Microwave/millimeter wave sensors

Robust Doppler-Based Gesture Recognition With Incoherent Automotive Radar Sensor Networks

Nicolai Kern[1*], Maximilian Steiner[1*], Ramona Lorenzin[1], and Christian Waldschmidt[1**]

[1] *Institute of Microwave Engineering, Ulm University, Ulm 89081, Germany*
[*] *Student Member, IEEE*
[**] *Senior Member, IEEE*

Manuscript received September 8, 2020; revised September 23, 2020; accepted October 15, 2020. Date of publication October 26, 2020; date of current version November 13, 2020.

Abstract—In this letter, the capabilities of an incoherent radar sensor network for robust Doppler-based gesture recognition are investigated, and a significant performance boost is demonstrated. A comprehensive dataset is recorded with an incoherent sensor network consisting of three time-synchronized 77GHz frequency-modulated continuous wave radars. Based on this dataset, we show that differential Doppler features obtained from the varying viewing angles result in a significant multistatic gain for classification, particularly for high intraclass variations and low Doppler frequencies. For the most complex dataset, cross-user validation accuracy of a convolutional neural network with optimized data fusion is improved by 7.4% to an overall value of 87.1%, which we regard to be high as gestures are not designed for distinguishability but reflect everyday control and communication signals.

Index Terms—Microwave/millimeter wave sensors, autonomous driving, gesture recognition, multistatic radar, radar sensor network.

I. INTRODUCTION

In recent years, millimeter-wave radar has become a crucial part of advanced driver-assistance system as a key enabler for autonomous operation, e.g., on motorways [1]. However, with the striving for fully autonomous driving, radar is required to provide a more compound environmental understanding beyond detection and ranging [2]. For example, it might provide robust intention recognition of urban traffic participants such as pedestrians also under adverse weather conditions. One key for intention recognition is the utilization of nonverbal communication toward the vehicle, such as hand and arm gestures.

The field of sensor-based gesture recognition has gained ground in recent years, mainly fueled by the rapid progress in deep learning. Radar has been successfully applied to this task, with the focus mainly on fine-grained gestures for human–machine interaction performed close to the sensor. Since high velocity resolution is the outstanding feature of typical radar sensors, the large majority of state-of-the-art algorithms exploits the gestures' velocity components which are perceived as time-varying Doppler frequencies by the radar. The Doppler information is utilized either applying convolutional neural networks (CNNs) to Doppler spectrograms [3], [4] or exploiting temporal evolutions in range-Doppler maps [5]–[7].

However, when moving radar-based gesture recognition to more complex scenarios such as nonverbal communication in urban traffic environments, algorithms operating on single-sensor Doppler information can run into problems. Communication gestures are not designed for distinguishability and are performed under different observation angles, resulting in some executions with low radial velocity components and, therefore, little Doppler information. Moreover, as gestures are performed under arbitrary orientations of pedestrians, radial velocity varies strongly, and large intraclass variations evolve.

In order to mitigate these issues, we propose the use of an incoherent radar sensor network consisting of three frequency-modulated continuous wave (FMCW) sensors placed on a horizontal line. By recording gestures under slightly different observation angles, the problem of unfavorable orientations is alleviated. Moreover, the differences in radial velocities perceived by the different sensors provide additional valuable information not accessible to a single sensor. Despite the expectable benefits, there is little literature on the use of sensor networks. Yang and Li [8] use two sensors to distinguish between hand gestures based on hand-crafted features, and [9] deploys a single sensor whose receive antennas are separated apart enough to form a multistatic arrangement. However, there is no investigation of the potential of modern radar sensor networks for classification accuracy and robustness. For that reason, this letter first motivates the sensor network approach by theoretical consideration of the multistatic arrangement, before validating the potential by training an optimized classifier on different subsets derived from a comprehensive multistatic dataset with high intraclass variations. Note that we refer to our system as multistatic since we exploit multistatic angles between different sensors, despite deviating from "classical" multistatic systems by having individual transmit-receive pairs colocated.

