Classification of Tracked Objects Using Multiple Frame Processing for Automotive Radar

Hassan, Mujtaba ; Fioranelli, Francesco ; Yarovoy, Alexander; Chen, Lihui ; Ravindranath, Satish; Wu, Ryan

DOI
10.23919/EuRAD61604.2024.10734928

Publication date
2024

Document Version
Final published version

Published in
Proceedings of the 2024 21st European Radar Conference (EuRAD)

Classification of Tracked Objects Using Multiple Frame Processing for Automotive Radar

Mujtaba Hassan[#*], Francesco Fioranelli[#], Alexander Yarovoy[#], Lihui Chen[$], Satish Ravindran[$], Ryan Wu[$]

[#]EEMCS, Delft University of Technology, The Netherlands
[*]Advanced Radar Solutions, NXP Semiconductors, Germany
[$]Advanced Radar Solutions, NXP Semiconductors, USA
{s.m.hassan, f.fioranelli, a.yarovoy}@tudelft.nl, {phil.chen_1, satish.ravindran, ryan.wu}@nxp.com

***Abstract*— A neural network (NN) based multi-frame classification approach is proposed to solve the problem of classification of tracked objects. Initially, a baseline tracker is implemented that uses the classification output of an object detection network for classification. Afterwards, two approaches for multi-frame classification are applied to perform classification of tracked objects. The first approach aggregates points from multiple frames and applies a single frame NN for classification, whereas the second approach uses bidirectional long short term memory (BiLSTM) layers to process points from multiple frames. Extensive experiments on the opensource 2D RadarScenes dataset showed a consistent increase in track performance when using either of the two techniques for multi-frame classification.**

***Keywords*— BiLSTM, classification, MOTA, radar, tracking.**

I. Introduction

In autonomous driving, multi-object tracking is a vital component of the perception stack that is required for several downstream tasks such as trajectory prediction and motion planning. Classification is an important sub-component of this tracking module. It not only allows a higher degree of scene understanding by enabling the system to recognize objects present in the scene, but can also assist in tracking [1].

A common method for tracking used frequently in autonomous driving research is to apply a NN based object detector that gives the location and class of each detected object, and then track these objects as specific classes [2]. A major problem with this approach for radar data is that single frame object classification is not robust, since radar data is sparse and contains a lot of clutter. In our previous work [3], we showed that tracking all objects together as a generic class can give better tracking performance as compared to tracking objects as specific classes. However, to determine the classification of the tracked object, in that work, we directly used the classification output of the object detection network which gave us incorrect classification results for many frames.

In this paper, we propose to solve the problem of classification of tracked objects by using multiple frames. The idea is that since we are tracking objects, we can use the points detected at previous frames to improve the classification of the tracked objects at the current timestamp. This will provide the classification network with more points that helps in improving the classification of the tracked objects. Fig. 1 shows a general pipeline of our proposed method. We investigated two different NN based methods for classification of tracked objects. In the first method, we aggregated points detected at different timestamps for each tracked object. This increases the number of detected points per tracked object. Afterwards, we used a NN inspired from [4] to provide classification of the tracked object. In the second method, we designed a multi-frame NN whereby we extracted features at each frame for the tracked object and then processed these features using BiLSTM layers [5] for classification. To the best of our knowledge, the approach of tracking all objects obtained from a NN object detector as generic class and classifying them using a NN multi-frame classifier is unseen, and is the novelty of this work.

We evaluated our proposed classification methods on the opensource 2D RadarScenes dataset [6]. Our proposed solution outperforms the baseline tracker using single frame classification by providing an increase in MOTA [7] of 4.19% for car class, 0.80% for pedestrian class, and 22.31% for cyclist class. The main contributions of the paper are as follows: (1) We designed a tracker that uses multiple frames for classification of tracked objects which showed a significant improvement in MOTA as compared to a tracker using single frame. (2) We compared the performance of trackers using two different multi-frame processing approaches for classification and showed that the BiLSTM based approach gives slightly better results than the point aggregation based approach except for the case of pedestrians.

The rest of this paper is organized as follows: Section II briefly reviews the background. Section III illustrates the approach used in this work. Section IV describes the experiments performed with a discussion on results. Section V concludes this work with a statement on future directions.

