2023 IEEE 10th International Workshop on Metrology for AeroSpace (MetroAeroSpace), 19-21 June 2023, Milan, Italy
DOI:10.1109/MetroAeroSpace57412.2023.10189934

Radar-Camera Fusion for Ground-based Perception of Small UAV in Urban Air Mobility

Cheng Huang, Ivan Petrunin and Antonios Tsourdos
School of Aerospace, Transport and Manufacturing, Cranfield University
Bedfordshire, MK43 0AL, UK
{cheng-huang.huang, i.petrunin, a.tsourdos}@cranfield.ac.uk

***Abstract*—The resilient surveillance of cooperative and non-cooperative aerial targets is critical for the safety and security of urban air mobility operations. Accurate detection, tracking, and trajectory prediction are essential to the subsequent tasks, e.g. tactical conflict prediction and resolution. Meanwhile, the combination of radar and camera is a classic option to provide perception services in different challenging environments. In this paper, a deep semantic association network is proposed for building relationships between the image detections and raw radar points, which then contributes to subsequent tasks, e.g. detecting, tracking, and predicting the small UAV with networked radar and camera systems. Various flight trials are conducted for collecting multi-sensor data, finally, training and testing results on this dataset demonstrate the outstanding performance of the proposed fusion workflow in comparison to single-sensor performance. At the same time, the 2D predictions in the sensor network are reconstructed to 3D trajectories for comparison and also reveal the improvements of the radar-camera fusion approach.**

***Index Terms*—sensor fusion, deep association, neural network, urban air mobility**

I. Introduction

The efficient and safe operation is critical for urban air mobility (UAM), which enables high-demand services such as cargo delivery and passenger transit in the low-altitude airspace. Cooperative unmanned aerial vehicles (UAV) rely on onboard sensors for communication and sharing their positions and intentions with the air traffic control system [1]. Whereas non-cooperative even malicious intruders would be dangerous for other flights if risks are not recognized and identified in advance. For the purpose of operation safety and security, all aerial objects should be detected and tracked in real-time for assessing risks posed to other airspace users, as well as people and infrastructures on the ground. In metropolitan regions, the performance of some cooperative sensors and GPS degrades because of the blocking of tall buildings and street canyons [2]. To assist the surveillance of cooperative non-cooperative targets, non-cooperative sensors, e.g. radar, optical camera, and infrared sensing are required to provide complementary information in addition to cooperative sensors. In this paper, networked frequency-modulated continuous wave (FMCW) radars and pan-tilt-zoom (PTZ) cameras construct the perception system. Optical cameras can capture the texture and visual information of targets, and radars complement the distance, speed, and additional information when the environment becomes challenging for optical cameras.

For fusing the radar and camera data, typical algorithms can be categorized into three levels: data level, feature level, and object level [3]. For the data-level fusion, the raw images and radar data are taken as the input into the same network and output consistent predictions [4] [5]. At the feature level, the intermediate features from separated sensors are concatenated for subsequent networks [6] [7], which can avoid the heterogeneity of raw data. The object level leverages the preliminary result of the single detector and additional strategies for fusion, e.g. evidence theory [8], gradient boosting [9] and deep association [3]. This work investigates object-level fusion and enables an efficient and universe pipeline regardless of the types of sensors.

The synchronized and calibrated data is processed by the YOLOv5-based detector at first. Raw radar points and processed image detections are then associated by a deep semantic association network which is trained with the contrastive loss. In this way, the correlated points from different sensors work as the measurements of the Kalman Filter to track and predict the target trajectory. As the trajectory management of the UAV must be in three-dimensional spaces, it is not sufficient to only predict trajectories in 2D image views. To this end, trajectory predictions in pairwise cameras with overlapping views will finally be recovered from 2D to 3D following a tracking-reconstruction scheme [10].

Contributions of this paper can be concluded as follows:

1) A deep semantic association network is designed for fusing radar points and image detections, then benefits the subsequent tasks, e.g. detection, tracking, and prediction.
2) Networked sensors are used to construct a radar-camera dataset for UAV flight scenarios, where different flight trajectories (circle, square, line, etc) at 4-6 meters altitude are planned. And the sensor network also contributes to reconstructing 3D trajectories from 2D observations in pairwise camera views.

