國立陽明交通大學
機器人碩士學位學程
碩士論文

Graduate Degree Program of Robotics
National Yang Ming Chiao Tung University
Master Thesis

利用與相機結合來提升毫米波雷達於多目標追蹤之可信度
Improving mmWave Radar Multi-Object Tracking Reliability Through Fusion with Camera

研究生：陳森（Kevin）

指導教授：胡竹生（HU, JWU-SHENG）

中華民國 一一三年 七月

July 2024

利用與相機結合來提升毫米波雷達於多目標追蹤之可信度
Improving mmWave Radar Multi-Object Tracking Reliability Through Fusion with Camera

研 究 生：陳森　　Student： Kevin

指導教授：胡竹生 博士　　Advisor： Dr. HU, JWU-SHENG

國立陽明交通大學

機器人碩士學位學程

碩士論文

A Thesis
Submitted to Graduate Degree Program of Robotics
College of Science
National Yang Ming Chiao Tung University
in partial Fulfillment of the Requirements
for the Degree of
Master of Science
in
Robotics

July 2024
Taiwan, Republic of China

中華民國 一一三 年 七 月

摘　要

本論文旨在開發了一種將攝影機和雷達感測器數據結合的系統，提高在多樣化與複雜的環境中跟蹤目標的穩定性及準確性。本研究使用貝葉斯融合和擴展卡爾曼濾波框架，將這些異類感測器的數據關聯，並將其無縫集成。研究展示了即使在複雜且動態變化的環境中，也能精確跟蹤和識別多個物體，同時實現實時性能。本研究利用來自攝影機的像素寬度和高度資訊，並透過演算法與雷達資訊融合生成真實世界的 3D 邊界框，且成功地實現了 0.1561 公尺的 RMSE 距離準確度和 0.0520 弧度的角度準確度結果。

關鍵字：**目標跟踪、毫米波雷達、貝葉斯融合、卡爾曼濾波**

Abstract

This master's thesis introduces a solution for the fusion of camera and radar sensors, aimed at enhancing the robustness and accuracy of object tracking in diverse and challenging scenarios. The proposed solution employs a method for correlating data from these heterogeneous sensors and integrates them seamlessly using the Bayesian Fusion and Extended Kalman Filter framework. The resulting algorithm demonstrates the capability to precisely track and identify multiple objects, even in complex and dynamically changing environments, all while achieving real-time performance. By utilizing width and height information in pixels from the camera, the algorithm generates real-world 3D bounding boxes by fusing with radar information. The proposed fusion method managed to achieve a range accuracy result of RMSE 0.1561 m and an angle accuracy of 0.0520 rad.

Keywords: **Object Tracking, mmWave Radar, Bayesian Fusion, Extended Kalman Filter**

Contents

List of Figures

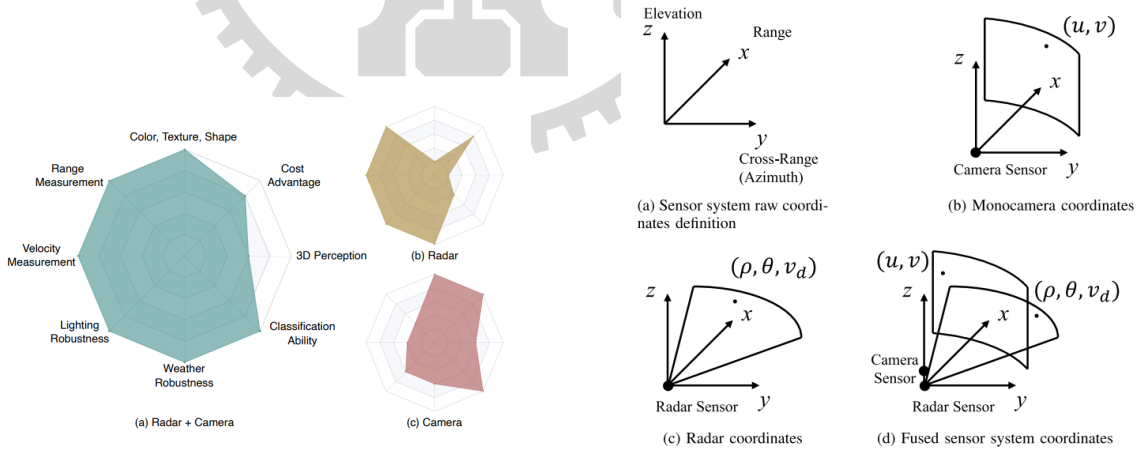
List of Tables

Chapter 1 Introduction

1.1 Research Background and Motivation

The inspiration for this thesis arose during the author's involvement with Unmanned Aerial Vehicles (UAVs). At the time, it was evident that existing obstacle-tracking and avoidance methods for UAVs faced significant challenges. The solutions in use were either incredibly unreliable or prohibitively expensive. For instance, cameras proved to be excellent for object tracking but struggled when it came to avoiding obstacles due to their limited range perception capabilities. Conversely, LiDAR excelled at obstacle detection, yet its reliability suffered in adverse weather conditions, and the associated costs could be exorbitant. Radar, while promising, also had its limitations, characterized by sparse and noisy data, making the identification and tracking of objects a formidable task, as illustrated in figure 1.1a.

The concept of fusing radar and camera technologies emerged as a promising approach. In theory, this fusion could provide accurate object tracking while maintaining the capability to avoid obstacles with great range accuracy. Nevertheless, this approach presents several challenges. Notably, radar and camera data exist in different planes, as illustrated in Figure 1.1b. This thesis shall explore a solution that can effectively process the raw data from both sensors and fuse them to create an ultimate sensor, applicable across a wide spectrum of use cases



(a) Charts of radar and camera characteristics[1]

(b) Radar and camera coordinate planes[2]

Figure 1.1：Radar camera calibration

1.2 Related Work

Which Sensor to Fuse

In recent years, the trend toward lidar-radar fusion has seen significant growth, as evidenced by numerous studies [3][4]. The fusion of radar and lidar presents a comparatively simpler scenario due to their ability to provide a Bird's Eye View (BEV), aligning their measurements within the same plane. Conversely, research efforts focusing on camera-lidar fusion [5][6] have emerged, yet both fusion approaches share a common drawback: the inherent cost inefficiency and susceptibility to adverse weather conditions, mainly rain, snow, and fog. Radar has the disadvantage of low resolution and noisy measurements, but is not affected by any weather or lighting conditions.

Radar-Camera Sensor Calibration

There are several methods of camera-radar calibration, most methods can be divided into three categories, mainly Pseudo Inverse (PI) [7], Direct Linear Transformation (DLT) [8], and Extrinsic Calibration (EC) [9]. Based on the experiment performed by Oh *et al.* [10], the method PI is not suitable for camera-radar calibration, while DLT may provide a better result than EC. The calibration method DLT is employed in this thesis due to its accuracy and simplicity.

Radar-Camera Fusion

Radar-camera fusion can be divided into four levels: object level, data level, feature level, and hybrid level [11]. **Data level fusion** [12] is where the raw data of radar and camera are fused before any pre-processing, also known as pixel-level fusion. The advantages of this method is that it provides redundancy, it also provides the least information loss. The disadvantage is without pre-processing, it is more prone noise and interference [13].

Feature level fusion [14], as the name implies, fuses features of both sensors, also known as middle-level fusion. This method transforms radar detection into an image form, its features are then extracted by an algorithm to be combined with the features of image from the camera. The advantage of this method is that it only focuses on the area of interest and disregards other information, making it one of the highest performance. The main disadvantage of this method is that it doesn't provide a solution for when the camera data is unreliable [11].

In **Object level fusion** [15], (also known as late-level fusion or decision-level fusion) the data is only fused after each sensor has completed its processing individually. The advantage of this fusion is that it has the flexibility of adding more sensors, and has better reliability performance [16]. The disadvantage of this method is that it consumes more time for pre-processing, it also can get confused with multiple targets in the same areas (i.e. when one object is obstructed), which will be mitigated in this thesis.