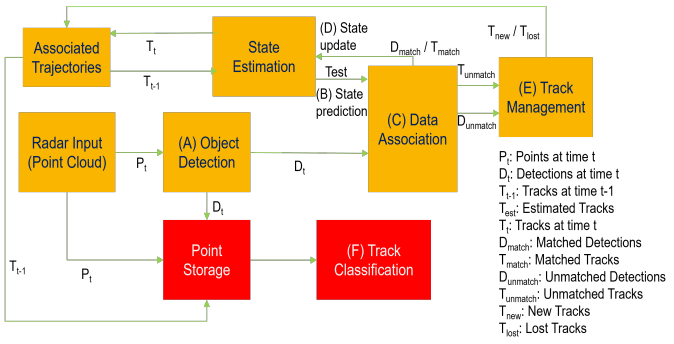


Fig. 1. Proposed multi-object tracker with multi-frame classification for automotive radar. The difference from baseline method is highlighted in red.

II. Background

Traditionally, mathematical models of extended objects are used for joint tracking and classification [8]. However, it is difficult to correctly model the different characteristics of the radar point cloud such as distribution of points, distribution of doppler, and radar cross section (RCS) based on hand crafted features. Conversely, a NN can learn to model these complex characteristics based on training data. As a result, most of the state of the art approaches for classification in automotive radar are based on deep learning [9]. [10] uses LSTM to classify objects on time series data but operate on spectrum instead of point cloud. Readers are referred to [11] for a detailed review.

III. Proposed Approach

We developed a multi-object tracker inspired by [2]. The key differences from the baseline model include: object detection using radar object detector for automotive radar point cloud instead of lidar point cloud; tracking objects as a generic class instead of tracking as specific classes; classification of tracked objects using multiple frames instead of single frame.

A. Proposed Processing Pipeline

Fig. 1 shows the pipeline of the proposed approach. Given an input point cloud, the first step is radar object detection. We used radar Pointpillars [12] that provides both localization and classification of the desired objects. However, we do not distinguish different classes of objects and instead track all of the detected objects together, which was found to provide better results in our previous work [3]. The next step is state estimation where we used a constant velocity motion model to predict track states at next frame. This is a reasonable model because of the high frame rate of the sensor and allowed us to use a linear Kalman filter [13]. Detected objects and predicted tracks that are within a certain gated distance to each other are then matched in the data association block. In order to reduce computation, a single hypothesis strategy was used to find only the optimal match. The states of the matched tracks are then corrected in the state update block based on the corresponding detections. The unmatched tracks and detections are passed to the track management module which decides track creation and deletion based on the times a new track is observed and an old track is missed respectively. Finally, tracked objects are classified using a multi-frame classification module, where multi-frame points are used by storing points for few frames.

B. Multi-frame Classifier using Point Aggregation Network

In the first proposed variant, the information coming from multiple frames is utilized by aggregating the points present inside the bounding box detection of each tracked object for a set of T frames. In this respect, a first-in-first-out (FIFO) queue of points was maintained with a size of T. The index of frames in the queue served as a timestamp of that set of points and the point cloud was appended with the timestamp. If there are no points for a tracked object at a particular frame, the respective index of that queue was given an empty value.

These set of points were then passed through a point cloud based NN inspired from lidar [4]. Fig. 2 shows an overview of the proposed scheme. Initially, the set of aggregated points are passed through feature extraction layers to obtain features. Afterwards, these discriminative features are passed through a set of fully connected layers which act as a cost minimization function to minimize the loss between predicted and ground truth classification values. Here, the key difference from [4] is that since the radar point cloud contains doppler, RCS and timestamp values in addition to location (x, y), the NN not only learns to classify the object based on the spatial distribution of detected points, but also learns to model the distribution of doppler and RCS features and to estimate track dynamics based on the timestamp values, which helps in classification.