II. Workflow of Radar and Camera Fusion

In this section, the workflow of multi-sensor fusion is demonstrated. The radar and camera data are processed at first. Detection and semantic association networks are then

© 2023 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

illustrated. And the tracking and prediction procedure is simply described.

A. Multi-sensor Calibration

Timestamp synchronization and spatial transformation are fundamental procedures for the fusion task. Because of the higher sample frequency of cameras than that of radars, the closest camera frame to each radar frame is selected for timestamp alignment.

In order to map radar points onto images, multi-sensor calibration is performed. We consider projecting points in radar frame $\mathcal{W}^r$ into image frame $\mathcal{W}^i$ directly without converting to world frame $\mathcal{W}$. The radar point $(r, \theta) \in \mathcal{W}^{rt}$ is firstly converted from polar to Cartesian coordinate system $\mathcal{W}^r$ by $x_r = r \cdot sin(\theta)$ and $y_r = r \cdot cos(\theta)$. z_r is set to 0 assuming the radar image plane is aligned with $x-y$ plane. In this case, the transformation from radar frame $\mathcal{W}^r$ to pixel $(u, v) \in \mathcal{W}^i$ can be formulated as:

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} R_{3\times3} & T_{3\times1} \\ O_{3\times1} & 1 \end{bmatrix} \begin{bmatrix} x_r \\ y_r \\ z_r \\ 1 \end{bmatrix} \tag{1}$$

where f_x, f_y, c_x, c_y represent focal lengths and principal points of the camera that are calibrated with a checkerboard. $R_{3\times3}$ and $T_{3\times1}$ are rotation and translation matrices.

With the corresponding point set in the radar and image coordinate frames, it is convenient to obtain the rotation and translation vectors from the SolvePnP algorithm [11].

B. Detection and Association

The overall architecture of radar-camera fusion is drawn in Fig. 1.

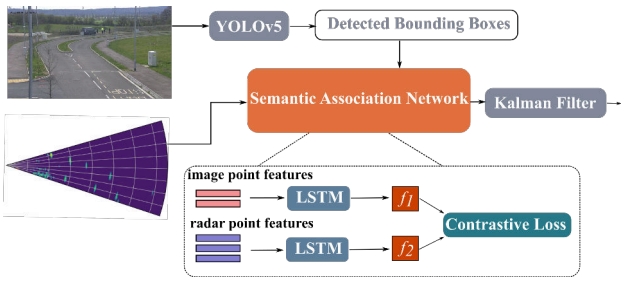


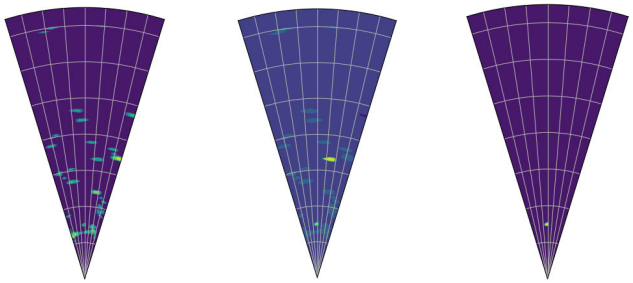
Fig. 1: Framework of camera and radar fusion.

1) Camera Detection: In the fusion framework, a YOLOv5 model is trained as the image detector, owning to its fast-speed inference and high accuracy. In the practical fusion procedure, each image is fed into the detector and generates detected bounding boxes. Depending on the image quality and target, not all targets can be correctly detected. And some false-positive results might be produced. As a result, single-camera detection is not so reliable. The features of detected bounding boxes (x and y coordinates of the center, and class prediction confidence) are collected for the next fusion task.