Real-world Application

When the author attended the Hannover Messe 2024 exhibition in Germany, he observed a real-world application very similar to his master's thesis. The sensor testing platform built by Hochschule Stralsund (fig 1.2) consists of a mmWave radar, an RGB camera, an infrared camera, a pair of LiDAR sensors, and a GPS. The main purpose of this platform is to assist a ship captain in spotting and tracking obstacles whilst determining whether the obstacle is in the ship's trajectory. What makes this platform unique is the complete combination of sensors, where each sensor compensates for the shortcomings of the others.

Figure 1.2：Maritime sensor platform by Hochschule Stralsund

1.3 Contribution

In this thesis, our primary objective is to harness the strengths of camera and radar sensors, while simultaneously mitigating their inherent limitations. Below are several notable contributions have been made:

1. **Extracting information from both camera and radar to produce real-world 3D Bounding Boxes:** The conventional approach for producing a Cartesian 3D Bounding Box typically involves the use of high-resolution 3D sensors, such as LiDAR. However, this thesis explores an alternative method by utilizing width and height information in pixels from the camera and leveraging sparse radar's range data as a key input.

2. **Keeping track of obstructed objects:** One of the limitations encountered by other radar-camera fusion systems is that the camera becomes less reliable when several objects obstruct each other. The thesis addresses this problem by utilizing the limited information from radar combined with predictions from the Kalman Filter.

3. **Improved Accuracy:** The experiments conducted in the later chapter demonstrate that the algorithm produces better results than using radar or camera alone in most cases. This thesis also compares its accuracy with other approaches and shows that it produces higher accuracy.

Chapter 2 Methodology

2.1 Configuration

In this experimental setup, two fundamental sensors are employed to acquire and analyze data. The Realsense D435i camera, known for its high-resolution image capture and depth-sensing capabilities, is utilized for visual data acquisition. Complementing the camera is the AWR1843boost mmWave radar, operating in the 77-81GHz frequency range. Both sensors are securely housed within a custom 3D-printed enclosure as seen in figure 2.1, which not only safeguards them but also minimizes external interference, ensuring the integrity of data acquisition. The mmWave radar config that is used is a 77-81GHz chirp, with settings balanced between range and resolution, collecting data at 20 frames per second.

Figure 2.1：Radar camera setup

2.2 Overview

Overview block diagram of the algorithm shown in figure 2.2.

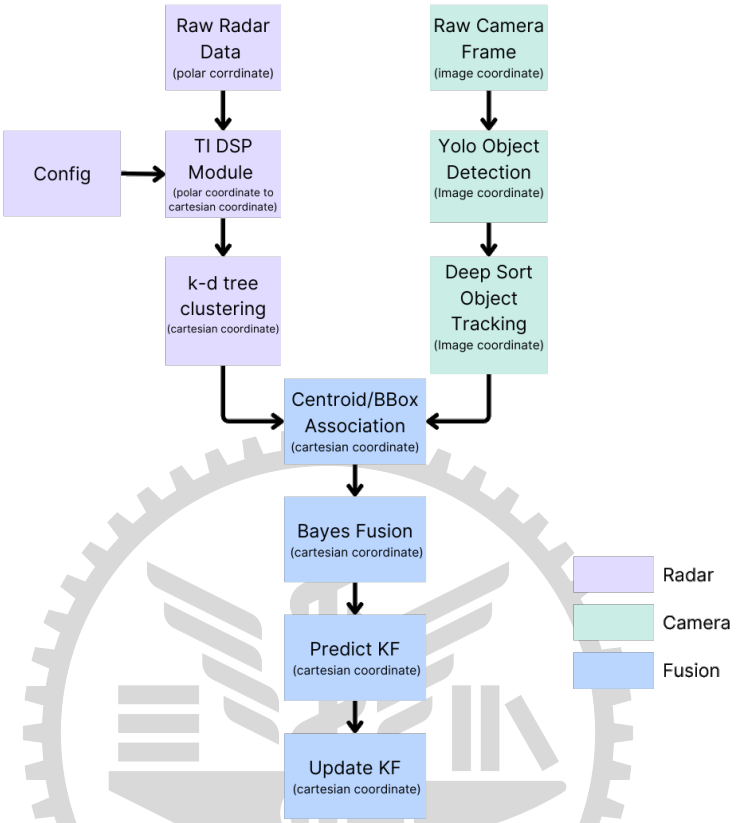


Figure 2.2：Radar camera Kalman Filter workflow

2.3 Calibration

2.3.1 Homography

Accurate fusion of both sensors requires the transformation of their measurements. In this scenario, given the challenges in mapping image coordinates to polar coordinates, homography method mentioned in this paper [8] is employed.

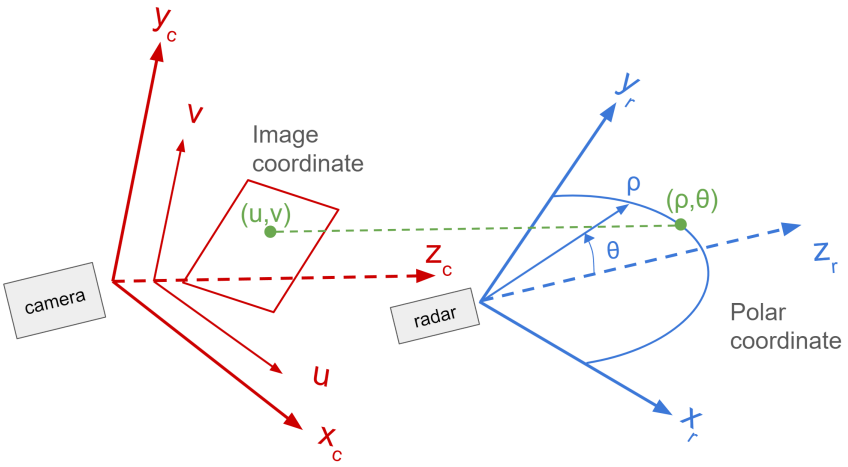


Figure 2.3：Homography of two coordinate system

Let radar coordinates be represented as (ρ, θ), and camera coordinates as (u, v), as described in fig 2.3. The rotation homography matrix $H = [h_{ij}]_{ij}$ for the camera with respect to the radar, originating from the origin and ω is an unknown constant; can be expressed as follows:

$$\omega \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{bmatrix} \begin{bmatrix} \rho \\ \theta \\ 1 \end{bmatrix} \tag{2.1}$$

The equation is similar to the linear least square problem, with enough data collected in the next section, it is solved with Singular Value Decomposition (SVD). Let $\omega = 1$ Solving equation 2.1 from table 2.2 yields :

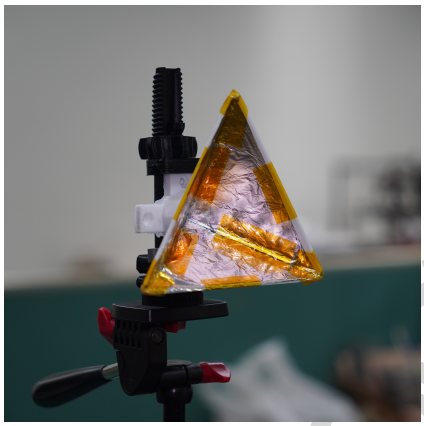
$$H = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{bmatrix} = \begin{bmatrix} 0 & 0.0003 & -0.2878 \\ 0.0002 & 0 & -0.151 \\ 0 & 0 & 1 \end{bmatrix} \tag{2.2}$$

Some assumptions have to be made to solve the equation:

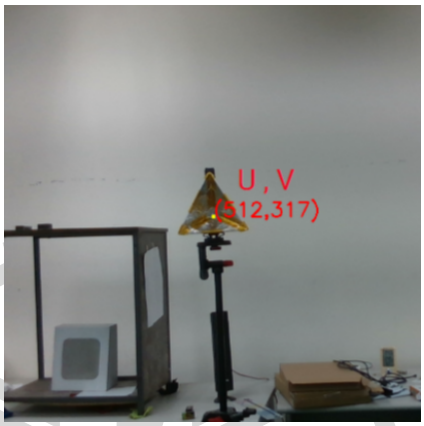
1. Radar's plane lies inside the camera plane (y_r=0), affine homography is assumed.
2. Sensor uncertainty (sensor noise) is not considered
3. This mapping method is only able to accurately project from image coordinates to azimuth in polar coordinates, which is expected

2.3.2 Calibration Approach

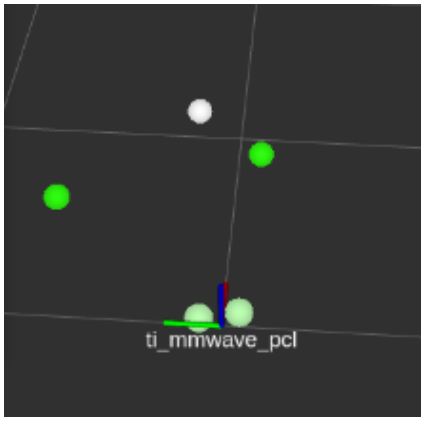
Data points from both radar and camera coordinates were collected with the corner reflector positioned at various locations. After data is collected, radar points are associated with image points based on equation 2.1. Image data was collected manually by reading the coordinates of the corner reflector from the camera's image. The mmWave radar was set to operate at 77-81 GHz chirp, with a configuration balanced between range and resolution, collecting data at 20 frames per second. As shown, the corner reflector represents the radar's reading of the strongest reflection (white dot).