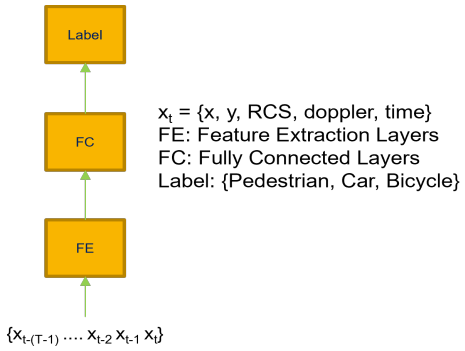


Fig. 2. Multi-Frame Classification using Point Aggregation Network

C. Multi-frame Classifier using BiLSTM Network

In the second proposed variant, instead of aggregating the points coming from multiple frames, a BiLSTM module was used to process the features coming from different timestamps. This is a common method used in NN research and allows to learn the complex relationship between features over different frames. Fig. 3 shows an overview of the proposed method. Initially, the set of points at each frames are passed through shared feature extraction layers to obtain features at each timestamp. These features are then aggregated using BiLSTM layers to learn discriminative features that help in separating the targets into different classes based on information from multiple timestamps. Afterwards, these discriminative features are passed through a set of fully connected layers which act as a cost minimization function to minimize the loss between predicted and ground truth classification values, thereby outputting classification labels that gives the least loss.

D. Training Procedure

In order to obtain the input for training multi-frame classifiers, the points present for each object within the ground truth were extracted. Given the track ID of each object, the points corresponding to the same object in other frames can be found. For the case of network using point aggregation, the points for a set of T frames were aggregated and each point was appended with its timestamp value. For the case of BiLSTM network, points were extracted from each ground truth track for a set of T frames and used as input to train the NN model.

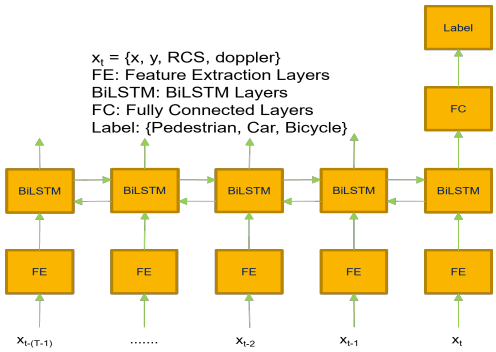


Fig. 3. Multi-Frame Classification using BiLSTM Network

E. Inference Procedure

For inference, points for each object at a particular frame were extracted by finding the points inside the bounding box of that object. The tracks which were matched to a detected object were given the points of that detected object for that particular frame, whereas tracks which were not matched to an object were given empty values for that frame. For the case of network using point aggregation, these points were then aggregated with the points of the track from previous T-1 frames by adding them into the FIFO queue. These set of points were then passed to the model to provide classification of tracked object. For frame number less than T, the classification was provided directly by the classification output from object detection network as in [3]. For the case of BiLSTM network, the points at different frames were passed to the model separately to provide classification labels. For frame number less than T, the classification was provided by classification output from object detection network as in [3].

IV. Experimental Results and Discussion

A. Dataset and Performance Metrics

We used the opensource RadarScenes dataset [6] for our experiments which provides ground truth label of classes and track ids for each moving point. RadarScenes is used because it is a large dataset, provides high quality annotations for individual points and has decent azimuth resolution of 0.5°-2°.

We used MOTA and F1 score from CLEAR MOT metrics [7] to evaluate our methods. This is because they are widely used by autonomous driving benchmarks i.e. NuScenes [14].

B. Results

1) Quantitative

Table 1, 2, 3 show a comparison of the quantitative performance of trackers using multiple frames with a baseline tracker inspired by [3] on car, pedestrian, and cyclist class. It can be observed that there is a consistent improvement in tracking metrics for trackers using multiple frames as compared to tracker using single frame for classification whereby MOTA increased by 4.19%, 0.80%, 22.31%, and F1 by 2.66%, 0.96%, 21.45% for car, pedestrian, and cyclist class respectively. The highest increase is observed for cyclist because a large portion of cyclist consist of wheels which have an irregular motion pattern that can confuse a single frame classifier. Using multiple frames, allow the classifier to capture the motion of the cyclist better, which helps in classification. However, it must be noted that the proposed method only improves classification, whereas other tracking modules such as data association and state estimation remain unchanged.

Table 1, 2, 3 also show that the performance of tracker using point aggregation for classification is slightly lower than the tracker using BiLSTM network except for the pedestrian class. This is because BiLSTM layers can learn to aggregate the features from multiple frames in a forward-backward direction rather than simple aggregation of points. The performance on pedestrian is worse since some pedestrians contain no detected points causing the BiLSTM network to fail to extract the essential features required for classification. On the other hand, the tracker aggregating the points from multiple frames will most likely contain at least few aggregated points from multiple frames, allowing it to extract the spatial and temporal features in a better way that will help in classification.

