Robust Multiobject Tracking Using MmWave Radar-Event-Camera Sensor Fusion

Leonard Haensel
HELLA GmbH & Co. KGaA
Lippstadt, Germany
leonard.haensel@forvia.com

Torsten Bertram
TU Dortmund University
Dortmund, Germany
torsten.bertram@tu-dortmund.de

***Abstract*— This paper introduces a novel track-to-track sensor fusion framework that integrates event-based camera tracks and radar detection tracks for robust multi-object tracking. Due to the lack of an available dataset combining real event-camera data with millimeter-wave (mmWave) radar in the automotive context, we utilize a state-of-the-art radar-camera dataset, adapting it to simulate event-based camera data. Event-based cameras, with their high temporal resolution and low energy consumption, provide asynchronous brightness change data, while radar sensors offer robustness to adverse environmental conditions and precise depth measurements. The proposed method leverages the complementary strengths of these two sensor modalities using a Global Nearest Neighbor (GNN) algorithm to associate tracks and maintain a single hypothesis about tracked objects. The event-camera's optical flow-based velocity and distance estimation is fused with radar's depth and lateral position measurements to enhance localization accuracy. This approach demonstrates the potential for improving tracking performance and robustness in autonomous driving scenarios.**

Keywords*— *autonomous driving, sensor fusion, event-based camera, target tracking, radar

I. Introduction

The rapid advancements in autonomous systems and intelligent vehicles have highlighted the critical importance of robust multi-object tracking for safe and efficient navigation. Sensor fusion has emerged as a pivotal approach to overcome the limitations of individual sensors by leveraging their complementary strengths. Among these, event-based cameras and millimeter-wave (mmWave) radar sensors offer unique advantages. Event-based cameras, inspired by biological vision systems, operate asynchronously by capturing brightness changes at a high temporal resolution with low energy consumption, making them ideal for dynamic environments. However, they face challenges in static or low-dynamic scenes due to sparse outputs [1]. On the other hand, mmWave radar sensors provide robust depth measurements and operate reliably under adverse lighting and weather conditions but suffer from lower angular resolution compared to optical sensors.

Despite the promise of combining event-based cameras and radar for enhanced multi-object tracking, there is currently no dataset that provides real event camera data alongside mmWave radar data in an automotive context. To address this limitation, this work utilizes artificial event data generated from the CARRADA Dataset [2]. The artificial events are created using upsampled camera frames processed through an event simulator, enabling the integration of event-based camera tracks with radar detection tracks [3], [4].

This paper is an extension of our previous paper in which we presented a preliminary study on distance and speed estimation using event-based vision [5]. Here we present a novel track-to-track fusion framework that combines artificial event-based camera tracks and radar detection tracks for enhanced multi-object tracking. By integrating these two sensor modalities, the proposed approach addresses key challenges such as improving localization accuracy, reducing missed detections, and ensuring robustness to single-sensor failures. The system employs a Global Nearest Neighbor (GNN) algorithm for track association. Radar's precise depth measurements are fused with the event camera's high-resolution lateral position data to achieve accurate 2D spatial localization. Additionally, coordinate transformations align sensor data into a unified spatial representation. The proposed methodology demonstrates significant improvements in tracking metrics such as Multi-Object Tracking Accuracy (MOTA), with reduced false negatives and enhanced robustness in different scenarios [6]. This work highlights the potential of combining bio-inspired vision systems with radar technology for applications in autonomous driving, robotics, and other safety-critical domains.

II. Related Work

The contribution at hand, builds on existing approaches to sensor data fusion, particularly in the context of combining event cameras and mmWave radar sensors for object detection and tracking. While the fusion of conventional cameras and radar sensors has already been extensively studied, the integration of event cameras represents a comparatively new research direction due to their asynchronous operation and high temporal resolution [7], [8], [9]. Safa et al. presented one of the first papers on the fusion of event camera and radar data in the context of Simultaneous Localization and Mapping (SLAM) [10]. In their study, a drone is used to both create a map of a warehouse and estimate its own position within this map. The sensor data is fused using a Spiking Neural Network (SNN), a bio-inspired architecture that processes information in the form of discrete pulses rather than relying on continuous data streams. This enables energy-efficient and robust data processing. The results show that the system is robust to illumination changes and in some cases achieves lower error rates compared to SLAM algorithms with conventional cameras.

