Radar-Lidar Fusion for Classification of Traffic Signaling Motion in Automotive Applications

Sabyasachi Biswas, *Student Member, IEEE,*
John E. Ball *Senior Member, IEEE* and Ali C. Gurbuz, *Senior Member, IEEE,*

2023 IEEE International Radar Conference (RADAR) | 978-1-6654-8278-3/23/$31.00 ©2023 IEEE | DOI: 10.1109/RADAR54928.2023.10371020

***Abstract*—Advanced driver-assisted system (ADAS) uses multiple sensors such as Radar, Lidar, or Cameras in vehicles to create a robust perception against challenging weather conditions and individual sensor failures. In typical conditions, Lidar and Camera can perceive the surrounding much better than the radar whereas, under low light or extreme weather conditions (fog, rain, snow) the radar outperforms both as it works independently of light source. These sensors in the ADAS system help to minimize driving errors by providing necessary information to the driver or taking automatic actions based on what it perceives. However, in some unstructured environments, which do not have any operational traffic lights present, a person via appropriate gesturing directs the traffic. The task of autonomous vehicles recognizing human body language and gestures in traffic-directing scenarios is significantly difficult. To overcome this challenge, based on the US traffic system, we present a new dataset collected of traffic signaling motions using millimeter-wave (mmWave) radar, camera, Lidar, and motion-capture system. Initial classification results from Radar microDoppler (μ-D) signature and Lidar data analysis using Multimodal Neural Network demonstrate that sensor fusion can not only very accurately (around 98%) classify traffic signaling motions in automotive applications but also outperforms the radar-based and lidar-based classification by around 7% and 4% respectively.**

***Index Terms*—Multimodal Neural Network, autonomy, traffic gesture classification, mmWave, ADAS, CNN.**

I. Introduction

OVER the last decade, due to the growing number of advancements in ADAS, modern vehicles are able to analyze various situations in their surrounding environments in more depth and detail [1], [2]. The use of multiple sensors such as lidar, radar, and camera in ADAS helps to create a robust perception of the surroundings and includes many active safety features. These features include adaptive cruise control (ACC) [3], lane departure warning (LDW) [4], blind spot detection (BSD) [5], forward collision warning (FCW) [6], automatic emergency braking (AEB) [7], pedestrian detection [8] etc. These features dramatically increase the effectiveness of ADAS to save lives. Furthermore, ADAS vehicles are being trained with machine learning techniques, which allow them to enhance their effectiveness over time by being exposed to new circumstances and data [9]. As these vehicles become more common, they will have access to an increasing quantity of data on real-world driving circumstances, which will help them become more competent at functioning in a variety of settings. However, ADAS vehicles are still in the early stages of development and are trained to operate in pristine road conditions. In our everyday lives, we see several scenarios where traffic relies on human guidance to navigate. For example, if a vehicle was going through a construction zone, entering and exiting a school or other high-traffic locations, or when automatic traffic signals were not functioning. Under all of these conditions, it is far more typical for someone to be entrusted with directing cars, either by signaling with an appropriate sign or by gesturing. This individual might be a traffic cop, a school official, or a construction worker. If the AV passes through any of these zones, it will need to rely on human instructions to navigate, and it should be able to identify and categorize human traffic directions autonomously. Moreover, in order to create a robust perception of the surrounding sensors such as radar, lidar, camera are being utilized as some of the main sensing systems. Each of these sensors has different strengths and weaknesses. For example, cameras can provide high-resolution images and color information, whereas lidar and radar are useful for obstacle avoidance, adaptive cruise control, etc [10]. by providing object detection, speed, and distance information. However, both the camera and lidar are affected by poor lighting conditions. On the other hand, the radar performs consistently as it works independently of the light source. So, by combining the data from these sensors, ADAS can create a more comprehensive and accurate view of the vehicle's surroundings.