2) Radar Perception: The radar used in this work returns Range-Azimuth-Doppler information. The in-phase element and the quadrature element of the radar return can be converted to the magnitude in the Range-Azimuth image. In addition, the velocity can be obtained based on Doppler information. It is difficult to identify the real target based on the magnitude or velocity due to the noisy background objects, reflective characteristics, and target motion. To cope with this, the denoised map $\mathcal{D}$ is generated by stacking the magnitude map $\mathcal{M}$ and velocity map $\mathcal{V}$ with a speed limitation:

$$\mathcal{D} = \mathcal{M} \odot (\forall\, |\mathcal{V}| > v_{min}) \tag{2}$$

where the point whose speed is less than v_{min} will be ignored. $\odot$ represents the element-wise multiplication. A denoised sample is shown in Fig. 2, we can observe objects obviously after removing noises. In this module, there is no radar detection approach applied. All possible points in the denoised map, whether false alarms or real object points, will be fed into the network for the point-level association.



(a) Magnitude map. (b) Velocity map. (c) Denoised map.

Fig. 2: Radar data denoise.

3) Semantic Association: The semantic association network is displayed in Fig. 1. The features of the bounding box and radar points are fed into the two-stream long short-term memory (LSTM) networks, respectively. Each branch outputs the 2d position in a special coordinate frame. To associate the correct radar point with the corresponding bounding box, the contrastive loss [12] is utilized as a loss function as in Eq. (3). Contrastive learning works for clustering data from heterogeneous sensors.

$$\begin{aligned} L_{contrastive}(x_{c_i}, x_{r_j}, y) = \frac{1}{2} y \left\| f(x_{c_i}) - f(x_{r_j}) \right\|_2^2 \\ + \frac{1}{2}(1-y) \left\{ max(0, m - \left\| f(x_{c_i}) - f(x_{r_j}) \right\|_2) \right\}^2 \end{aligned} \tag{3}$$

where y is the binary label. If the camera bounding box x_{c_i} (in the image coordinate system) and radar point x_{r_j} (in the radar coordinate system) are from the same object, we set $y = 1$, and the distance between encoded features $f(x_{c_i})$ and $f(x_{r_j})$ is minimized. Otherwise, $y = 0$, and the different-class radar points and the bounding box will be separated by the network. As a consequence, the radar points and image bounding boxes are clustered according to the attribute. Especially, the network outputs $f(x_{c_i})$ and $f(x_{r_j})$

are located in a special two-dimensional embedding space $\mathbb{R}^2$ learned by the network.

Another important purpose of the semantic association is cross-validation. If the bounding box is predicted accurately, the radar points nearby will reinforce the confidence of detections, otherwise, the real detection should be assessed by tracking and prediction components.

Finally, the nearest distance is used to find the closed radar points to target camera point in the embedding space $\mathbb{R}^2$. The indices of all points will be recorded to retrieval the original data. For instance, if points $f(x_{c_1})$, $f(x_{r_1})$ and $f(x_{r_2})$ in the embedding space $\mathbb{R}^2$ are related. Then indices c_1, r_1 and r_2 help to obtain the associated camera detection x_{c_i} and raw radar points $x_{r_j}(j = 1, 2)$. The associated raw data will be used for tracking and forecasting in the next stage.

C. Tracking and Prediction

With reliable detection results, the typical Kalman Filter (KF) is leveraged to estimate target states and predict the trajectory. From the deep semantic association in the previous section, the associated radar point $x_{r_j}(j = 1, \ldots, n)$ and image bounding box x_c can be obtained. Raw radar points are projected to the image view and denoted by $x'_{c_j}(j = 1, \ldots, n)$. The fused measurement x_f is calculated by the weighted distance between the center of valid radar points and the target box center:

$$x_f = \alpha \cdot \frac{1}{n} \sum_{j=1}^{n} x'_{c_j} + (1 - \alpha) \cdot x_c \tag{4}$$

where $\alpha = max\left(0, \frac{1}{n}\sum_{j=1}^{n} x'_{c_j} - x_c\right)$ is determined by the distance value. The weight changes dynamically according to the distribution of radar points and the target bounding box. It is necessary to mention that the KF here works only as the tracker and predictor instead of fusion, as fusion is achieved by the deep association and weighted distance.

III. Experiments

In this section, the radar-camera dataset is collected for training and testing. And the training accuracy of YOLOv5 detection and semantic association are presented. The performance of trajectory prediction is also evaluated in test scenarios.