II. INCOHERENT RADAR SENSOR NETWORK

A. Impact of Multistatic Geometry

Considering a setup with two sensors spanning a multistatic angle α_{MS}, the sensors will observe different radial velocities from a directed motion with velocity $\vec{v}$. This is illustrated in Fig. 1 for the simple case of a target moving in the sensors' azimuthal plane. For a target with radial and tangential velocity components v_{t}^0 and v_{r}^0 relative to node N_0, the velocity vector in N_0's coordinate system $\vec{v}^0 = (v_{\mathrm{r}}^0, v_{\mathrm{t}}^0)^T$ is transformed into its counterpart in N_1's coordinate system $\vec{v}^1$,

Corresponding author: Nicolai Kern (e-mail: nicolai.kern@uni-ulm.de).
Associate Editor: J. M. Corres.
Digital Object Identifier 10.1109/LSENS.2020.3033586

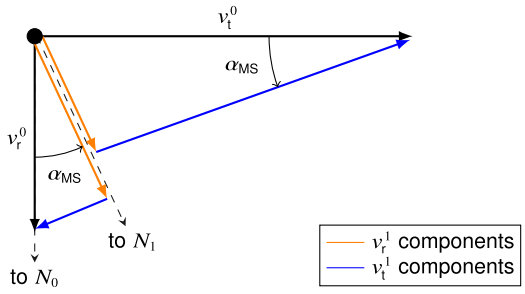

Fig. 1. Impact of the multistatic angle α_{MS} on the transformation of velocity components across the sensors.

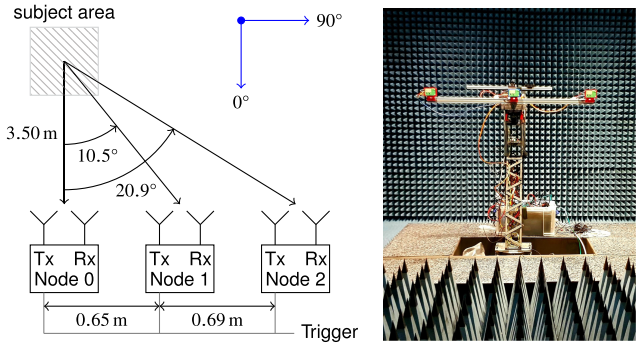

Fig. 2. (a) Sensor network geometry for the multistatic gesture measurement. The blue arrows indicate the orientations of the subjects during the measurements. (b) Photograph of the sensor network.

according to

$$\begin{pmatrix} v_r^1 \\ v_t^1 \end{pmatrix} = \begin{pmatrix} \cos\alpha_{MS} & \sin\alpha_{MS} \\ -\sin\alpha_{MS} & \cos\alpha_{MS} \end{pmatrix} \begin{pmatrix} v_r^0 \\ v_t^0 \end{pmatrix}. \quad (1)$$

Here, v_r^1 and v_t^1 are the radial and tangential velocity components with respect to the second sensor N_1. It can be seen that the Doppler response observed by N_1 is the sum of scaled versions of the radial and tangential motion component w.r.t. N_0, with the scaling obviously depending on the multistatic angle α_{MS}. Thus, observation under different viewing angles can enable a neural network to exploit information encoded in the tangential motion v_t^0 not recognizable by a single sensor alone. Furthermore, it enables the system to identify differences in the envelope scaling of motion patterns across the sensors, which allows the inference of information about the directions of motions. Clearly, both points benefit from high Doppler resolution, particularly for small multistatic angles. In addition, deploying a sensor network can increase radar cross-sectional diversity and allows the observation of body parts occluded by the torso for some of the sensors.

B. System Setup

As measurement system, we utilize an incoherent radar sensor network consisting of three identical FMCW radar sensors placed along a horizontal line (cf. Fig. 2). From the parameters specified in Table 1, range and velocity resolution can be specified as $\Delta R = 15.0$ cm and $\Delta v = 8.1$ cm s^{-1}, respectively. Each sensor deploys a 1×4 antenna array. We denominate the system as incoherent, as we do not establish phase synchronization between the different sensors [10]. However, in order to suppress interference between the nodes, the sequences of FMCW ramps are time-synchronized via an external trigger signal.

TABLE 1. Radar Parameters.

Parameter	Value	Description
f_c	76.5 GHz	center frequency
B	1 GHz	bandwidth
T_c	128 µs	up-ramp time
T_{RRI}	190 µs	ramp repetition interval
N_c	128	number of chirps per sequence
N_s	512	number of samples per chirp
Δf	60 MHz	FDM frequency offset
T_{block}	71.4 ms	block repetition time

Furthermore, the sensors transmit at slightly different frequencies in a frequency-division multiplexing (FDM) mode [11], which allows the suppression of the bistatic responses. As a result, each radar sensor digitizes and processess monostatic responses only, i.e., reflections of its own transmit signal.