This paper proposes a multimodal neural network architecture to conduct the initial studies toward understanding the success of varying sensors including mm-wave automotive radar and 3D Lidar systems in order to recognize gestures from human traffic directors. We created a novel dataset consisting of 12 different motions based on the most commonly utilized to control traffic in the US traffic system (See Section II-B and Figure 3). First, in order to create a robust dataset, data were collected from 14 participants in the lab environment using RGB-Depth cameras, Lidars, mm-wave radars, and a motion-capture system. In this dataset, only single human directors using pre-defined classes of directions within the line of sight of the sensor suite are considered.

For understanding human traffic directions, we first show the effectiveness of mm-wave radar and Lidar separately. We transformed time-frequency radar data to micro-doppler μ-D signature images and converted 3D Lidar point cloud data into image frames. Afterward, both the processed radar and lidar data were classified using the developed unimodal CNN-2D and CNN-3D architectures respectively. Finally, we implemented two multimodal fusion strategies, data-level, and feature-level, to observe the classification performance. Using data-level fusion we observed performance improvement

978-1-6654-8278-3/23/$31.00 ©2023 IEEE

around 6% and 3% over the radar-based and lidar-based classification respectively. Finally, using feature-level fusion, we further increased the classification accuracy by 1%.

The paper is organized as follows: the experimental setup and dataset are provided in Section II, data processing steps for radar and lidar are discussed in Section III, then the unimodal and multimodal architectures were briefly described in Section IV followed by the performance comparison discussion in Section V. Finally, Section VI concludes the paper and discusses future works.

II. Experimental Setup and Dataset

A. Experimental Setup

The data collection for this study utilizes six different sensors for kinematic movement and visual data to achieve the goal of motion analysis and classification. Of the six sensors used, four are intended for automotive or short-range automotive applications. These sensors include TI AWR2243 mmWave radar, a TI AWR1642 mmWave radar, an Intel Realsense L515 LiDAR, and an Intel Realsense D435 camera. The remaining two sensors are Ouster OS-1 360° scanning LiDAR and MotionMonitor motion capture system with Vicon motion capture cameras respectively. These six sensors are managed across three distinct data collection settings. The AWR2243 sensor collects raw voltage using TI's mmWave Studio software; the AWR1642, L515, D435, and OS-1 sensors collect radar scan, RGB, and point cloud information using the Robot Operation System (ROS); and the MotionMonitor device captures three-dimensional positioning data for important centroids on the participants' upper bodies, as well as rotation and flexion of joints and the head. Figure 1 depicts the setup for the automobile sensors as well as the layout of the laboratory used for data collection, while Figure 2 represents the setup for the four short-ranged sensors utilized in the experiment.

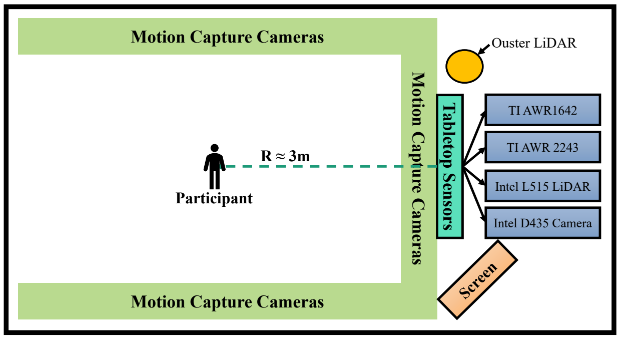


Fig. 1: Experimental Setup

All the sensors except the motion capture system were positioned on top of a table with an elevation of 1 meter prior to data collection. Participants were placed 3 meters in front of the radar. A computer monitor was positioned to the left of the radar, but out of the radar's field of vision (FOV). The display provided prompts indicating which gestures needed to be articulated. The data was gathered for 155 seconds. During this period, four distinct motions were done with five repetitions each. There was a one-second pause between each iteration. After each gesture, a 10-second interval was allowed to preview the following motion. Each sample lasted around 5 seconds.