A. Dataset

To fulfill the radar-camera fusion, the dataset is fundamental for the study. We conducted several flight trials with a small UAV over the Multi-User Environment for Autonomous Vehicle Innovation (MUEAVI) road at Cranfield University. A set of heterogeneous sensors including radar, camera, and camera are integrated into the MUEAVI system. And we can easily collect the flight data with this system. In our flight trials, 3 radars and 7 cameras are used, and the layout of these sensors is displayed in Fig. 3. The camera has nearly infinite view distance, whereas the radar can detect objects within 150 meters.

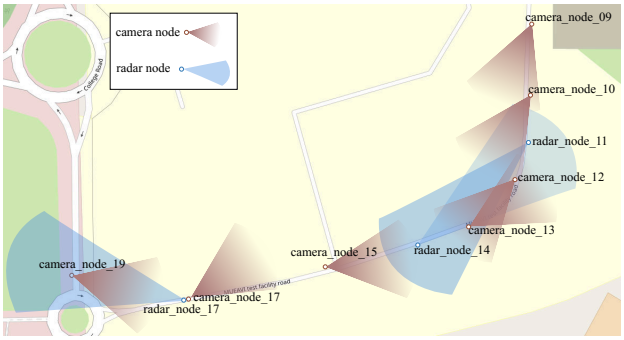


Fig. 3: Parts of radars and cameras layout on MUEAVI road.

The radar and camera can be used for fusion only if they have a common perception area. For instance, radar 17 only has the intersecting region with camera 19. In this case, we can just fuse the aligned data between radar 17 and camera 19, and there is no valid data for fusing radar 17 and camera 09. Data from radar 11 can be used for fusing with multiple cameras because it has overlapped observation areas with 6 different cameras. The advantage of overlapping views is the convenience of recovering 3D position from pairwise 2D views.

To this end, the realistic flight is performed by a single UAV to collect the data. We then select some data sequences from the flight trials to construct the dataset. Only those frames in which the target simultaneously appears in camera and radar are used. As shown in TABLE II, the dataset consists of 4201 camera and radar frames in total for all radar-camera combinations.

TABLE I: Frame number of each radar-camera combination

Camera Node \ Radar Node	11	14	17
09	89	-	-
10	657	728	-
12	501	452	-
13	70	-	-
15	773	673	-
17	102	84	-
19	-	-	72

The initial update rate of the radar is 15.625 Hz assuming 128 Doppler bins and then downgraded to 5-6 Hz to increase the Doppler resolution. And the camera capture frequency is 30 Hz. As the radar and camera are temporally aligned based on the radar sampling frequency, the synchronized data finally works in 5-6 frames per second.

B. Image Detector Training Result

The first step in our fusion framework is training an image detector with YOLOv5. And all 4201-frame images captured from different cameras are randomly split into train sets and test sets. It is expected to obtain a generalized model without considering the specific camera information.

The pre-trained image detector is then applied to the test set for evaluation. We can observe the 97.7% accuracy from TABLE II. And the means average precision (mAP) for an IoU (intersection over union) threshold of 0.5 achieves 94.8%, in addition to the 60.2% mAP over different IoU thresholds, from 0.5 to 0.95. Thus the trained model can be regarded as an accurate image detector for the fusion task.

TABLE II: Test results of trained YOLOv5 model

Epoch	Precision	Recall	mAP@.5	mAP@.5:.95
300	0.977	0.901	0.948	0.602

C. Semantic Association Training Result

The semantic association is the critical component in the fusion framework. The idea is to cluster the camera and radar points from the same target. After training, we can view the distributions from Fig. 4 that the projected features of radar (green color) and camera (blue color) points belonging to the drone are initially separated remotely. In contrast, radar and camera features in the right subplot can be pushed to be closed after the association. Especially, the long-tailed distribution of radar features arises from the performance difference of various radar configurations. Especially, x and y axes in Fig. 4 are determined by the 2D embedding space $\mathbb{R}^2$ of the network and show the distribution distinction.