C. Radar Signal Processing

Following the derivation in [12], a transmitted chirp at node n can be expressed as

$$s_{t,n}(t) = \sin\left(2\pi\left(f_{c,n}t + \frac{B}{2T_c}t^2\right) - \phi_{T0}\right) \quad (2)$$

with the center frequency $f_{c,n}$ slightly different for each node due to FDM operation. The monostatic receive signal $s_{r,n} = s_{t,n}(t - \tau_n)$ is downconverted by mixing with a sensor's transmit signal. With the delay between transmit and receive signal expressed as $\tau_n = 2(R_n + v_{r,n}t)$, the intermediate frequency signal can be obtained as

$$s_{IF,n} = \sin\left(2\pi\left(\frac{2f_{c,n}R_n}{c} + \left(\frac{2f_c v_{r,n}}{c} + \frac{2BR_n}{T_c c}\right)t\right)\right). \quad (3)$$

For chirp sequence modulation, N_c chirps are transmitted with ramp repetition interval T_{RRI}. Sampling along the chirps and applying a 2-D fast Fourier transform, range-Doppler matrices can be obtained, with the target located in its corresponding range and Doppler bin. We sum up the absolute-value per-channel matrices at each sensor noncoherently to increase the signal-to-noise ratio, eventually resulting in three range-Doppler matrices. For each measurement, N_{block} chirp sequence blocks are recorded, resulting in the overall sequence of range-Doppler maps $s^k_{2D,n,i,j}$ with $k = 0, ..., N_{block} - 1$, $i = 0, ..., N_c - 1$, $j = 0, ..., N_s - 1$. Each range-Doppler matrix is collapsed along range dimension via a maximum operation, resulting in Doppler frequency vectors $\vec{f}^k_{D,n}$, $k = 0, ..., N_{block} - 1$, with entries $f^k_{D,n,i} = \max_j s^k_{2D,n,i,j}$. Then, the Doppler frequency vectors are concatenated along the chirp sequence blocks to obtain short-time Fourier transform (STFT)-like, Doppler spectrograms $DP_n = [\vec{f}^0_{D,n}, ..., \vec{f}^{N_{block}-1}_{D,n}]$, $n = \{0, 1, 2\}$. Finally, the power spectrograms are converted to logarithmic scale to better capture the high dynamic range of the measurements for the classifier.

III. NEURAL NETWORK ARCHITECTURE

A. Classifier With In-Net Data Fusion

Due to the image-like character of spectrograms of all kind, CNNs have been shown to be well suited as classification algorithm. In order to obtain a simple-structured yet powerful classifier, we adapt the architecture of VGG-16 [13] to obtain the modular classifier in Fig. 3. The model consists of $\log_2(d_I) - 1$ convolutional modules, with d_I

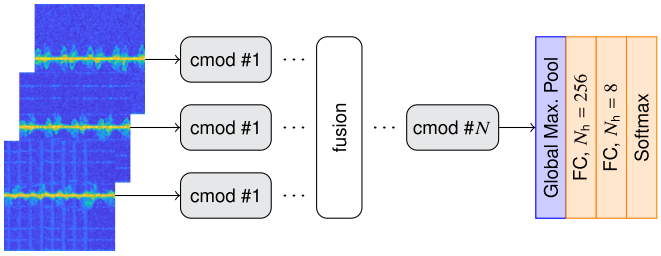

Fig. 3. Architecture of the classifier. The CNN consists of stacked convolutional modules, each comprising N_{conv} convolutional layers. The spectrograms from the nodes are fused within the network at a variable position, and different fusion schemes are investigated.

being the edge length of the quadratic input samples. For spectrograms of size 64×64, the CNN comprises five convolutional modules. Each convolutional module (*cmod*) contains N_{convs} convolutional layers followed by one 2×2 max-pooling layer and doubles the number of filters, starting from a predefined initial width. We tuned the overall depth and width of the convolutional stack via N_{convs} and the initial width with data stacked at the input until accuracy saturated, which happened for $N_{\text{convs}} = 2$ and an initial width of 16. Thus, the network comprises 10 convolutional layers, and the five modules have widths of $(16 - 32 - 64 - 128 - 256)$ filters.

In addition, we test different fusion positions by inserting a fusion layer, while all prior modules keep shared weights. Within the fusion layer, different fusion schemes for the per-sensor feature maps are tested—namely, *max-pool*; *cat*: concatenation of feature maps along channel dimension; *conv*: concatenation of feature maps followed by a 1×1 convolutional layer for downsampling along channel dimension, reducing the number of feature maps again; and *cat-ext*: concatenation of feature maps while doubling the number of filters in the subsequent convolutional module, such that the number of filters is ensured to be larger than the number of feature maps. This helps to extract the maximum amount of information from the stacked feature maps.