Table 1. Results on RadarScenes Dataset for Car Class

Tracking Metrics (%)	**Tracker Using Single Frame Classification**	**Tracker Using Multi-Frame Point Aggregation**	**Tracker Using Multi-Frame BiLSTM Classification**
MOTA	68.45	71.29	72.64
F1	87.42	89.43	90.08

Table 2. Results on RadarScenes Dataset for Pedestrian Class

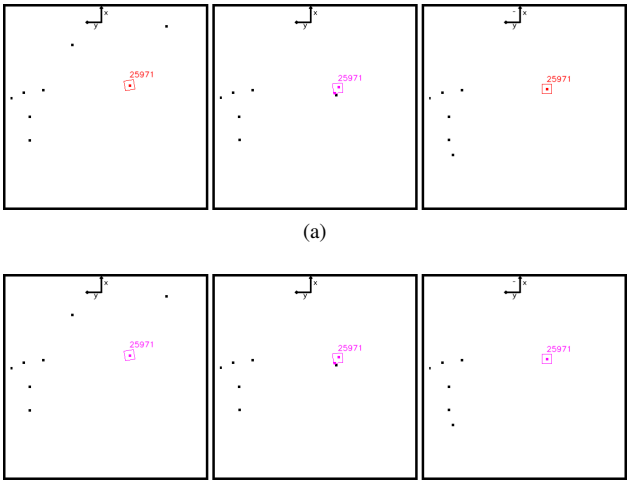
Tracking Metrics (%)	**Tracker Using Single Frame Classification**	**Tracker Using Multi-Frame Point Aggregation**	**Tracker Using Multi-Frame BiLSTM Classification**
MOTA	44.19	45.34	44.99
F1	65.98	67.21	66.94

Table 3. Results on RadarScenes Dataset for Cyclist Class

Tracking Metrics (%)	**Tracker Using Single Frame Classification**	**Tracker Using Multi-Frame Point Aggregation**	**Tracker Using Multi-Frame BiLSTM Classification**
MOTA	34.10	50.55	56.41
F1	57.28	74.33	78.73

2) Qualitative

Fig. 4 provides a visualization of the performance between the baseline tracker that uses single frame and the proposed tracker that uses multiple frames for classification on a sequence of images containing a cyclist moving in front of ego car. Fig. 4a shows that the classification from a single frame fluctuates between a pedestrian and a cyclist. This is because there is only a single detection point for these frames, making it difficult for a single frame classifier to extract useful features required for classification. However, Fig. 4b shows that the multi-frame classifier gives a consistent correct prediction of the cyclist class. This is because usage of multiple frames allows the classifier to capture the motion of the cyclist better. This will help the classifier to recognise the object as cyclist.



(a)

(b)

Fig. 4. Classification output of a tracked cyclist on a test scene using different classification approaches. (a) Single Frame Classification: Cyclist (label in pink) is classified incorrectly as pedestrian (label in red) for frames 1 and 3. (b) Multiple Frame Classification: Cyclist is classified correctly for all frames.

C. Influence of Aggregation Time

An investigation to determine the optimum time duration, required for classification was performed by training the classifiers using different time windows. Fig. 5 shows a plot of the classification accuracy against time. It can be observed that there is an increase in accuracy upon increasing the time duration since more information about the track is available. However, the curve saturates reaching a plateau at time duration = 0.36s (equivalent to 6 frames), showing that a further increase in time duration will not help in classification.

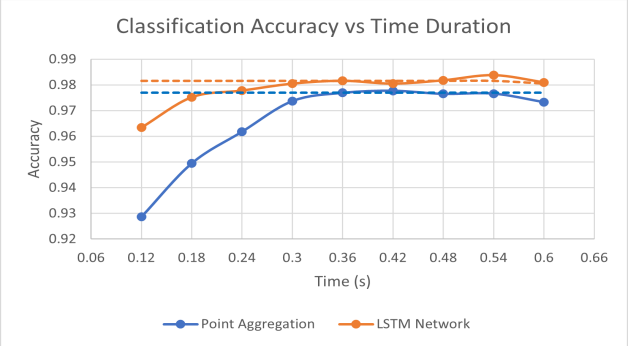


Fig. 5. Relationship between Classification Accuracy and Time Duration

V. Conclusion

This work addressed the problem of classification of tracked objects using multiple frames. Two methods were designed; one used NN based on point aggregation and another utilized BiLSTM layers to process multi-frame information. Both the aforementioned approaches showed a consistent increase in tracking performance compared to the baseline tracker using single frame for classification of tracked objects.

