UDC 519.8

Götz Grimmer[1], Clemens Klöck [2], Steffen Schober [3]

MCW-RADAR ROAD USER DETECTION, BASED ON EDGE DEVICES DATA PROCESSING WITH QUANTIZED NEURAL NETWORK

Introduction. Traffic monitoring in urban and highway environments is crucial for road safety, congestion control, and smart city applications. Traditional monitoring approaches, such as camera-based surveillance or inductive loops, face limitations in terms of privacy concerns, environmental dependencies, and computational requirements [1]. Cameras are highly susceptible to lighting conditions, occlusions, and privacy restrictions, while inductive loops are costly to install and maintain. To address these challenges, radar-based sensing has gained significant attention due to its capability to detect moving objects in various weather and lighting conditions without compromising privacy. Frequency-Modulated Continuous Wave (FMCW) radar technology, in particular, provides high-resolution range-doppler data that enables the classification of different road users. However, processing radar data in real time requires efficient computing resources, making edge-based processing a viable solution. This paper introduces an embedded edge device utilizing a dual 77GHz FMCW radar system combined with a quantized neural network for real-time road user classification. The proposed system operates independently, reducing the need for centralized computing infrastructure while providing accurate and efficient detection of pedestrians, cyclists, and vehicles [5]. The key contributions of this work include:

[1] M.Sc., Researcher, Department of Computer Science, Esslingen University of Applied Sciences, Germany.

[2] Dr., Professor, Department of Computer Science, Esslingen University of Applied Sciences, Germany.

[3] Dr., Professor, Department of Computer Science, Esslingen University of Applied Sciences, Germany.

- Development of a low-cost, embedded radar-based detection system using edge computing.
- Implementation of a quantized Time Convolutional Network (TCN) for sequential radar data processing.
- Optimization of the neural network through Quantization-Aware Training (QAT) and Post-Training Quantization (PQT).
- Performance evaluation demonstrating real-time classification accuracy and efficient computational utilization on a Raspberry Pi 4B.

System Architecture. The proposed system consists of two synchronized Texas Instruments AWR.1642BOOST radar modules, operating at 77GHz with a 4GHz bandwidth. These radars are configured in a MIMO (Multiple Input Multiple Output) setup, enhancing detection accuracy and robustness. The radar sensors are connected to a Raspberry Pi 4B via a CAN-FD interface [2]. Each radar sensor captures range-doppler matrices with 256 range bins and 64 velocity bins, allowing the system to track moving objects within a 20m range. The raw radar frames are preprocessed before being passed to the neural network for classification. The processing pipeline includes noise filtering, range-doppler transformation, and normalization to improve classification performance.

Machine Learning Model. To classify detected objects into pedestrians, cyclists, vehicles, or empty road segments, a Time Convolutional Network (TCN) is employed. Unlike traditional Recurrent Neural Networks (RNNs), which store temporal dependencies within hidden states, TCNs utilize causal convolutions over multiple time steps, making them more efficient for sequential data processing.

- The neural network consists of multiple convolutional layers, with residual connections improving gradient propagation. The network structure includes:
- Feature Extraction Layer: Initial convolutional layers to extract key features from the radar range-Doppler matrices.
- Temporal Convolutional Layers: Stacked convolutional layers with dilated filters to capture sequential dependencies.
- Fully Connected Layer: Final dense layer for classification into predefined categories.

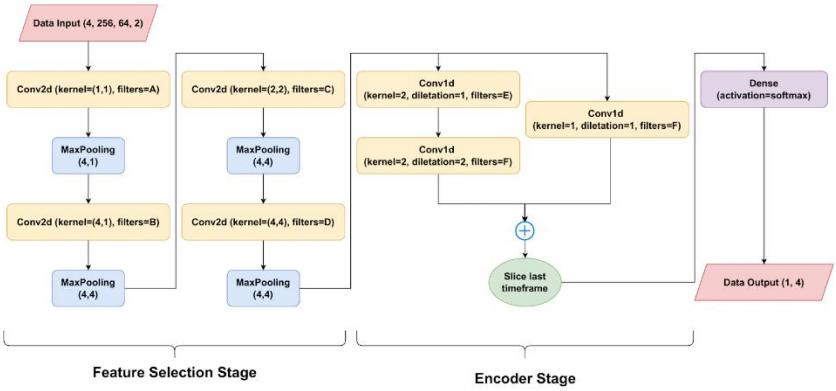


Figure 1: Model structure of quantized sequence model to recognized patterns inside a stream of range-doppler matrices.

