E. Performance evaluation methods for multiple-target-tracking algorithms. In *Signal and Data Processing of Small Targets 1991*; International Society for Optics and Photonics: Bellingham, WA, USA, 1991; Volume 1481, pp. 371–383.
59. Rahmathullah, A.S.; Garcia-Fernandez, A.F.; Svensson, L. Generalized optimal sub-pattern assignment metric. In Proceedings of the 20th International Conference on Information Fusion, Fusion 2017—Proceedings, Xi’an, China, 10–13 July 2017; pp. 1–8. [[CrossRef](#)]
60. Bernardin, K.; Stiefelhagen, R. Evaluating multiple object tracking performance: The CLEAR MOT metrics. *Eurasip J. Image Video Process.* **2008**, *2008*, 246309. [[CrossRef](#)]
61. Weng, X.; Wang, J.; Held, D.; Kitani, K. 3D multi-object tracking: A baseline and new evaluation metrics. In Proceedings of the IEEE International Conference on Intelligent Robots and Systems, Las Vegas, NV, USA, 25–29 October 2020; pp. 10359–10366. [[CrossRef](#)]
62. Kreiss, S.; Bertoni, L.; Alahi, A. PifPaf: Composite fields for human pose estimation. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 11969–11978. [[CrossRef](#)]
63. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-CNN. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 2961–2969. [[CrossRef](#)] [[PubMed](#)]