Overall, the proposed approach improves classification of tracked objects, but does not improve other components such as state estimation and data association. So, for future work, this motivates us to use other approaches such as object detectors with velocity estimation that can improve the state estimation performance and use low level radar cube features that can give an improved classification and data association.

Acknowledgment

This work is part of the IPCEI ME/CT and is funded by the European Union Next Generation EU, the German Federal Ministry for Economic Affairs and Climate Action, the Bavarian Ministry of Economic Affairs, Regional Development and Energy, the Free State of Saxony with the help of tax revenue based on the budget approved by the Saxon State parliament and the Free and Hanseatic City of Hamburg.

References

[1] Y. Bar-Shalom, T. Kirubarajan, and C. Gokberk, "Tracking with classification-aided multiframe data association," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 41, no. 3, pp. 868–878, 2005. DOI: 10.1109/TAES.2005.1541436.

[2] X. Weng, J. Wang, D. Held, and K. Kitani, "3D multi-object tracking: A baseline and new evaluation metrics," in *IEEE International Conference on Intelligent Robots and Systems*, 2020. DOI: 10.1109/IROS45743.2020.9341164.

[3] M. Hassan, F. Fioranelli, A. Yarovoy, and S. Ravindran, "Radar multi object tracking using dnn features," in *2023 IEEE International Radar Conference (RADAR)*, 2023, pp. 1–6. DOI: 10.1109/RADAR54928.2023.10371032.

[4] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "PointNet++: Deep hierarchical feature learning on point sets in a metric space," in *Advances in Neural Information Processing Systems*, vol. 2017-December, 2017.

[5] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, 1997, ISSN: 08997667. DOI: 10.1162/neco.1997.9.8.1735.

[6] O. Schumann, M. Hahn, N. Scheiner, *et al.*, "Radarscenes: A real-world radar point cloud data set for automotive applications," in *Proceedings of 2021 IEEE 24th International Conference on Information Fusion*, 2021.

[7] K. Bernardin and R. Stiefelhagen, "Evaluating multiple object tracking performance: The CLEAR MOT metrics," *Eurasip Journal on Image and Video Processing*, vol. 2008, 2008, ISSN: 16875176. DOI: 10.1155/2008/246309.

[8] J. Lan and X. R. Li, "Tracking of maneuvering non-ellipsoidal extended object or target group using random matrix," *IEEE Transactions on Signal Processing*, vol. 62, no. 9, pp. 2450–2463, 2014. DOI: 10.1109/TSP.2014.2309561.

[9] M. Ulrich, C. Glaser, and F. Timm, "DeepReflecs: Deep Learning for Automotive Object Classification with Radar Reflections," in *IEEE National Radar Conference - Proceedings*, vol. 2021-May, 2021. DOI: 10.1109/RadarConf2147009.2021.9455334.

[10] T. Akita and S. Mita, "Object tracking and classification using millimeter-wave radar based on lstm," in *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*, 2019, pp. 1110–1115. DOI: 10.1109/ITSC.2019.8917144.

[11] F. J. Abdu, Y. Zhang, M. Fu, Y. Li, and Z. Deng, "Application of deep learning on millimeter-wave radar signals: A review," *Sensors*, vol. 21, no. 6, 2021, ISSN: 1424-8220. DOI: 10.3390/s21061951. [Online]. Available: https://www.mdpi.com/1424-8220/21/6/1951.

[12] A. Palffy, E. Pool, S. Baratam, J. F. Kooij, and D. M. Gavrila, "Multi-Class Road User Detection with 3+1D Radar in the View-of-Delft Dataset," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, 2022, ISSN: 23773766. DOI: 10.1109/LRA.2022.3147324.

[13] R. E. Kalman, "A new approach to linear filtering and prediction problems," *Journal of Fluids Engineering, Transactions of the ASME*, vol. 82, no. 1, 1960, ISSN: 1528901X. DOI: 10.1115/1.3662552.

[14] H. Caesar, V. Bankiti, A. H. Lang, *et al.*, "Nuscenes: A multimodal dataset for autonomous driving," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2020. DOI: 10.1109/CVPR42600.2020.01164.