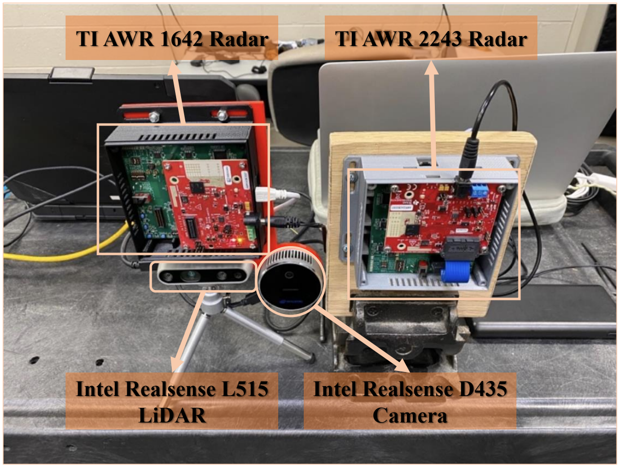


Fig. 2: Table Setup

B. Dataset

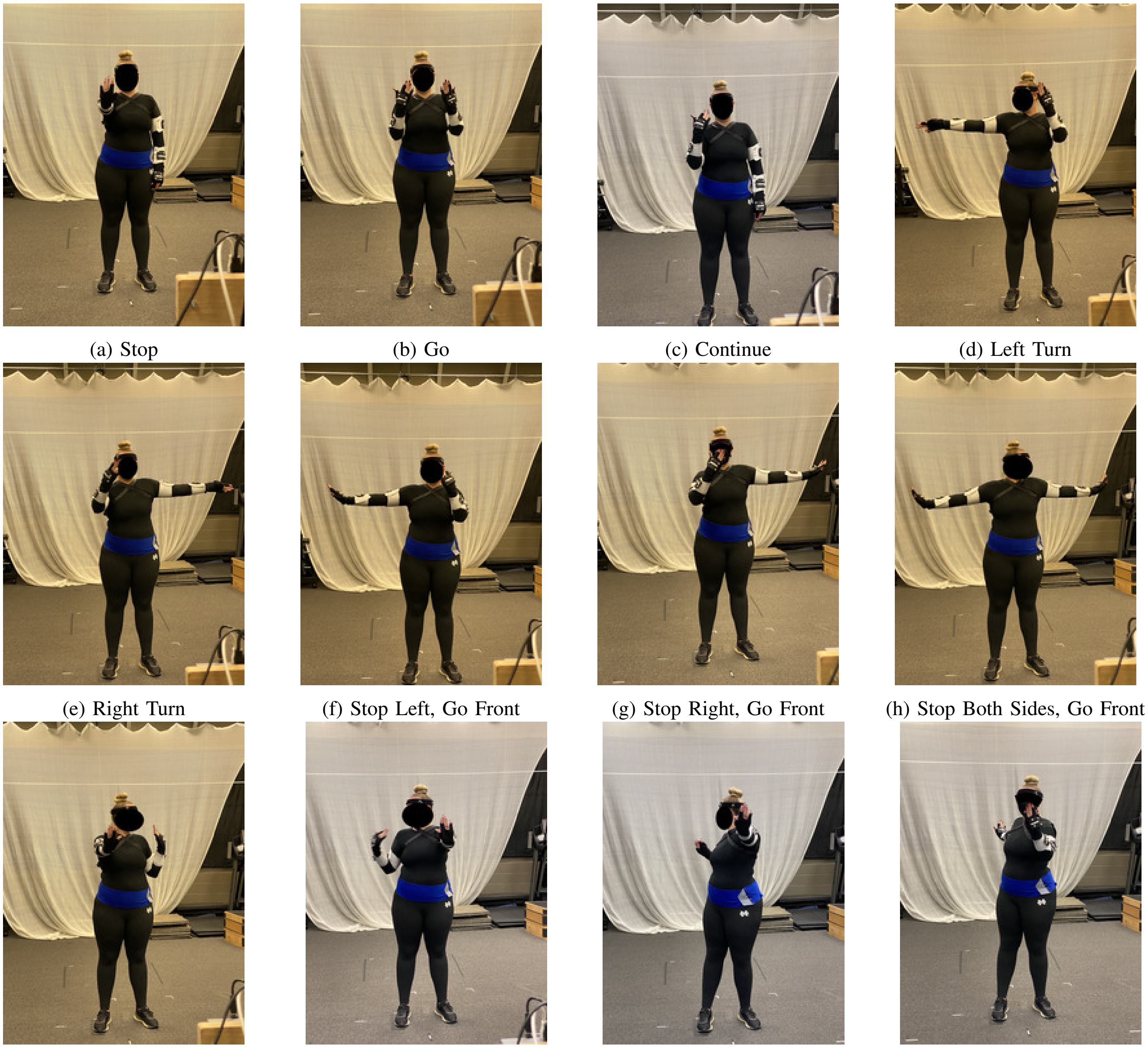
A dataset of 12 traffic signaling movements based on the US traffic system is developed for this study. These gestures are intended to direct an incoming vehicle to stop, move from the halting point in one of three directions, or have vehicles in any given place wait for additional traffic to enter the road. The gestures entail movement of not just the participants' arms but also their hands, as well as head rotation to stare in specified directions. Figure 3 depicts each gesture made by a participant, as well as a brief description of each motion. Previously, this dataset were used to measure the effectiveness of automotive radar for traffic signaling motion classification [11].