(a) Radar Corner Reflector



(b) Corner reflector measurement from camera



(c) Corner reflector reading from radar

Figure 2.4：Radar camera calibration

2.3.3 Calibration Result

Calibration with the homography and approach produces a respectable result, it is able to accurately translate the image coordinate into polar coordinate. Albeit, there is a limitation to only being able to translate the angle, due to the camera's limitation perceiving depth.

Method	**RMSE**
Proposed	0.0465 rad, 0.0271 m
Domhof, Kooij, and Gavrila[17]	0.025 m

Table 2.1：calibration method RMSE compared

The table below shows the raw data collected from the camera and radar, which is then compared to the homography result of transforming image coordinates to polar coordinates for azimuth.

u (pixel)	**v** (pixel)	**radar azimuth** θ (rad)	**camera azimuth** homography result (rad)	**abs error** (rad)
639	297	0.03125508785	0.001141552511	0.03011353534
589	299	0.06254076689	0.05821917808	0.004321588807
564	303	0.09388787093	0.08675799087	0.007129880062
493	304	0.1886163813	0.1678082192	0.02080816213
433	306	0.2526802453	0.2363013699	0.01637887542
356	307	0.3178237065	0.3242009132	0.006377206754
240	317	0.4183463717	0.4566210046	0.03827463284
200	315	0.4528166082	0.502283105	0.04946649681
100	316	0.5235987508	0.6164383562	0.09283960532
23	321	0.5600753183	0.7043378995	0.1442625813
973	230	-0.3178237119	-0.3801369863	0.06231327441
945	241	-0.2850964515	-0.348173516	0.06307706449
885	270	-0.252680245	-0.2796803653	0.02700012035
848	293	-0.1886163813	-0.2374429224	0.04882654106
801	318	-0.1568928728	-0.1837899543	0.02689708157
390	272	0.2850964456	0.2853881279	0.0002916822178
432	281	0.252680245	0.2374429224	0.01523732258
467	289	0.2205332655	0.1974885845	0.02304468102
461	299	0.2205332494	0.2043378995	0.01619534981
485	313	0.1886163867	0.1769406393	0.01167574748
512	317	0.1568928748	0.1461187215	0.01077415335
515	328	0.1568928728	0.1426940639	0.01419880884
533	330	0.125327834	0.1221461187	0.00318171529
1103	265	-0.4528165732	-0.5285388128	0.07572223958
875	282	-0.2205332494	-0.2682648402	0.04773159083

Table 2.2：Camera data converted into polar coordinate compared to radar measurements (RMSE 0.0465 rad)

2.3.4 Generating 3D BBox

The camera's bounding box detection result is initially expressed in pixels. To transform these pixel coordinates into a meaningful real-world 3D bounding box, it is essential to acquire the distance of the object, a measurement obtained from the radar. Once the distance to the object is known, the width of the object can be readily derived using homography and trigonometry, and converted into cartesian coordinates.

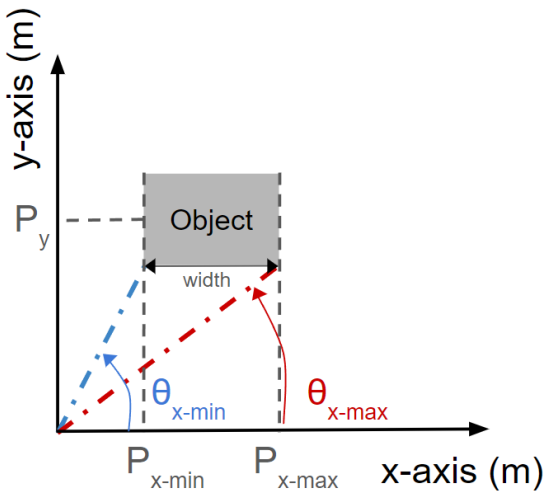


Figure 2.5：Object's width to polar coordinates relation

Cartesian distance to the object is obtained from range (ρ) data from radar's polar coordinate:

$$p_y = \rho sin(\theta) \tag{2.3}$$

Where x_{min},x_{max},y_{min}, and y_{max} the size of the bbox in pixel obtained from Yolo detection, is then converted into θ_{x-max} and θ_{x-min} using homography equation.

$$\begin{aligned} \text{width} &= P_{x-max} - P_{x-min} \\ &= p_y \tan(\theta_{x-max}) - p_y \tan(\theta_{x-min}) \\ &= p_y(\tan(\theta_{x-max}) - \tan(\theta_{x-min})) \end{aligned} \tag{2.4}$$

The same equation can also be applied to derive the height of the detected image.

$$\begin{aligned} \text{height} &= P_{y-max} - P_{y-min} \\ &= p_y \tan(\theta_{y-max}) - p_y \tan(\theta_{y-min}) \\ &= p_y(\tan(\theta_{y-max}) - \tan(\theta_{y-min})) \end{aligned} \tag{2.5}$$

2.4 Data Pre-Processing

2.4.1 mmWave Radar Data Pre-Processing

Given that radar data is inherently sparse and noisy, its data needed to be filtered. For this purpose, k-d tree is employed to cluster the pointcloud. A k-d tree, short for k-dimensional tree, is a hierarchical data structure used for efficient multidimensional data organization and search operations. It arranges data points in k-dimensional space, such as spatial coordinates, in a binary tree structure.

2.4.2 Image Data Pre-Processing

Image Recognition and Tracking

YOLOv3 is utilized to generate bounding boxes (BBox) and regions of interest (ROI)[18]. Subsequently, DeepSORT is applied for tracking the generated bounding boxes for objects, enabling robust and efficient object tracking throughout the analysis[19].

2.5 Radar-camera Data Association

Prior to fusion, radar clusters have to be associated with tracked objects from deepsort. First, centroids of radar clusters are mapped into image coordinates. Second, is to find the theoretical error of radar's measurement [2], which is radar's resolution, determined by equation 2.6 Finally, if the distance between the calculated radar centroid and the image detection BBox lies within the theoretical boundary , we can safely assume both measurements belong to the same object of interest.

$$\Delta\theta = \frac{c_0}{f_c d N_{RX} N_{TX} \cos(\theta_i)} \tag{2.6}$$

where

$$\begin{aligned} f_c &= \text{center frequency} \\ \lambda &= \text{carrier signal wavelength} \\ d &= \lambda/2 \\ N_{RX} &= \text{Number of receiving antenna} \\ N_{TX} &= \text{Number of transferring antenna} \\ \theta_i &= \text{angle of interest} \end{aligned}$$

2.6 Radar-camera Data Synchronization

Given the disparate update rates of radar and camera data, coupled with the inherent time required for data processing, effective synchronization of measurements becomes crucial. Notably, image processing emerges as the most time-intensive step in this synchronization process. The algorithm strategically utilizes radar data corresponding to the processed image, storing it for fusion in later stages, as depicted in Figure 2.6.