Müller et al. presented a paper focusing on the fusion of event camera and radar data for body gesture recognition of ground controllers on the apron of an airport [11]. For this purpose, a specially developed dataset was created that includes different body poses at different distances, angles and environmental conditions. The sensor data was fused using an artificial neural network, which led to a significant improvement in recognition accuracy. In addition, the authors presented a novel coding method that efficiently compresses the raw radar data, reducing the data rate and simplifying implementation. This study highlights the potential of fusing radar and event data for more accurate and robust detection systems.

Wang et al. presented a system for precise and low-latency drone localization that combines the advantages of mmWave radar and event cameras [12]. In contrast to traditional approaches that use frame cameras, the event camera harmonizes better with the mmWave sensor technology due to its higher sampling rate. The presented system, mmWave-Event-Localization (mmE-Loc), integrates two central modules: a consistency-based collaborative tracking and a graph-informed adaptive optimization. These modules utilize the temporal consistency and spatial complementarity of the sensors to extract precise measurements and enable efficient sensor data fusion. In experimental tests conducted under real-world conditions, mmE-Loc showed superior results in terms of localization accuracy and latency compared to existing methods.

III. Proposed Methods

In this contribution, we extend a previously proposed method for estimating real-world distances and velocities of objects based on artificially generated events [5]. For this purpose, we present an object-based sensor fusion approach, which combines the distance and velocity estimates of the previous paper with the corresponding radar data. [2], [5]. First, each processing branch is considered individually by radar and artificial event data before the respective tracks of the two branches are fused using a GNN approach in combination with a cross-covariance fusion algorithm. To provide an overview of the system, Fig. 1. illustrates the different processing steps of the respective branches.

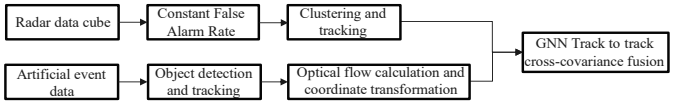

Fig. 1. General system overview of the tracking-fusion framework in conjunction with the radar and artificial event branch.

The initial step involves generating a point cloud from the raw radar data stored in the radar data cube. This is achieved using the Constant False Alarm Rate (CFAR) algorithm, which identifies the most intense reflections while accounting for their local context [13]. After that we transform the range-angle represented points into a cartesian coordinate system. The resulting points then correspond to a sparse point cloud in a 2D coordinate space similar to the bird's eye view representation. The representation is enhanced using the recorded Doppler for each point. The 3D point cloud combines the cartesian coordinates of the reflected point and its doppler value, which corresponds to the radial velocity component. The distance-based cluster algorithm Density Based Spatial Clustering of Applications with Noise (DBScan), which was already proposed by Nordenmark et al. for radar data processing, is used to group the points into clusters [14]. The center of gravity is calculated by a weighted averaging of the points within the cluster, whereby the weights are derived from the inverse of the measurement noise covariance matrices. The clustered object is then tracked over time using a Joint Probabilistic Data Association (JPDA) tracker [15]. Additionally, a Kalman filter (KF) is used to initialize and update the state estimates for each track. To determine the characteristics of each cluster a weighted average is applied to the included radar points. This procedure is inspired by Zhang et al. and follows the equation in (1) [16].

$$\dot{x} = \frac{\sum wx}{\sum w} \tag{1}$$

Here, $\dot{x}$ and x are the characteristics of the clustered radar object. The characteristics include distance, position and radial velocity. The weight w represents the reflective power of the respective radar point within the cluster.