Finally, 840 samples were gathered for 12 different gestures, for a total of 70 samples per gesture from 14 different subjects. For now, we are only considering classification using TI AWR2243 automotive FMCW radar and Ouster OS1-32 lidar. The next sections will briefly discuss the data processing methods for both of these sensors.

III. Data Processing

A. AWR2243 Radar Data Processing

The Texas instrument's (TI) AWR2243 radar is a frequency modulated continuous wave (FMCW) radar system that operates between 77 GHz to 81 GHz. This transmits chip signals in the direction of radar field of view [12]. At first, the transmitted signal is reflected from the target, in our case humans. Then the signal received by the radar is a frequency-shifted, time-delayed version of the transmitted signal. The



(a) Stop (b) Go (c) Continue (d) Left Turn

(e) Right Turn (f) Stop Left, Go Front (g) Stop Right, Go Front (h) Stop Both Sides, Go Front

(i) Stop Front, Go Right (j) Stop Front, Go Left (k) Stop Front, Go Back (l) Stop Back, Go Front

Fig. 3: 12 traffic signalling motions to control traffic in the US traffic system

kinematic features of each human target gesture result in a time-varying sequence of Micro-motions [13] e.g. vibrations and rotations. Each gesture produces its own distinct patterns, which can be evaluated through time-frequency analysis. The μ-*D* spectrogram is then calculated from the square modulus of the Short-Time Fourier Transform (STFT) of the continuous-time input signal $x[n]$ and may be described in terms of the window function, $h[n]$.

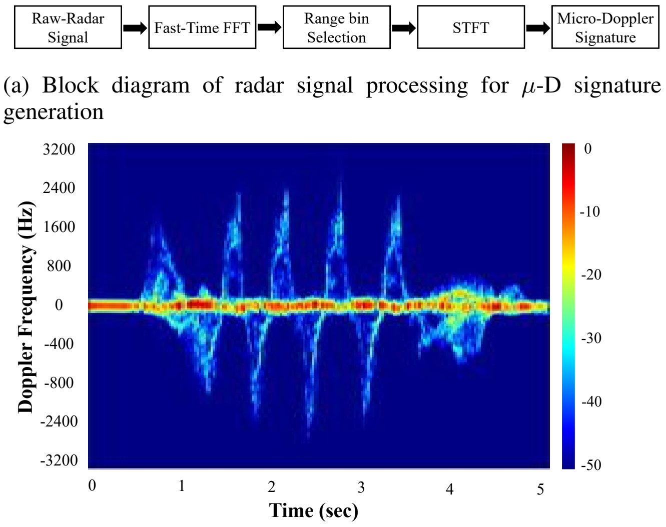
$$\mathrm{STFT}[x[n]]_{m,\omega} = X[m,\omega] = \sum_{n=-\infty}^{\infty} x[n]h[n-m]e^{-j\omega n} \quad (1)$$

$$\mathrm{Spectrogram}[x[n]]_{m,\omega} = |X[m,\omega]|^2 \quad (2)$$

Figure 4a visualizes the procedure of generating a μ-D spectrogram from raw radar data and Figure 4b provides a μ-D signature example of traffic sign 'Continue' using an image of scaled colors. At 0 Hz, positive Doppler frequencies are depicted above the horizontal axis, whereas negative Doppler frequencies are represented below the horizontal axis.