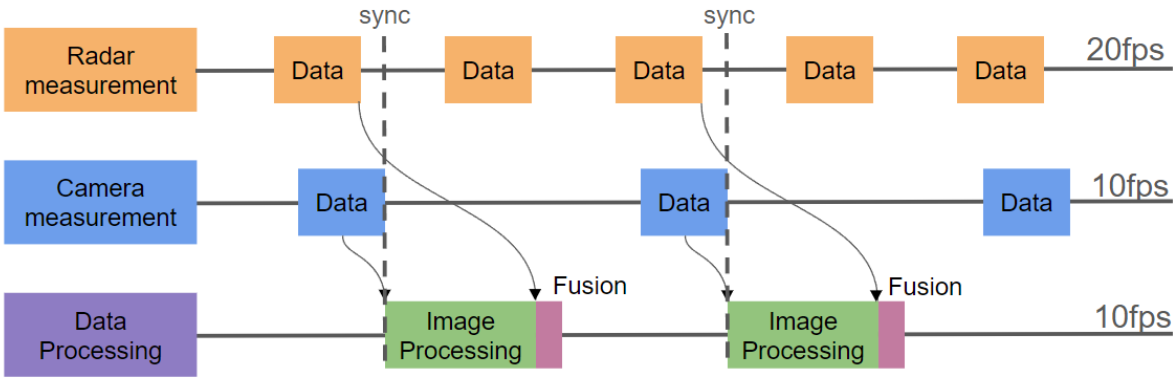


Figure 2.6：Camera and Radar Data Synchronizing

Chapter 3 Problem Statement

3.1 Fusion Algorithm

Fusing measurements from two heterogeneous sensors that have few cross-correlations can be challenging. To take inputs from two different sensors and give them weight, we use Bayes Fusion in this thesis. Bayes fusion takes the noises of both sensors to determine their reliability at a given point of the measurement. Thus giving us a reliable way to give scores of trust to each sensor.

The coordinates that undergo fusion from both radar and image sensors are the horizontal coordinates, specifically the radar's azimuth and the image's "u" coordinate. This constraint arises from the fact that these are the only coordinates where both sensors provide measurements, as illustrated in Figure 1.1b. It's important to note that radar's elevation measurement resolution is limited and subject to noise. Additionally, the image sensor does not provide depth information.

If X is the real position of the object, then Bayes' theorem predicts that the probability of the fused position is shown in equation 3.1 [20].

$$P_{prob}(\frac{P}{X}) = \frac{e^{\frac{-(P-X)^T R^{-1}(P-X)}{2}}}{2\pi R^(0.5)} \tag{3.1}$$

Applying Bayes' fusion, the value of the measured measurements is provided by equation 3.2.

$$\theta_{bayes} = \frac{\frac{\theta_{radar}}{R_{radar}} + \frac{\theta_{cam}}{R_{cam}}}{\frac{1}{R_{radar}} + \frac{1}{R_{cam}}} \tag{3.2}$$

where

$$\begin{aligned} \theta_{bayes} &= \text{fused position} \\ \theta_{radar} &= \text{radar azimuth angle} \\ \theta_{cam} &= \text{camera azimuth angle} \\ R_{radar} &= \text{radar covariance} \\ R_{cam} &= \text{camera covariance} \end{aligned}$$

$$\frac{1}{R} = \frac{1}{R_1} + \frac{1}{R_2} \tag{3.3}$$

$$\mathbf{R}_{radar} = \sigma^2_{radar_x} \tag{3.4}$$

$$\mathbf{R}_{cam} = \sigma^2_{cam_u} \tag{3.5}$$

3.2 Motion Model

3.2.1 Predict

The state matrix used in this Kalman Filter [21] is from the single sensor maneuvering tracking, with constant velocity. It has four elements and is defined with position and velocity, which projects onto the x-axis and y-axis. State vector also includes the width and the height of the object[22].

$$\mathbf{x} = \begin{bmatrix} p \\ v \\ w \\ h \end{bmatrix} = \begin{bmatrix} p_x \\ p_y \\ v_x \\ v_y \\ w \\ h \end{bmatrix} \tag{3.6}$$

where

$p_x =$ position in x-axis
$p_y =$ position in y-axis
$v_x =$ velocity in x-axis
$v_y =$ velocity in y-axis
$w =$ width(meter)
$h =$ height(meter)

Transition Matrix is expressed as follows:

$$\mathbf{F} = \begin{bmatrix} 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \tag{3.7}$$

Thus state vector is updated as below:

$$\mathbf{x}_k = \mathbf{F}_k \mathbf{x}_{k-1} + \mathbf{w} \tag{3.8}$$

Error covariance update:

$$\mathbf{P}_k = \mathbf{F}_k \mathbf{P}_{k-1} \mathbf{F}_k^T + \mathbf{Q}_k \tag{3.9}$$

3.2.2 Update

Both radar and camera undergo the same update procedure within the Kalman filter framework. Nevertheless, variations exist in their measurement data, leading to distinct measurement matrices and noise covariance. Radar measurements encompass two parameters: azimuth and range, both presented in polar coordinates. These radar coordinates are derived from the centroids of the k-d tree clusters. Meanwhile, the width and height measurements are taken from equation 2.4 and equation 2.5.

$$\mathbf{z} = \begin{bmatrix} \rho \\ \theta_{bayes} \\ w \\ h \end{bmatrix} \tag{3.10}$$

After the measurement matrix $\mathbf{z}_k$ is obtained, it is subtracted from the previously predicted values by the Kalman Filter.

$$\mathbf{y}_k = \mathbf{z} - \mathbf{H}_k \mathbf{x}_k \tag{3.11}$$

Subsequently, the matrix $\mathbf{G}_k$ is calculated to determine the trustworthiness of the current measurement, taking into account measurement noise.

$$\mathbf{G}_k = \frac{\mathbf{P}_k \mathbf{H}_k^T}{\mathbf{H}_k \mathbf{P}_k \mathbf{H}_k^T + \mathbf{R}_k} \tag{3.12}$$

Finally, the current state estimation is updated based on this adjusted measurement.

$$\mathbf{P}_k = (\mathbf{I} - \mathbf{G}_k \mathbf{H}) \mathbf{P}_k \tag{3.13}$$

3.2.3 Non-linearity

Since measurements and Bayes fusion results are obtained and calculated in polar coordinates, the application of the Extended Kalman Filter becomes crucial to linearize these results into Cartesian coordinates. $h(\mathbf{x}_k)$ matrix represents conversion of polar coordinate to the cartesian coordinate, which is defined as follows:

$$h(\mathbf{x}_k) = \begin{bmatrix} \rho \\ \theta_{bayes} \\ w \\ h \end{bmatrix} = \begin{bmatrix} \sqrt{p_x^2 + p_y^2} \\ \tan^{-1}(\frac{p_x}{p_y}) \\ p_y(\tan(\theta_{x-max}) - \tan(\theta_{x-min})) \\ p_y(\tan(\theta_{y-max}) - \tan(\theta_{y-min})) \end{bmatrix} \tag{3.14}$$

Because the solution comprises higher-dimensional components than the equations, a first-order Jacobian matrix is employed on **H**:

$$\mathbf{H} = \left[\frac{\partial h(\mathbf{x}_k)}{\partial \mathbf{x}_k}\right] = \begin{bmatrix} \frac{\partial \rho}{\partial p_x} & \frac{\partial \rho}{\partial p_y} & \frac{\partial \rho}{\partial v_x} & \frac{\partial \rho}{\partial v_y} & \frac{\partial \rho}{\partial w} & \frac{\partial \rho}{\partial h} \\ \frac{\partial \theta}{\partial p_x} & \frac{\partial \theta}{\partial p_y} & \frac{\partial \theta}{\partial v_x} & \frac{\partial \theta}{\partial v_y} & \frac{\partial \theta}{\partial w} & \frac{\partial \theta}{\partial h} \\ \frac{\partial w}{\partial p_x} & \frac{\partial w}{\partial p_y} & \frac{\partial w}{\partial v_x} & \frac{\partial w}{\partial v_y} & \frac{\partial w}{\partial w} & \frac{\partial w}{\partial h} \\ \frac{\partial h}{\partial p_x} & \frac{\partial h}{\partial p_y} & \frac{\partial h}{\partial v_x} & \frac{\partial h}{\partial v_y} & \frac{\partial h}{\partial w} & \frac{\partial h}{\partial h} \end{bmatrix} \tag{3.15}$$