In Figure 1 you see the structure of the model we applied. A preprocessing stage has to extract the features of range-doppler matrices and convert it into a 2D tensor. That is accomplished through a small YOLO network, which extracts the features and acts as an „embedding" stage which processes independent of the time dependency. The stage is called Feature Selection Stage. Its only task is to extract features of the range-doppler matrix, later the time reference will be depicted only through the TCN, called Encoder Stage [4]. To optimize computational efficiency, a quantization process is applied to the model, reducing the required memory footprint and inference time. Quantization-Aware Training (QAT) ensures that the neural network maintains accuracy while converting floating-point operations to lower-bit precision. Post-Training Quantization (PQT) is further applied to reduce model size to 78kB without significantly affecting accuracy [5].

Experimental Evaluation. The system was evaluated using a dataset comprising 2,106 radar measurements. The dataset includes various traffic scenarios, such as:

- Single pedestrian, cyclist, or vehicle moving within the radar range.
- Pedestrian and cyclist moving together, with the cyclist overtaking the pedestrian.
- Pedestrian and vehicle appearing in the same scene, with the vehicle overtaking the pedestrian.

Each radar frame is labeled and used for training, validation, and testing. The dataset is divided into 70% training, 20% validation, and 10% testing subsets. The quantized model achieves an overall classification Area Under the Curve (AUC) of 0.95.

Performance of predictions are shown in Figure 2, the confusion matrix is evaluated on the test dataset.

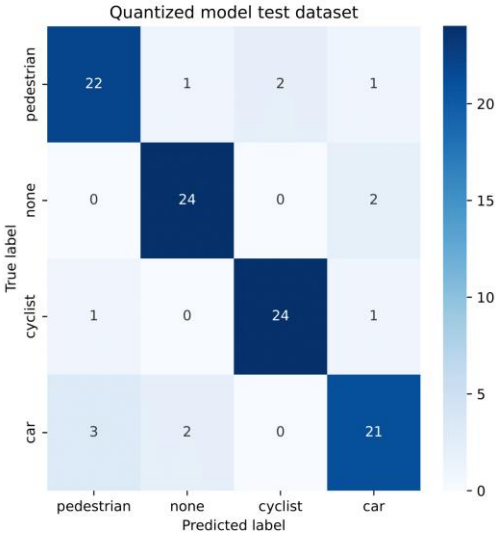


Figure 2: Confusion matrix on test dataset with balanced dataset.

However, a slight misclassification occurs when pedestrians and vehicles are in close proximity, likely due to overlapping radar signatures. The embedded system operates efficiently on the Raspberry Pi 4B, utilizing only 25% of the available computational resources per inference cycle. This low computational load enables real-time operation, with radar frames processed at a rate of 4 frames per second (FPS). Additional optimizations, such as reducing the radar frame size or implementing hardware acceleration, could further improve performance.

Conclusions. This study presents an edge-based radar detection system capable of real-time road user classification using a quantized neural network. By integrating a dual FMCW radar setup with an optimized machine learning pipeline, the system achieves high classification accuracy while maintaining a low computational footprint. The quantized neural network significantly reduces memory usage and inference time, making it suitable for low-cost embedded platforms. The proposed system offers several advantages over traditional traffic monitoring methods, including:

- **Privacy Preservation:** Unlike camera-based systems, radar-based detection does not capture personally identifiable images.

- **Environmental Robustness:** The system functions reliably in various lighting and weather conditions.
- **Reduced Computational Load:** On-device processing minimizes reliance on cloud-based computing, reducing latency and bandwidth requirements.

Future work will focus on expanding the dataset to improve classification accuracy, integrating additional sensor modalities (e.g., LiDAR, infrared cameras), and developing a cooperative edge-based sensor network for enhanced traffic monitoring. Further optimization of the neural network architecture will also be explored to improve real-time performance.

REFERENCES

1. Nimac P, Krpič A, Batagelj B, Gams A. Pedestrian traffic light control with crosswalk FMCW radar and group tracking algorithm. Sensors. 2022;22(5):1754.
2. Macenski S, Foote T, Gerkey B, Lalancette C, Woodall W. Robot operating system 2: Design, architecture, and uses in the wild. Science Robotics. 2022;7(66):eabm6074.
3. Gholami A, Kim S, Dong Z, Yao Z, Mahoney MW, Keutzer K. A survey of quantization methods for efficient neural network inference. Low-power computer vision. Chapman and Hall/CRC; 2022. p. 291–326.
4. Scherer M, Magno M, Erb J, Mayer P, Eggimann M, Benini L. TinyRadarNN: Combining spatial and temporal convolutional neural networks for embedded gesture recognition with short range radars. IEEE Internet of Things Journal. 2021;8(13):10336–46.
5. Klein LA, Mills MK, Gibson DR, others. Traffic detector handbook: Volume I. Turner-Fairbank Highway Research Center; 2006.