B. Lidar Data Processing

For lidar point clouds, spatial resolution changes with distance. Using the change in distance, distance normalization on a statistical outlier filter is used to remove the outliers [14]. Then the point cloud dataset was transformed into a voxel grid [15]. Finally, this voxelized data was projected on a 2D



(a) Block diagram of radar signal processing for μ-D signature generation

(b) A microDoppler Spectrogram example of the traffic sign 'Continue'

Fig. 4: Block diagram and example of μ-D signature generation

plane [16]. This is considered as a single frame. A total of 60 frames were captured for a single data sample during 5s of data collection creating a lidar video.

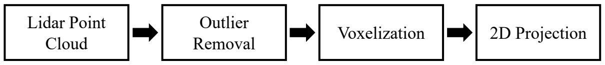


Fig. 5: Block diagram of lidar data processing

IV. The CNN Architectures

Since, we are using μ-D spectrogram images and voxelized lidar videos for classification, two unimodal 2D and 3D CNN models were developed respectively. Also, two multimodal neural network architectures were developed for data-level and feature-level fusions. For classification, the dataset was split into 80% training and 20% testing.

A. Unimodal Neural Network Architectures

The CNN-2D architecture for radar μ-D spectrogram classification has four convolution layers, each with kernel size 3×3, stride 1×1, and comprising of 12, 12, 16, and 16 filters respectively. A 2×2 max-pooling, batch normalization, and activation ReLU are performed after each convolutional layer. Then, the tensor is then flattened, fed into a dense layer with a size of 64, dropped out by 0.2, and activated with ReLU. Finally, the model ends with the softmax classifier. 64×64 spectrogram images were used as the input to the model.

The CNN-3D architecture for lidar video shares the same architectural behavior as CNN-2D. 64×64×60 lidar videos were used as the input to the model.

B. Multimodal Neural Network Architectures

1) Data-level fusion: For the data level fusion, the spectrogram images was added as an additional channel to the CNN-3D input. So, the combined input dimensions for the CNN input then becomes 64×64×61. The CNN-3D model for data-level fusion shares the same architecture as the CNN-3D model used for lidar video classification.

2) Feature-level fusion: The feature-level fusion architecture was obtained by concatenating the features collected from the CNN-2D architecture for radar spectrogram and CNN-3D architecture for lidar video. This was followed by a dense layer with a size of 64, a relu activation layer, and a dropout layer of 0.2. Finally, the model ends with a softmax classification.

The developed unimodal and multimodal CNN architectures are shown in figure 6.