Thus giving the transition matrix as follows:

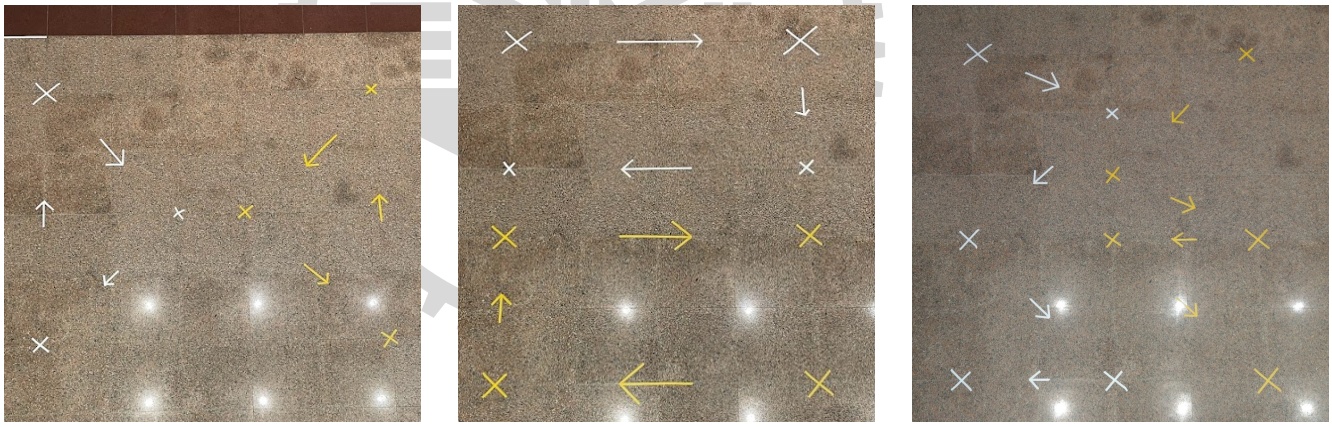
$$\mathbf{H} = \begin{bmatrix} \frac{p_x}{\sqrt{p_x^2+p_y^2}} & \frac{p_y}{\sqrt{p_x^2+p_y^2}} & 0 & 0 & 0 & 0 \\ -\frac{p_y}{p_x^2+p_y^2} & \frac{p_x}{p_x^2+p_y^2} & 0 & 0 & 0 & 0 \\ 0 & (\tan(\theta_{x-max}) - \tan(\theta_{x-min})) & 0 & 0 & 1 & 0 \\ 0 & (\tan(\theta_{y-max}) - \tan(\theta_{y-min})) & 0 & 0 & 0 & 1 \end{bmatrix} \tag{3.16}$$

Chapter 4 Evaluation

4.1 Experiment

4.1.1 Experiment Setup

Our experiments aim to illustrate how the strengths of one sensor compensate for weaknesses in the other, demonstrating the potential of sensor fusion. The weakness inherent in radar lies in its challenge to recognize objects due to sparse measurements. Although a camera offers high-resolution measurements, it lacks depth perception. However, radar compensates for this limitation by providing highly accurate range measurements, effectively mitigating the shortcomings of the camera. To test the performance of the algorithm, 3 specific scenarios were tested: 1. Control scenario, when object 1 (white) and object 2 (yellow) do not crosspath. 2. One object conceals another and continues their original trajectory after departing. 3. One object conceals another and changes trajectory when departing.



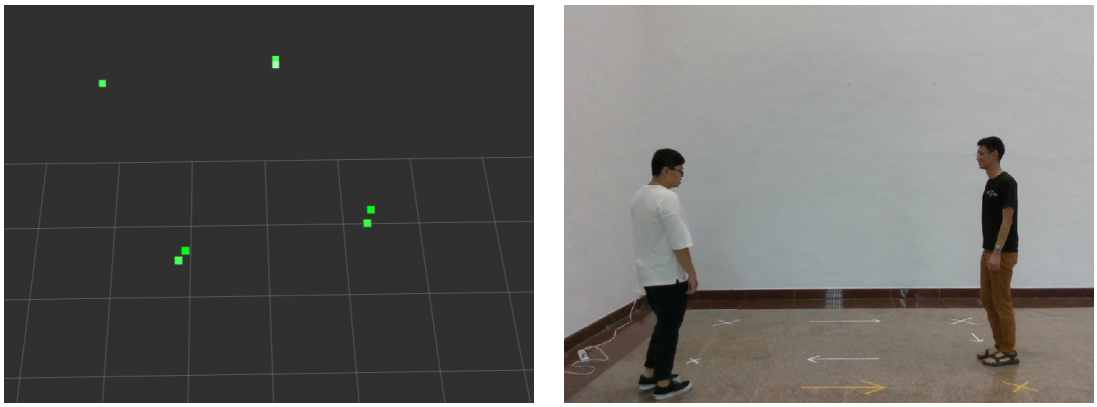
(a) scenario 1 (b) scenario 2 (c) scenario 3

Figure 4.1：Bird Eye View of multi-object tracking

4.1.2 Challenges to Overcome

Blind spot of both sensors. Challenges the algorithm to correctly predict objects when data is obstructed.

From figure 4.3a can be seen that both object clusters disappear.



(a) Raw radar data (b) Image frame

Figure 4.2：Object 1 and object 2 before crossingpath

(a) Raw radar data (b) Image frame

Figure 4.3：Object 1 covers object 2

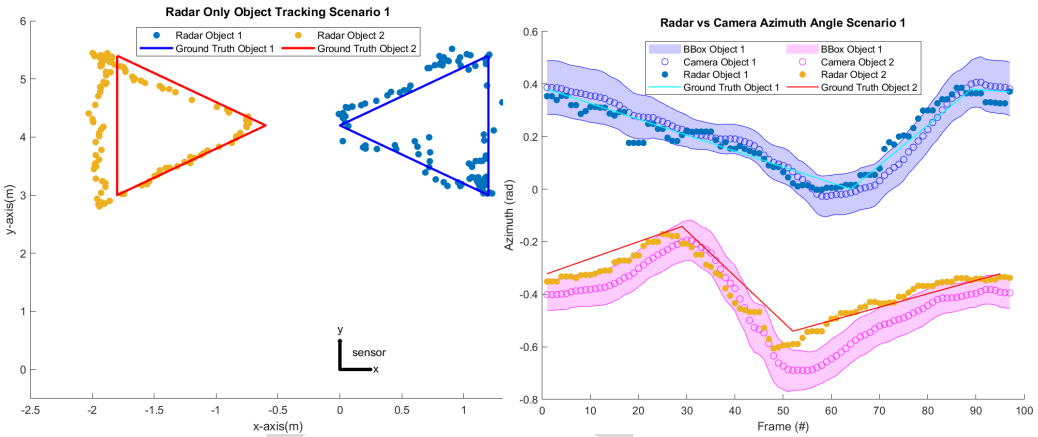
(a) Raw radar data (b) Image frame

Figure 4.4：Object 1 and object 2 seperates

4.2 Experiment Result

4.2.1 Scenario 1

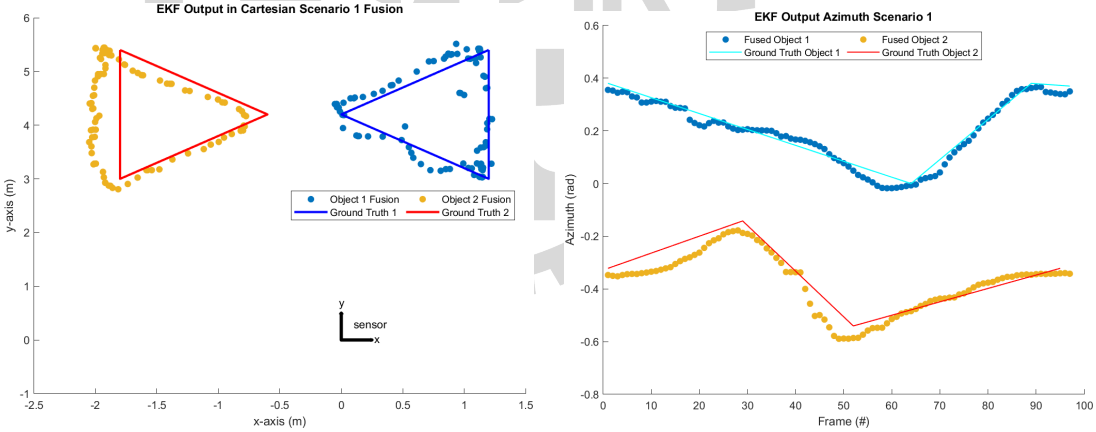
Scenario 1 is the control when object 1 and object 2 do not crosspath.