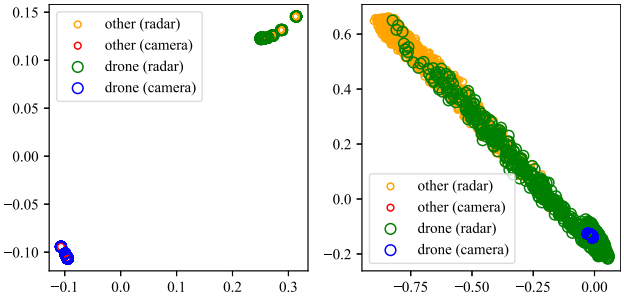


Fig. 4: Feature distributions before and after training the deep association network.

The semantic association can determine the class of all radar points. As is depicted in Fig. 5, all radar points from radar 11 and the image bounding box from camera 9 can be associated semantically, and the correct points from the drone are clustered. Those points are then drawn in the image frame for visualization.

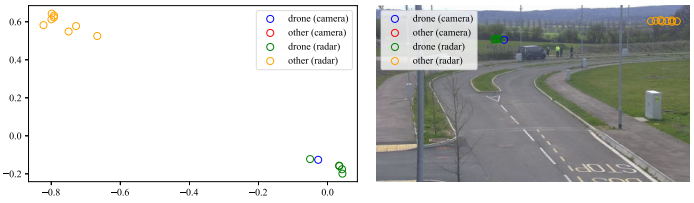


Fig. 5: Association example (left: association result; right: points in image frame).

D. Case Studies

With the performance of the image detector and semantic association guaranteed, the tracking and prediction performance after fusion is also necessary for evaluation with the Kalman Filter.

The root-mean-square error (RMSE) metric is calculated for measuring the difference between estimated trajectory and manually annotated ground truth, which is formulated by:

$$RMSE = \left[\sum_{i=1}^{N}(p_i - o_i)^2/N\right]^{1/2} \tag{5}$$

where p_i and o_i are estimated positions and corresponding ground truth of the i-th point in the trajectory. The unit of RMSE here is *pixel* in the image coordinate system.

Besides the KF predictor (*Fusion_KF*) with fused measurement input, as a comparison, another two KF predictors are also applied for single-sensor measurement, i.e. *Cam_KF* for camera prediction and *Rad_KF* only for radar input.

1) Evaluation in 2D image views: We start from evaluating the trajectory prediction in 2D image frames. In the plotting such as Fig. 6, the term "*Measurement*" in the figure denotes the fused input to *Fusion_KF*. Predicted trajectories from *Fusion_KF*, *Cam_KF* , *Rad_KF* and ground truth are drawn in the left coordinate frame and the right image frame. The x and y coordinates in these figures are pixel positions in the original image coordinate.

Quantity values of RMSE in TABLE III demonstrate the prediction performance of single-type sensors and various combinations of radar and camera. We can observe the better-fused results than predicting with the single-type sensor. An important reason is that the single sensor misses the object in some frames, and the KF predictors are not updated in these frames. Whereas the fused measurements can avoid this issue to some extent except the target signals of different sensors are lost at the same time.

TABLE III: RMSE Comparison (unit: *pixel* in image frame)

Camera Node	Radar Node	*Cam_KF*	*Rad_KF*	*Fusion_KF*
09	11	54.536	59.075	**54.170**
10	11	82.501	109.993	**82.134**
10	14	86.307	101.737	**86.103**
12	11	72.418	97.935	**72.156**
12	14	76.341	101.811	**76.186**
13	11	**87.572**	179.072	88.075
15	11	**78.827**	82.835	78.837
15	14	**38.519**	41.438	38.559
17	11	**53.773**	54.773	53.845
17	14	71.685	72.948	**71.675**
19	17	111.191	199.352	**110.826**

For visualization, predicted trajectories from camera 10, camera 12 and radar 14 are demonstrated in Fig. 6 and Fig. 7. The drone moves along with a circular path. Radars fail to get reflective signals at critical turning points. As a result, we can see that the predicted radar trajectories are not of good

quality. The fusion is then managing the advantage of camera measurements to improve the prediction performance.