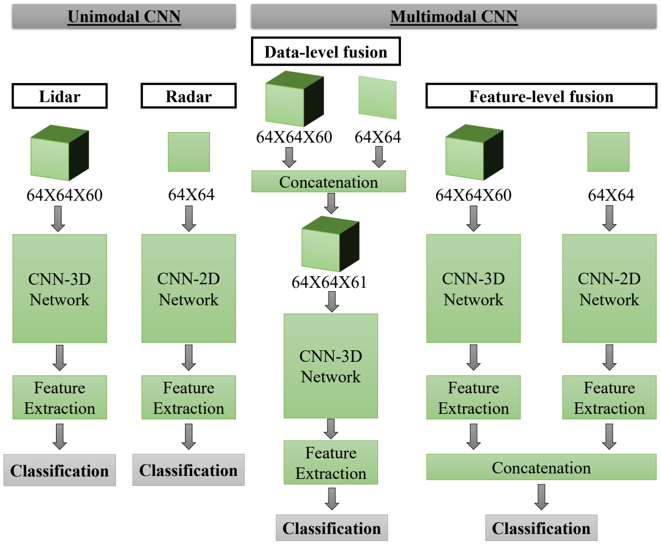


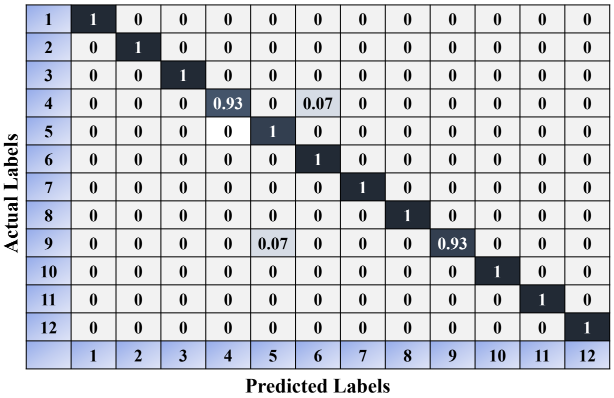
Fig. 6: Block Diagrams for CNN architectures

V. Performance Analysis

For performance evaluation, the dataset was split into 80% training and 20% training. The spectrogram images and lidar videos were saved as 64×64 and 64×64×60 respectively. The comparison results between unimodal and multimodal CNN architectures are shown in Table I. From the table, it can be seen that both multimodal CNNs are performing better than unimodal ones. The feature level fusion architecture is showing the best testing accuracy of 98.81% which is 7.14% and 4.17% higher than the CNN 2D and 3D models used for radar-based and lidar-based classifications respectively. It also outperforms the data-level fusion architecture by 1%. Though the dataset was very challenging as performed gestures varied from person to person, the confusion matrix in Figure 7 shows a very promising performance for all the classes for feature-level fusion classification. While the radar-based classification showed some inconsistencies in classifying some of the classes, fusion-based classifications were pretty consistent throughout. This shows a lot of potentials as it might be an important feature to be included in the ADAS system.

TABLE I: Performance Comparison

Network	Testing Accuracy	Precision	Recall	F1 Score
CNN-2D (Radar)	91.67	91.68	91.76	91.60
CNN-3D (Lidar)	94.64	95.12	94.64	94.65
CNN-3D (Data-Fusion)	97.62	97.73	97.62	97.59
CNN-2D+3D (Feature Fusion)	98.81	98.89	98.81	98.81



1	1	0	0	0	0	0	0	0	0	0	0	0
2	0	1	0	0	0	0	0	0	0	0	0	0
3	0	0	1	0	0	0	0	0	0	0	0	0
4	0	0	0	0.93	0	0.07	0	0	0	0	0	0
5	0	0	0	0	1	0	0	0	0	0	0	0
6	0	0	0	0	0	1	0	0	0	0	0	0
7	0	0	0	0	0	0	1	0	0	0	0	0
8	0	0	0	0	0	0	0	1	0	0	0	0
9	0	0	0	0	0.07	0	0	0	0.93	0	0	0
10	0	0	0	0	0	0	0	0	0	1	0	0
11	0	0	0	0	0	0	0	0	0	0	1	0
12	0	0	0	0	0	0	0	0	0	0	0	1
	1	2	3	4	5	6	7	8	9	10	11	12

Fig. 7: Confusion matrix for feature level fusion classification

VI. Conclusion and Future Work

The purpose of this study is to give an initial study towards understanding the success of sensor fusion in gesture recognition obtained from human traffic directors. The results show that in terms of decision-making for ADAS vehicles, the sensor fusion techniques are much more efficient than sensors working independently. The feature-level fusion technique had a testing accuracy of almost 99% even with a limited number of data samples. Our initial study was performed in the lab environment and with specific guidance. Also, multi-stage processing was required to utilize CNN architectures for classification. Moreover, data collected from the depth camera and Motion capture are yet to be explored. In future works, we will try to develop models that can work on raw radar, lidar, and camera data directly for real-time classification. Real-time fusion techniques between these sensors should open more doors to exploration of human body motions used in automotive applications.