At the same time, the optical flow is determined from the artificial event data using the trajectory-based methods from our previous work [5]. A feature vector is calculated from this, which contains the estimated position, radial speed, detected object class and unique track. This feature vector then serves as the basis for the fusion with the previously calculated features of the radar pipeline. Similar to the radar tracks, a KF is used to initialize and update the state estimates for each track.

With regard to the assignment of the individual tracks, the distance between the radar targets and the event camera objects is calculated first. The Mahalanobis distance was chosen as the distance metric [17]. The advantage of this method is that it takes the correlation and scaling of the data into account compared to the Euclidean distance. The distance metric d_{ecr} is defined as follows and is shown in (2):

$$d_{ecr} = (x_{ec} - x_r)^T s^{-1} (x_{ec} - x_r) \tag{2}$$

Here, x_{ec} and x_{r} are the characteristics of the event camera ($_{\mathrm{ec}}$) and radar objects ($_{\mathrm{r}}$). The states include longitudinal position, lateral position and longitudinal velocity as well as the corresponding track ID. d_{ecr} is the distance matrix, which represents the similarity between two objects in vector space. s is the covariance matrix, which is used to scale the distance matrix based on the distribution of the data. The next step is the actual fusion, where the radar objects are assigned to their corresponding event camera objects. The GNN algorithm was selected for association. The algorithm finds the data pair with the minimum total distance. The Hungarian method was used to find an optimal solution [18]. This results in three possible outcomes, which are divided into matched detection-track pairs, unmatched tracks and unmatched detections. For the matched detection track pairs, the actual track fusion is performed using a cross-covariance fusion algorithm that

considers the uncertainties of the different sensor modalities. The cross-covariance fusion algorithm uses a weighted combination of the states, where the weights are determined by the inverse covariance matrices. The mathematical definition is shown in (3).

$$x_f = s_f{}^{-1}(s_r{}^{-1}x_r + s_{ec}{}^{-1}x_{ec}) \quad (3)$$

Here, x_f represents the merged state of both weighted states of x_r and x_{ec}. This is then weighted again with the combined covariance matrix s_f from both states, resulting in the fused track weighted from both uncertainties. If there are no assigned measurement or recognition values for a unreached track, the track is updated with the predicted value of the KF to provide continuous estimated tracking. Additionally, a running detection loss counter is incremented by a single frame. If the radar track is undetectable for 20 consecutive frames, the track is deleted. By generating the artificial event data and the associated tenfold upsampling process, the detection loss counter on the part of the event data was increased to 200. For the third scenario, i.e. unmatched detections, new track IDs are assigned and tracks are initialized. While all of the prediction stages start immediately from the next frames for new tracks, they are not displayed until they are classified as reliable tracks. This is the case when the track has been previously confirmed on 5 consecutive radar frames or 50 event frames and follows the model of Sengupta et al. [19].

IV. Experiments

To ensure an objective and standardized evaluation, we rely on the widely recognized Clear Multi-Object Tracking (MOT) metrics, which were specifically developed for the evaluation of multi-object tracking systems [6]. We explicitly limit ourselves here to multi-object tracking accuracy (MOTA) and refer to our previous paper for multi-object tracking precision (MOTP) [5]. MOTA is a metric that determines the overall accuracy of the tracking system, taking into account various sources of error. MOTA combines the number of missed objects or unrecognized objects (FN), the number of false alarms or incorrectly recognized objects (FP) and the number of ID switches (IDSW) into a uniform metric. The IDSW is an incorrect assignment of track IDs over time. The relationship is shown in (4) and always refers to the ground truth value within the scene that is currently displayed on image *n*.

$$MOTA = 1 - \frac{\sum_n FN_n + \sum_n FP_n + \sum_n IDSW_n}{\sum_n GT_n} \quad (4)$$

The individual terms represent the respective proportions of FN, FP and IDSW. These are defined as a rate and indicate the proportion of the respective error sources to the ground truth value and are shown in (5).