When fusing the camera 15, radar 11 and radar 14, we get worse fusion results in contrast to *Cam_KF* in TABLE III. In these two scenarios, the error from radar measurements affects the final fusion quality. Similar issue also happens when fusing camera 17 and radar 11 as in in TABLE III. The enlarged radar projection error is caused by the drone movement relative to the radar. The radar is not capable of detecting the drone well at specific angle and altitude. However, for camera 17 and radar 14, as well as camera 19 and radar 17, the fused performances still surpass the single-sensor prediction.

2) Evaluation in 3D space: Then we assess the performance after trajectories are predicted in 2D image frames. The corresponding trajectories in pairwise 2D camera views are then reconstructed to 3D space following the workflow in [10]. Three scenarios different maneuvering scenarios are selected to explain the performance. For Scenario 1, flight trajectories in Fig. 6 and Fig. 7 are observed from the same flight trial. In other words, corresponding trajectories in camera 10 and camera 12 can be reconstructed to 3D space as plotted in Fig. 8. For Scenario 2, corresponding predictions in camera 10 (Fig.9) and camera 15 (Fig. 10) can be rebuilt to the 3D trajectory in Fig. 11. Similarly, for Scenario 3, 2D trajectories in Fig. 12 observed from camera 15 and Fig. 13 from camera 12 view contribute the recovered 3D trajectory in Fig.14.

In this way, predictions from single sensor and fusion are all reconstructed to 3D space, the RMSE in the realistic scale are compared in TABLE. IV. It is convenient to find that the fused trajectory in the three scenarios are still closer to the ground truth.

Finally, we can conclude that the fusion framework is effective to perform drone detection, tracking, and prediction tasks in different camera and radar views. And various combinations of radars and cameras in the sensor network are effective for the surveillance of UAM .

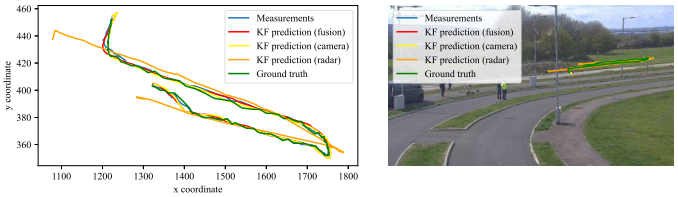


Fig. 6: Predictions in image frame (camera 10 and radar11).

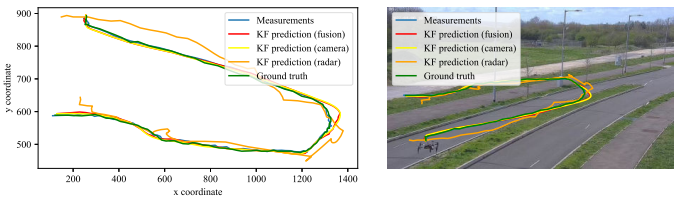


Fig. 7: Predictions in image frame (camera 12 and radar11).

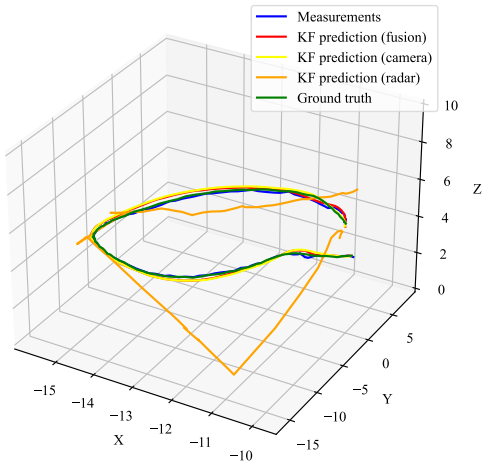


Fig. 8: Predictions in 3D frame (scenario 1).

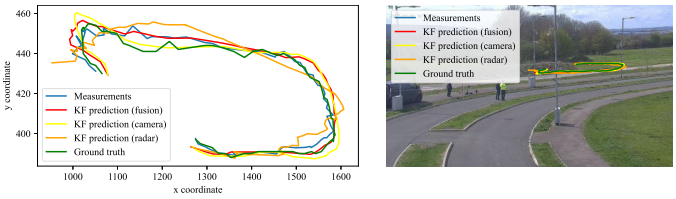


Fig. 9: Predictions in image frame (camera 10 and radar14).