(a) Radar Only Object Tracking Scenario 1

(b) Raw radar and image data after data association

Figure 4.5：Raw Data of Scenario 1



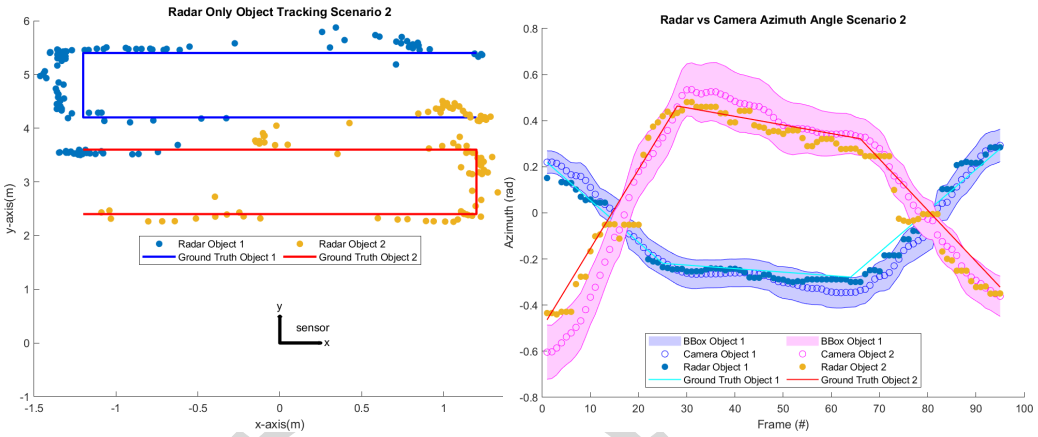
(a) EKF Output in Cartesian Scenario 1

(b) EKF Output Azimuth Scenario 1

Figure 4.6：EKF Output of Scenario 1

4.2.2 Scenario 2

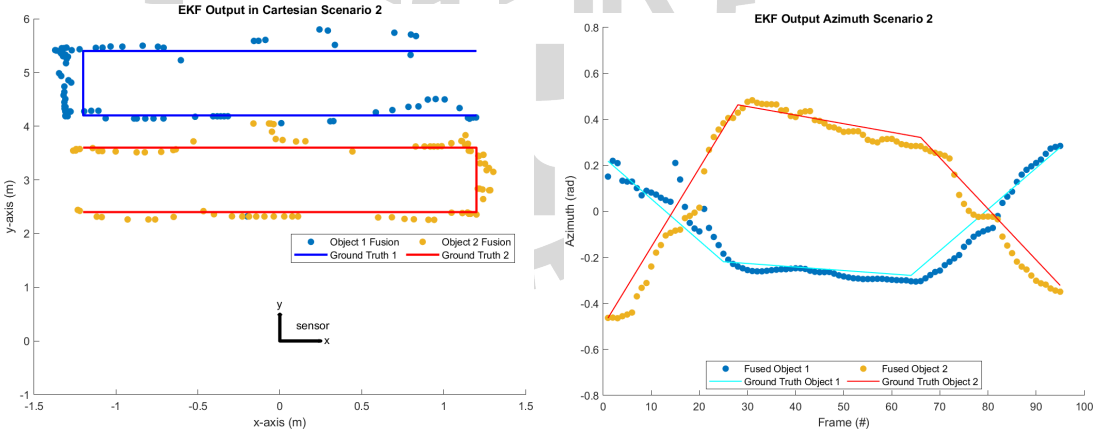
In scenario 2, object 1 crosses path with object 2 in a straight line horizontally. After concealing object 2 from both camera and radar, both objects continue their original trajectory. In figure 4.7a can be seen that object 1 and object 2 switch places. The algorithm is able to track and identify objects 1 and 2 correctly (figure 4.8a).



(a) Radar Only Object Tracking Scenario 2

(b) Raw radar and image data after data association

Figure 4.7：Raw Data of Scenario 2



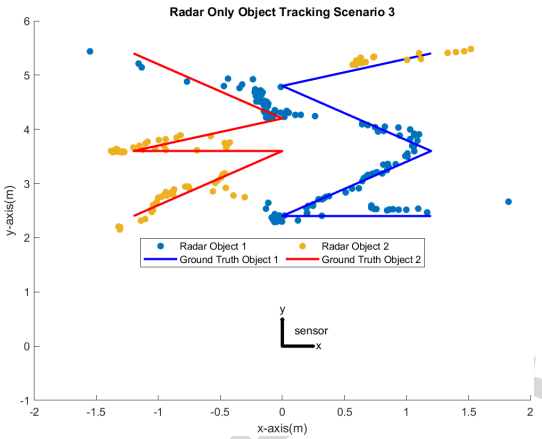
(a) EKF Output in Cartesian Scenario 2

(b) EKF Output Azimuth Scenario 2

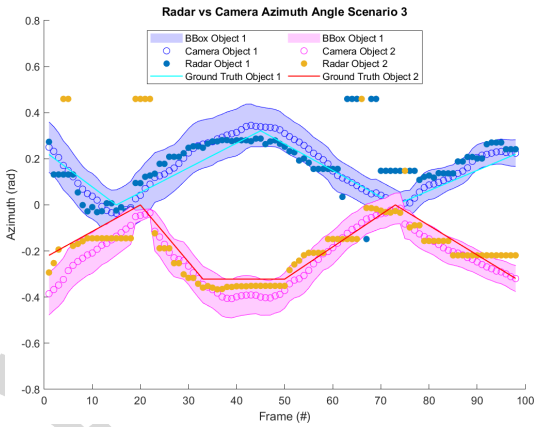
Figure 4.8：EKF Output of Scenario 2

4.2.3 Scenario 3

In scenario 3, object 1 covers object 2, and in the next few frames, object 2 covers object 1. In this scenario, both objects change trajectory when departing from each other. In figure 4.9a can be seen that object 1 and object 2 switch places. The algorithm is able to track and identify objects 1 and 2 correctly (figure 4.10a).

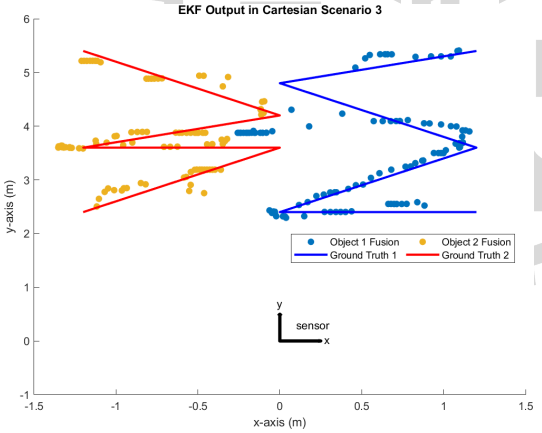


(a) Radar Only Object Tracking Scenario 3

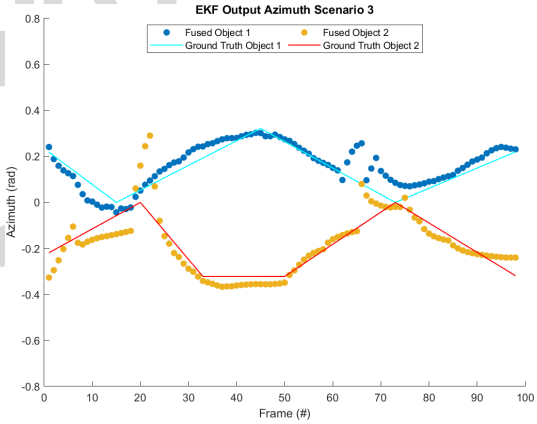


(b) Raw radar and image data after data association

Figure 4.9：Raw Data of Scenario 3



(a) EKF Output in Cartesian Scenario 3



(b) EKF Output Azimuth Scenario 3

Figure 4.10：EKF Output of Scenario 3

4.3 Object Tracking Performance Evaluation

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{4.1}$$

4.3.1 3D Bounding Box Error Evaluation

The accuracy of the 3D Bounding Box depends on the precision of both the radar's range and, notably, Yolo's Bounding Box (BBox). Additionally, the accuracy of the width and height of the bounding box is contingent upon capturing the subject of interest entirely within the camera frame. In our scenarios, the accuracy of the subjects' height is reliable only when they are 3.7 meters or farther away. Width prediction performance is not analyzed since our setup does not have a reliable way to collect subjects' widths. The tables below demonstrate that the algorithm can predict the subject's height with accuracy within 0.1 meters throughout the experiment.