$$FNR = \frac{\sum_n FN_n}{\sum_n GT_n}, FPR = \frac{\sum_n FP_n}{\sum_n GT_n}, IDSWR = \frac{\sum_n IDSW_n}{\sum_n GT_n} \quad (5)$$

To generate the ground truth values, manual annotation of the upsampled CARRADA test dataset with over 21,530 images in relation to the artificial events is required. In addition, the resulting 2,153 radar point clouds were labeled and manually corrected according to the annotation pipeline of Zhang et al. [20]. The results are shown in detection. Once a match is
TABLE I. The individual scenes 1 - 6 are listed chronologically according to their recording date in the test dataset. The metrics show that the track-to-track fusion method proposed here leads to a significant improvement compared to the tracking of the individual sensors. This is particularly visible in relation to the FNR (missed objects). The average improvement here is 7.94 % compared to pure event-based tracking. Only the first scenario shows a deterioration in performance. The high uncertainty of the radar can only be compensated for to a limited extent. The higher uncertainties result in each case from a more complex scene structure in which different vehicles and pedestrians interact with each other and the reduced horizontal resolution of the radar cannot resolve the closely spaced objects. The elevated FPR in the sensor fusion system arises when radar and event camera detections from the same physical object are misclassified as independent entities due to association failures between the two sensing modalities. One reason is the high volatility and discontinuity of radar points, causing the center of gravity of the radar point cluster to frequently jump back and forth in front of and behind the target. Another critical factor is the extraction of the distance and velocity information based on the artificial event data from our previous work. The deviation between the estimated and the real radar value is sometimes more than 4 m discrepancy in the distance and up to 5 m/s in the radial velocity component. With regard to the IDSW, we have a significant improvement in direct comparison to the event data. This can mainly be explained by the robust detection of static objects by the radar. The event object detector has problems detecting static objects over a longer period of time due to the lack of changes. Here, the combination of both tracks leads to an average improvement of 4.21 %. Only in highly concentrated scenes with many different objects can the high uncertainty on the part of the radar no longer be compensated for, similar to the detection of the FP. The combination using the approach presented here leads to an improvement of 7.23 % MOTA compared to the purely event-based approach and 20.17 % compared to pure radar detection and thus represents a significant improvement in tracking.

V. Conclusions and Outlook

The contribution at hand presents a method for object-based sensor fusion. To our knowledge, this is the first contribution in the field of sensor fusion using event and mmWave data in regard to the automotive sector. Due to the lack of a suitable dataset, artificial event data is used. First, both sensor modalities are considered separately and the corresponding features are calculated. On the one hand, this involves generating a point cloud based on the raw radar data as well as clustering and tracking the object. On the other hand, the features are determined from the artificial event data using a trajectory-based approach. These features are then projected into a common feature space. Subsequently, data association is performed using a GNN approach in conjunction with the Mahalanobis distance and the Hungarian method. This results in three different outputs: matched detection-track pairs, unmatched tracks and unmatched detection. Once a match is

TABLE I. CLEAR-MOT Metrics for the CARRADA testset

Scene	Event-Camera Only				Radar Only				Sensor Fusion			
	***FNR* (%)**	***FPR* (%)**	***IDSWR* (%)**	***MOTA* (%)**	***FNR* (%)**	***FPR* (%)**	***IDSWR* (%)**	***MOTA* (%)**	***FNR* (%)**	***FPR* (%)**	***IDSWR* (%)**	***MOTA* (%)**
1	1.56	6.11	5.33	87.00	55.11	5.14	2.11	37.64	3.20	10.13	3.33	83.34
2	24.71	6.43	7.25	61.61	27.34	3.74	1.13	67.79	3.61	19.47	1.28	75.64
3	19.57	11.72	8.39	60.32	24.35	6.19	2.19	67.27	9.87	10.29	2.83	77.01
4	6.14	13.34	4.20	76.32	39.52	4.17	1.15	55.16	2.15	23.60	1.34	72.91
5	13.79	10.57	5.17	70.47	44.41	3.78	2.64	49.17	3.78	12.11	2.87	81.24
6	11.58	13.82	8.55	66.05	27.57	4.15	1.17	67.11	7.11	15.88	1.98	75.03
Overall[a]				**70.30**				**57.36**				**77.53**

[a.] Average value

found, the actual fusion is performed using a cross-covariance fusion approach. The evaluation shows that a combination of both sensors has a significant improvement over the individual modalities explicitly in the avoidance of FN and the overall MOTA performance.