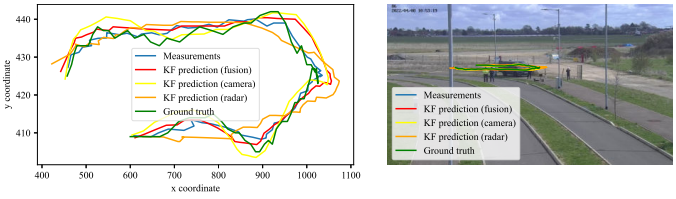


Fig. 10: Predictions in image frame (camera 15 and radar14).

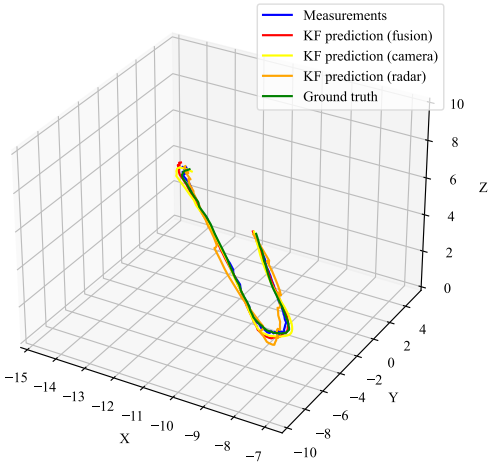


Fig. 11: Predictions in 3D frame (scenario 2).

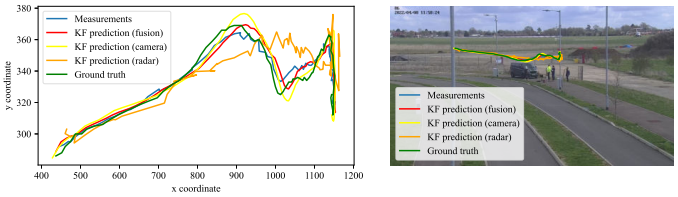


Fig. 12: Predictions in image frame (camera 15 and radar14).

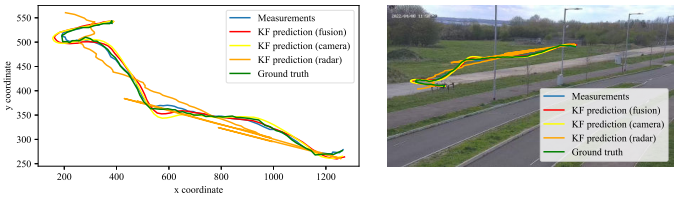


Fig. 13: Predictions in image frame (camera 12 and radar14).

IV. CONCLUSIONS

In this work, the detection, tracking, and trajectory prediction of the small UAV are studied with the radar-camera fusion, to promote the surveillance of cooperative and non-cooperative aerial objects, as a consequence, ensuring flight safety and security in urban air mobility. A fast fusion framework is illustrated, with a YOLOv5-based image detector, a semantic association network, and the Kalman Filter for tracking and prediction. Especially, the deep association network is the kernel component for relating the correct points in radar and camera frames. Without correct association, the prediction will be polluted by error-fused measurements. And the results from different flight trials and various combinations of camera and radar can prove the effectiveness of the fusion framework.

There are still some improvements to be performed in the future. For instance, current short-term trajectory prediction is not able to be used by tactical conflict detection since the drone in the urban environment can maneuver fast, and the intention-based long-term trajectory prediction is required to forecast the trajectory conflict in advance.

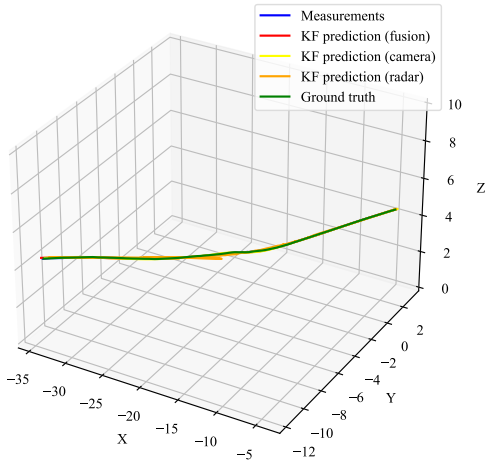


Fig. 14: Predictions in 3D frame (scenario 3).