Scenario	**Height (m)**			
	Object 1 Ground Truth = 1.67m		**Object 2** Ground Truth = 1.80m	
	Height	Covariance	Height	Covariance
Scenario 1	1.6954	0.0072	1.8190	0.0333
Scenario 2	1.7677	0.0061	1.9494	0.2058
Scenario 3	1.7583	0.1554	1.8470	0.0932

Table 4.1：All scenarios BBox prediction compared

4.3.2 Azimuth Error Evaluation

The table presents azimuth errors (Root Mean Square Error - RMSE) for three different scenarios, considering two objects in each scenario. The RMSE (rad) columns provide a comparison of errors in radians for the fusion of radar and camera data, as well as individual errors for camera-only and radar-only measurements. Despite the difference in theta accuracy between radar and camera, the algorithm effectively leverages the higher accuracy of the camera's angle while incorporating data from both sources. The "Fusion" column represents the azimuth error when the algorithm fuses data from both radar and camera. In some cases (e.g., Scenario 1), the fusion result outperforms both individual camera and radar measurements, showcasing the benefits of integrating information from both sources. In some cases, the fusion result can lag behind the camera's result, but not by a large margin.

Scenario	RMSE (rad)					
	Object 1			Object 2		
	Fusion	Camera	Radar	Fusion	Camera	Radar
Scenario 1	**0.0242**	0.0312	0.0358	**0.0443**	0.0977	0.0493
Scenario 2	0.0566	**0.0555**	0.1228	**0.0644**	0.0917	0.0749
Scenario 3	0.0512	**0.0366**	0.1158	0.0672	**0.0611**	0.1537

Table 4.2：All scenarios azimuth error compared

4.3.3 Cartesian Error Evaluation

From Scenario 1, it becomes apparent that all components—fusion and radar commendable performance. In scenario 2, where the algorithm showcases its strengths, it combines the best of both sensors to produce better a result. Conversely, in Scenario 3, radar struggles significantly in tracking object 2 as it enters its blind spot. Nevertheless, the algorithm maintains its performance by incorporating camera data. The collective analysis across all scenarios yields an average RMSE of 0.1561m.

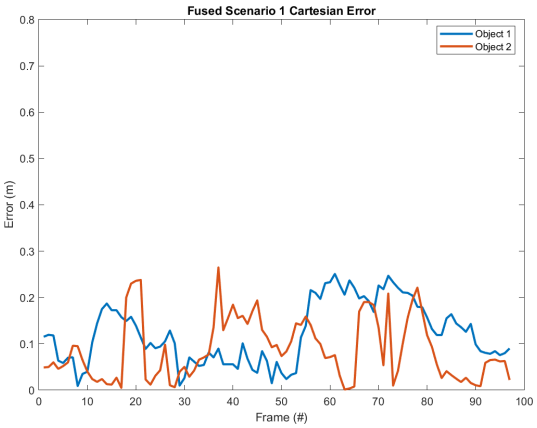
Scenario	RMSE (m)			
	Object 1		Object 2	
	Fusion	Radar	Fusion	Radar
Scenario 1	0.1388	**0.1124**	**0.105**	0.1201
Scenario 2	**0.1206**	0.1228	**0.1687**	0.2014
Scenario 3	**0.0903**	0.0928	**0.3075**	0.3592

Table 4.3：Cartesian RMSE compared

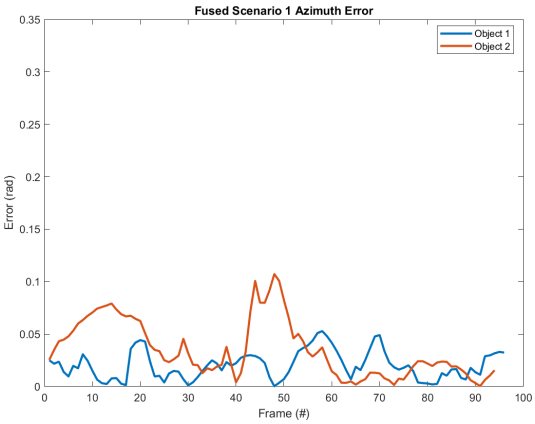
Comparison of errors between the author's proposed method and other radar-camera tracking algorithms. Please note that the author did not benchmark the results of other algorithms with the same dataset.

Method	Radar	Object	Max Range	RMSE
Proposed	AWR1843	human	6m	**0.1561 m**
Kang and Kum[23]	RT3002	vehicle	40m	0.18 m
Kim, Kim, and Kum[24]	AWR1642	vehicle	50m	0.1828 m
Zewge, Kim, Kim, *et al.*[25]	IWR1443	human	6m	0.2486 m
Zhang and Cao[2]	AWR1642	human	10m	0.2902 m