Nevertheless, explicitly overcrowded scenarios show that the reduced horizontal resolution of the radar can only be compensated for to a limited extent. More complex approaches such as deep learning or a lower fusion level on both sensor modalities are required here [9]. Another important point is the generation of features from the event data. Here, our first publication already shows a deviation from the radar features. A more sophisticated approach should also be considered here. For example, instance segmentation could be used [21]. At the same time, there is also the possibility of clustering the event point cloud to be more robust [22]. In addition, validation with real events and adaptation to dynamic scenes is essential for practical use in the context of autonomous driving.

REFERENCES

[1] G. Gallego *et al.*, "Event-Based Vision: A Survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 1, pp. 154–180, Jan. 2022, doi: 10.1109/TPAMI.2020.3008413.

[2] A. Ouaknine, A. Newson, J. Rebut, F. Tupin, and P. Pérez, "CARRADA Dataset: Camera and Automotive Radar with Range-Angle-Doppler Annotations," May 26, 2021, *arXiv*: arXiv:2005.01456. Accessed: Feb. 23, 2024. [Online]. Available: http://arxiv.org/abs/2005.01456

[3] Y. Hu, S. -C. Liu and T. Delbruck, "v2e: From Video Frames to Realistic DVS Events," 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Nashville, TN, USA, 2021, pp. 1312-1321, doi: 10.1109/CVPRW53098.2021.00144.

[4] H. Jiang, D. Sun, V. Jampani, M.-H. Yang, E. Learned-Miller, and J. Kautz, "Super SloMo: High Quality Estimation of Multiple Intermediate Frames for Video Interpolation," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2018, pp. 9000–9008. doi: 10.1109/CVPR.2018.00938.

[5] L. Haensel and T. Bertram, "A Preliminary Study on Distance and Speed Estimation Using Event-Based Vision," in *2024 IEEE 18th International Conference on Application of Information and Communication Technologies (AICT)*, Sep. 2024, pp. 1–5. doi: 10.1109/AICT61888.2024.10740303.

[6] K. Bernardin and R. Stiefelhagen, "Evaluating Multiple Object Tracking Performance: The CLEAR MOT Metrics," *EURASIP J. Image Video Process.*, vol. 2008, no. 1, Art. no. 1, Dec. 2008, doi: 10.1155/2008/246309.

[7] A. Palffy, J. F. P. Kooij, and D. M. Gavrila, "Detecting Darting Out Pedestrians With Occlusion Aware Sensor Fusion of Radar and Stereo Camera," *IEEE Trans. Intell. Veh.*, vol. 8, no. 2, pp. 1459–1472, Feb. 2023, doi: 10.1109/TIV.2022.3220435.

[8] Y. Kim, S. Kim, J. Shin, J. W. Choi, and D. Kum, "CRN: Camera Radar Net for Accurate, Robust, Efficient 3D Perception," Aug. 16, 2023, *arXiv*: arXiv:2304.00670. Accessed: Dec. 19, 2023. [Online]. Available: http://arxiv.org/abs/2304.00670

[9] S. Yao *et al.*, "Radar-Camera Fusion for Object Detection and Semantic Segmentation in Autonomous Driving: A Comprehensive Review," 2023, doi: 10.48550/ARXIV.2304.10410.