TABLE IV: RMSE Comparison of 3D trajectories (unit: *meter*)

Scenario	1	2	3
Camera Node 1	10	10	15
Radar Node 1	11	14	14
Camera Node 2	12	15	12
Radar Node 2	11	14	14
Cam_KF	0.145	0.211	0.096
Rad_KF	2.035	0.428	1.477
Fusion_KF	**0.122**	**0.200**	**0.067**

ACKNOWLEDGMENT

This research was partially supported by grants from the Funds of China Scholarship Council (202008420248).

REFERENCES

[1] "Urban air mobility (uam) - concept of operations v1.0," Federal Aviation Administration, Tech. Rep. 1-37, 2020.

[2] S. Hening, C. A. Ippolito, K. S. Krishnakumar, V. Stepanyan, and M. Teodorescu, "3d lidar slam integration with gps/ins for uavs in urban gps-degraded environments," in *AIAA Information Systems-AIAA Infotech@ Aerospace*, 2017, p. 0448.

[3] X. Dong, B. Zhuang, Y. Mao, and L. Liu, "Radar camera fusion via representation learning in autonomous driving," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 1672–1681.

[4] F. Nobis, M. Geisslinger, M. Weber, J. Betz, and M. Lienkamp, "A deep learning-based radar and camera sensor fusion architecture for object detection," in *2019 Sensor Data Fusion: Trends, Solutions, Applications (SDF)*. IEEE, 2019, pp. 1–7.

[5] Y. Cheng, H. Xu, and Y. Liu, "Robust small object detection on the water surface through fusion of camera and millimeter wave radar," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 15 263–15 272.

[6] L.-q. Li and Y.-l. Xie, "A feature pyramid fusion detection algorithm based on radar and camera sensor," in *2020 15th IEEE International Conference on Signal Processing (ICSP)*, vol. 1. IEEE, 2020, pp. 366–370.

[7] R. Nabati and H. Qi, "Centerfusion: Center-based radar and camera fusion for 3d object detection," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 1527–1536.

[8] P. Liu, G. Yu, Z. Wang, B. Zhou, and P. Chen, "Object classification based on enhanced evidence theory: Radar–vision fusion approach for roadside application," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–12, 2022.

[9] K. Kowol, M. Rottmann, S. Bracke, and H. Gottschalk, "Yodar: uncertainty-based sensor fusion for vehicle detection with camera and radar sensors," *arXiv preprint arXiv:2010.03320*, 2020.

[10] Z. Wu, N. I. Hristov, T. H. Kunz, and M. Betke, "Tracking-reconstruction or reconstruction-tracking? comparison of two multiple hypothesis tracking approaches to interpret 3d object motion from several camera views," in *2009 Workshop on Motion and Video Computing (WMVC)*. IEEE, 2009, pp. 1–8.

[11] V. Lepetit, F. Moreno-Noguer, and P. Fua, "Epnp: An accurate o (n) solution to the pnp problem," *International journal of computer vision*, vol. 81, no. 2, pp. 155–166, 2009.

[12] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, vol. 2. IEEE, 2006, pp. 1735–1742.

https://dspace.org/

Radar-camera fusion for ground-based perception of small UAV in urban air mobility

Huang, Cheng; Petrunin, Ivan; Tsourdos, Antonios

2023-07-27
Institute of Electrical and Electronics Engineers (IEEE)
https://doi.org/10.1109/MetroAeroSpace57412.2023.10189934
Huang C, Petrunin I, Tsourdos A. (2023) Radar-camera fusion for ground-based perception of small UAV in urban air mobility. In: 2023 IEEE International Workshop on Metrology for AeroSpace (MetroAeroSpace), 19-21 June 2023, Milan, Italy

Downloaded from a DSpace repository. https://dspace.org/