Table 4.4：Comparison of other tracking method

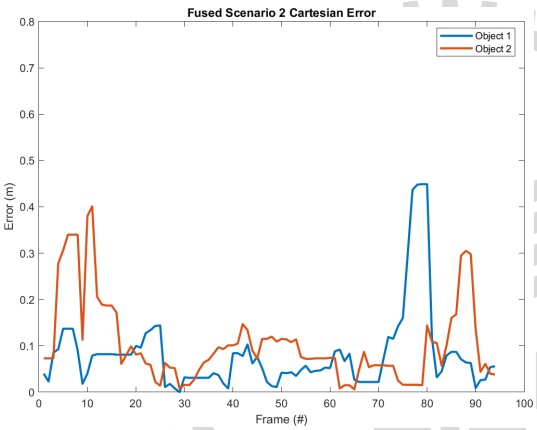


(a) Fused Scenario 1 Cartesian Error

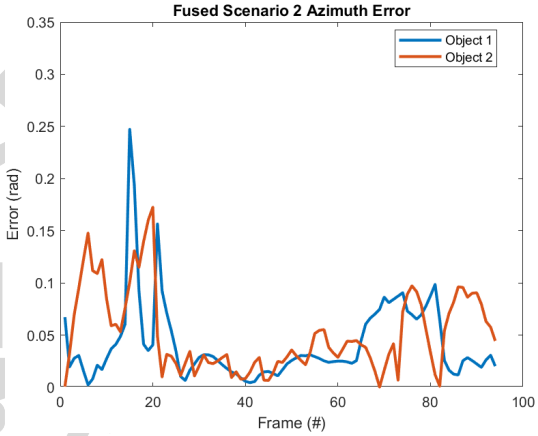


(b) Fused Scenario 1 Azimuth Error

Figure 4.11：EKF Error of Scenario 1

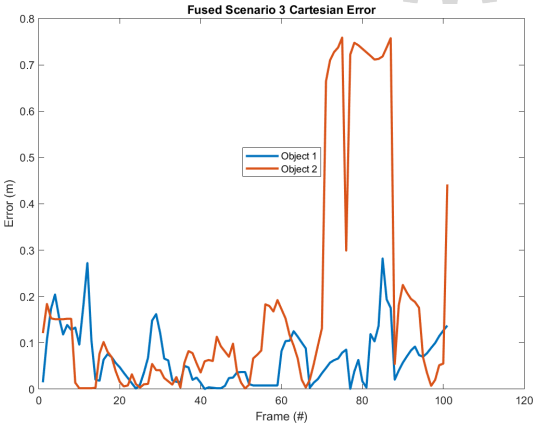


(a) Fused Scenario 2 Cartesian Error

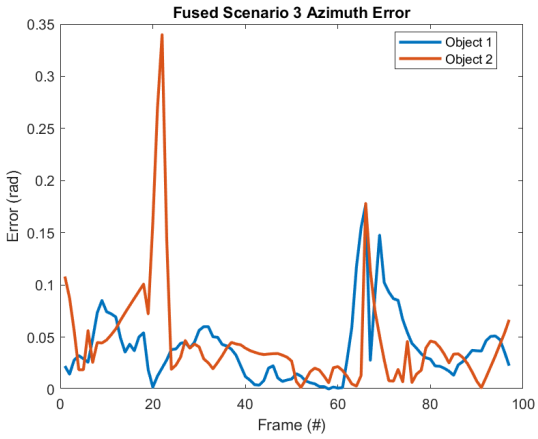


(b) Fused Scenario 2 Azimuth Error

Figure 4.12：EKF Error of Scenario 2



(a) Fused Scenario 3 Cartesian Error



(b) Fused Scenario 3 Azimuth Error

Figure 4.13：EKF Error of Scenario 3

Chapter 5 Conclusion

5.1 Limitation and Future Work

In the future, addressing certain aspects could enhance the thesis's outcomes. For instance, exploring methods that account for width data from radar clusters to improve width accuracy. Introducing algorithms considering radar point reflection intensity might also improve accuracy. Further validation with better sensors, i.e. LiDAR, considering the limitations of ground truth, particularly for human-tracked objects, would increase the thesis's contribution.

References

[1] S. Yao, R. Guan, X. Huang, *et al.*, "Radar-camera fusion for object detection and semantic segmentation in autonomous driving: A comprehensive review," *IEEE Transactions on Intelligent Vehicles*, pp. 1–40, 2023. DOI: 10.1109/tiv.2023.3307157. [Online]. Available: https://doi.org/10.1109%2Ftiv.2023.3307157.

[2] R. Zhang and S. Cao, "Extending reliability of mmwave radar tracking and detection via fusion with camera," *IEEE Access*, vol. 7, pp. 137 065–137 079, 2019. DOI: 10.1109/ACCESS.2019.2942382.

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

[4] 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 *2016 IEEE 22nd International Conference on Embedded and Real-Time Computing Systems and Applications (RTCSA)*, 2016, pp. 104–104. DOI: 10.1109/RTCSA.2016.20.

[5] X. Chen, H. Ma, J. Wan, B. Li, and T. Xia, *Multi-view 3d object detection network for autonomous driving*, 2017. arXiv: 1611.07759 [cs.CV].

[6] B. Li, T. Zhang, and T. Xia, *Vehicle detection from 3d lidar using fully convolutional network*, 2016. arXiv: 1608.07916 [cs.CV].

[7] T. Wang, N. Zheng, J. Xin, and Z. Ma, "Integrating millimeter wave radar with a monocular vision sensor for on-road obstacle detection applications," *Sensors*, vol. 11, no. 9, pp. 8992–9008, 2011, ISSN: 1424-8220. DOI: 10.3390/s110908992. [Online]. Available: https://www.mdpi.com/1424-8220/11/9/8992.

[8] D. Y. Kim and M. Jeon, "Data fusion of radar and image measurements for multi-object tracking via kalman filtering," *Information Sciences*, vol. 278, pp. 641–652, 2014, ISSN: 0020-0255. DOI: https://doi.org/10.1016/j.ins.2014.03.080. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0020025514003715.

[9] Z. Ji and D. Prokhorov, "Radar-vision fusion for object classification," in *2008 11th International Conference on Information Fusion*, 2008, pp. 1–7.

[10] J. Oh, K.-S. Kim, M. Park, and S. Kim, "A comparative study on camera-radar calibration methods," in *2018 15th International Conference on Control, Automation, Robotics and Vision (ICARCV)*, 2018, pp. 1057–1062. DOI: 10.1109/ICARCV.2018.8581329.

[11] S. Yao, R. Guan, X. Huang, *et al.*, "Radar-camera fusion for object detection and semantic segmentation in autonomous driving: A comprehensive review," *IEEE Transactions on Intelligent Vehicles*, vol. 9, no. 1, pp. 2094–2128, 2024. DOI: 10.1109/TIV.2023.3307157.

[12] K. Bansal, K. Rungta, and D. Bharadia, *Radsegnet: A reliable approach to radar camera fusion*, 2022. arXiv: 2208.03849 [cs.CV].

[13] Z. Wei, F. Zhang, S. Chang, Y. Liu, H. Wu, and Z. Feng, *Mmwave radar and vision fusion for object detection in autonomous driving: A review*, 2022. arXiv: 2108.03004 [cs.CV].

[14] S. Chadwick, W. Maddern, and P. Newman, *Distant vehicle detection using radar and vision*, 2019. arXiv: 1901.10951 [cs.RO].

[15] H. Jha, V. Lodhi, and D. Chakravarty, "Object detection and identification using vision and radar data fusion system for ground-based navigation," in *2019 6th International Conference on Signal Processing and Integrated Networks (SPIN)*, 2019, pp. 590–593. DOI: 10.1109/SPIN.2019.8711717.

[16] Y. Zhou, Y. Dong, F. Hou, and J. Wu, "Review on millimeter-wave radar and camera fusion technology," *Sustainability*, vol. 14, no. 9, 2022, ISSN: 2071-1050. DOI: 10.3390/su14095114. [Online]. Available: https://www.mdpi.com/2071-1050/14/9/5114.

[17] J. Domhof, J. F. Kooij, and D. M. Gavrila, "An extrinsic calibration tool for radar, camera and lidar," in *2019 International Conference on Robotics and Automation (ICRA)*, 2019, pp. 8107–8113. DOI: 10.1109/ICRA.2019.8794186.

[18] J. Redmon and A. Farhadi, *Yolov3: An incremental improvement*, 2018. arXiv: 1804.02767 [cs.CV].

[19] N. Wojke, A. Bewley, and D. Paulus, "Simple online and realtime tracking with a deep association metric," in *2017 IEEE International Conference on Image Processing (ICIP)*, IEEE, 2017, pp. 3645–3649. DOI: 10.1109/ICIP.2017.8296962.

[20] R. Kunjumon and G. S. Sangeetha Gopan, "Sensor fusion of camera and lidar using kalman filter," in *Intelligent Systems*, A. Sheth, A. Sinhal, A. Shrivastava, and A. K. Pandey, Eds., Singapore: Springer Singapore, 2021, pp. 327–343, ISBN: 978-981-16-2248-9.

[21] R. E. Kalman, "A New Approach to Linear Filtering and Prediction Problems," *Journal of Basic Engineering*, vol. 82, no. 1, pp. 35–45, Mar. 1960, ISSN: 0021-9223. DOI: 10.1115/1.3662552. eprint: https://asmedigitalcollection.asme.org/fluidsengineering/article-pdf/82/1/35/5518977/35_1.pdf. [Online]. Available: https://doi.org/10.1115/1.3662552.

[22] P. Kmiotek and Y. Ruichek, "Representing and tracking of dynamics objects using oriented bounding box and extended kalman filter," in *2008 11th International IEEE Conference on Intelligent Transportation Systems*, 2008, pp. 322–328. DOI: 10.1109/ITSC.2008.4732695.

[23] D. Kang and D. Kum, "Camera and radar sensor fusion for robust vehicle localization via vehicle part localization," *IEEE Access*, vol. 8, pp. 75 223–75 236, 2020. DOI: 10.1109/ACCESS.2020.2985075.

[24] J. Kim, Y. Kim, and D. Kum, "Low-level sensor fusion for 3d vehicle detection using radar range-azimuth heatmap and monocular image," Nov. 2020.

[25] N. S. Zewge, Y. Kim, J. Kim, and J.-H. Kim, “Millimeter-wave radar and rgb-d camera sensor fusion for real-time people detection and tracking,” in *2019 7th International Conference on Robot Intelligence Technology and Applications (RiTA)*, 2019, pp. 93–98. DOI: 10.1109/RITAPP.2019.8932892.