[10] A. Safa *et al.*, "Fusing Event-based Camera and Radar for SLAM Using Spiking Neural Networks with Continual STDP Learning," Oct. 09, 2022, *arXiv*: arXiv:2210.04236. Accessed: Dec. 08, 2023. [Online]. Available: http://arxiv.org/abs/2210.04236

[11] L. Müller *et al.*, "Aircraft Marshaling Signals Dataset of FMCW Radar and Event-Based Camera for Sensor Fusion," in *2023 IEEE Radar Conference (RadarConf23)*, San Antonio, TX, USA: IEEE, May 2023, pp. 01–06. doi: 10.1109/RadarConf2351548.2023.10149465.

[12] H. Wang *et al.*, "Ultra-High-Frequency Harmony: mmWave Radar and Event Camera Orchestrate Accurate Drone Landing," Feb. 20, 2025, *arXiv*: arXiv:2502.14992. Accessed: Feb. 27, 2025. [Online]. Available: http://arxiv.org/abs/2502.14992

[13] H. Rohling, "Radar CFAR Thresholding in Clutter and Multiple Target Situations," *IEEE Trans. Aerosp. Electron. Syst.*, vol. AES-19, no. 4, pp. 608–621, Jul. 1983, doi: 10.1109/TAES.1983.309350.

[14] S. Lim, S. Lee, and S.-C. Kim, "Clustering of Detected Targets Using DBSCAN in Automotive Radar Systems," in *2018 19th International Radar Symposium (IRS)*, Jun. 2018, pp. 1–7. doi: 10.23919/IRS.2018.8448228.

[15] T. Fortmann, Y. Bar-Shalom, and M. Scheffe, "Sonar tracking of multiple targets using joint probabilistic data association," *IEEE J. Ocean. Eng.*, vol. 8, no. 3, pp. 173–184, Jul. 1983, doi: 10.1109/JOE.1983.1145560.

[16] L. Ruoyu, Z. Darui, Y. Hang, W. Daihan, B. Ning, and Z. Jianguang, "A Data-Driven Radar Object Detection and Clustering Method Aided by Camera," SAE International, Warrendale, PA, SAE Technical Paper 2020-01–5035, Feb. 2020. doi: 10.4271/2020-01-5035.

[17] R. De Maesschalck, D. Jouan-Rimbaud, and D. L. Massart, "The Mahalanobis distance," *Chemom. Intell. Lab. Syst.*, vol. 50, no. 1, pp. 1–18, Jan. 2000, doi: 10.1016/S0169-7439(99)00047-7.

[18] X. Tian, T. Yuan, and Y. Bar-Shalom, "Track-to-Track Fusion in Linear and Nonlinear Systems," in *Advances in Estimation, Navigation, and Spacecraft Control*, D. Choukroun, Y. Oshman, J. Thienel, and M. Idan, Eds., Berlin, Heidelberg: Springer, 2015, pp. 21–41. doi: 10.1007/978-3-662-44785-7_2.

[19] A. Sengupta, L. Cheng, and S. Cao, "Robust Multiobject Tracking Using Mmwave Radar-Camera Sensor Fusion," *IEEE Sens. Lett.*, vol. 6, no. 10, pp. 1–4, Oct. 2022, doi: 10.1109/LSENS.2022.3213529.

[20] L. Zhang *et al.*, "PeakConv: Learning Peak Receptive Field for Radar Semantic Segmentation," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Vancouver, BC, Canada: IEEE, Jun. 2023, pp. 17577–17586. doi: 10.1109/CVPR52729.2023.01686.

[21] T. Bolten, R. Pohle-Fröhlich, and K. Tönnies, "Instance Segmentation of Event Camera Streams in Outdoor Monitoring Scenarios:," in *Proceedings of the 19th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, Rome, Italy: SCITEPRESS - Science and Technology Publications, 2024, pp. 452–463. doi: 10.5220/0012369100003660.

[22] A. Mondal, S. R, J. H. Giraldo, T. Bouwmans, and A. S. Chowdhury, "Moving Object Detection for Event-based Vision using Graph Spectral Clustering," in *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, Oct. 2021, pp. 876–884. doi: 10.1109/ICCVW54120.2021.00103.