VII. Acknowledgements

The data collection work was funded in part by the National Science Foundation (NSF) Awards #1931861, and #2047771. Human studies research was conducted under Mississippi State University Institutional Review Board (IRB) Protocol # IRB-21-256.

References

[1] M. Hasenjäger and H. Wersing, "Personalization in advanced driver assistance systems and autonomous vehicles: A review," in *2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)*, 2017, pp. 1–7.

[2] J. Nidamanuri, C. Nibhanupudi, R. Assfalg, and H. Venkataraman, "A progressive review: Emerging technologies for adas driven solutions," *IEEE Transactions on Intelligent Vehicles*, vol. 7, no. 2, pp. 326–341, 2022.

[3] M. M. Brugnolli, B. A. Angélico, and A. A. M. Laganá, "Predictive adaptive cruise control using a customized ecu," *IEEE Access*, vol. 7, pp. 55 305–55 317, 2019.

[4] Y. A. Ahmed, A. T. Mohamed, and A. M. B. Aly, "Robust lane departure warning system for adas on highways," in *2022 4th Novel Intelligent and Leading Emerging Sciences Conference (NILES)*, 2022, pp. 321–324.

[5] S.-M. Chang, C.-C. Tsai, and J.-I. Guo, "A blind spot detection warning system based on gabor filtering and optical flow for e-mirror applications," in *2018 IEEE International Symposium on Circuits and Systems (ISCAS)*, 2018, pp. 1–5.

[6] S. Kumar, V. Shaw, J. Maitra, and R. Karmakar, "Fcw: A forward collision warning system using convolutional neural network," in *2020 International Conference on Electrical and Electronics Engineering (ICE3)*, 2020, pp. 1–5.

[7] A. Dixit, P. D. Devangbhai, and C. R. Kumar, "Modelling and testing of emergency braking in autonomous vehicles," in *2021 Innovations in Power and Advanced Computing Technologies (i-PACT)*, 2021, pp. 1–6.

[8] R. Ayachi, M. Afif, Y. Said, and A. B. Abdelaali, "pedestrian detection for advanced driving assisting system: a transfer learning approach," in *2020 5th International Conference on Advanced Technologies for Signal and Image Processing (ATSIP)*, 2020, pp. 1–5.

[9] A. Moujahid, M. ElAraki Tantaoui, M. D. Hina, A. Soukane, A. Ortalda, A. ElKhadimi, and A. Ramdane-Cherif, "Machine learning techniques in adas: A review," in *2018 International Conference on Advances in Computing and Communication Engineering (ICACCE)*, 2018, pp. 235–242.

[10] M. E. Warren, "Automotive lidar technology," in *2019 Symposium on VLSI Circuits*, 2019, pp. C254–C255.

[11] S. Biswas, B. Bartlett, J. E. Ball, and A. C. Gurbuz, "Classification of traffic signaling motion in automotive applications using fmcw radar," in *2023 Radar Conference, San Antonio, Texas, USA*, 2023.

[12] L. Piotrowsky, T. Jaeschke, S. Kueppers, J. Siska, and N. Pohl, "Enabling high accuracy distance measurements with fmcw radar sensors," *IEEE Transactions on Microwave Theory and Techniques*, vol. 67, no. 12, pp. 5360–5371, 2019.

[13] V. C. Chen, D. Tahmoush, and W. J. Miceli, *Radar micro-Doppler signatures*. Institution of Engineering and Technology, 2014.

[14] X. Wang, P. Ma, L. Jiang, L. Li, and K. Xu, "A new method of 3d point cloud data processing in low-speed self-driving car," in *2019 IEEE 3rd Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC)*, 2019, pp. 69–73.

[15] C. Kunert, T. Schwandt, and W. Broll, "Efficient point cloud rasterization for real time volumetric integration in mixed reality applications," in *2018 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, 2018, pp. 1–9.

[16] S. Chamorro, J. Collier, and F. Grondin, "Neural network based lidar gesture recognition for realtime robot teleoperation," in *2021 IEEE International Symposium on Safety, Security, and Rescue Robotics (SSRR)*, 2021, pp. 98–103.