B. Gesture Dataset

We measured radar responses of eight gestures that might be used, e.g., in nonverbal communication in traffic scenarios, illustrated in Fig. 4. Extensive data were recorded from 45 subjects and under two extreme cases of the subjects' orientation (cf. Fig. 2) to produce high intraclass variations. Each measurement was conducted twice so that the dataset comprises $45 \cdot 8 \cdot 2 \cdot 2 = 1440$ measurements in total. A single measurement comprises 80 frames, corresponding to a measurement duration of 5.71s. The Doppler spectrograms DP_n are subsampled by moving a window with a length of 2s and step width of 0.5s over the measurements. This step results in a significant increase in training samples, thereby also serving as a means of data augmentation and fostering temporal translation invariance. All spectrograms are resampled to a size of 64×64 and normalized to $[0, 1]$.

In addition, to investigate the impact of large intraclass variations and low Doppler frequency scenarios on classification accuracy, we derive several subsets from the dataset: S_0 contains only gestures recorded under 0°, i.e., with face toward the sensor plane. This subset serves as baseline, as it reflects the way in which gestures are typically recorded in the literature. S_1 contains gestures recorded under both orientations $(0^\circ, 90^\circ)$ and therefore introduces large intraclass variations. Finally, in S_2, the orientation for each gesture is chosen to provide low Doppler frequency content, i.e., with its main motion component tangential to the sensor plane.

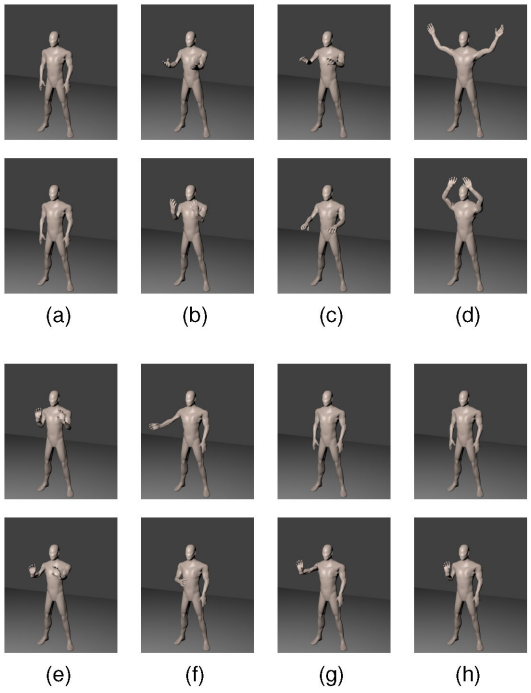

Fig. 4. Illustration of start (top row) and end (bottom row) pose for the eight arm gestures that constitute the gesture set. (a) Fly. (b) Come closer. (c) Slow down. (d) Wave. (e) Push away. (f) Wave through. (g) Stop. (h) Thank you.

TABLE 2. Cross-User Validation Accuracy and Enhancement w.r.t. Using Sensor 0 Only for Varying Combinations of Radar Nodes and Spectrograms Concatenated at the Input.

Set/Nodes	0	1	2	0,1	0,2	0,1,2
S_0	84.1 %	84.7 %	84.0 %	87.6 %	89.0 %	89.8 %
ΔS_0	-	0.6 %	−0.1 %	3.5 %	4.9 %	5.7 %
S_1	80.0 %	79.0 %	80.0 %	82.9 %	86.0 %	86.0 %
ΔS_1	-	−1.0 %	0.0 %	2.9 %	6.0 %	6.0 %
S_2	75.7 %	74.9 %	76.9 %	78.9 %	83.5 %	84.4 %
ΔS_2	-	−0.8 %	1.2 %	3.2 %	7.8 %	8.7 %

IV. EXPERIMENTAL RESULTS

The approach was validated by training the CNN on the different data subsets. For training, we generate nine fixed but random disjoint user subsets, each containing data from five participants. These subsets were used to apply leave-N-out cross-validation, performing the training ninefold while alternating the validation set. The results are averaged to obtain the cross-user validation accuracy that allows assessment of the cross-user generalization. All results are obtained by training with stochastic gradient descent with momentum with a learning rate of 0.003 and L2 regularization with a weight decay parameter of $\lambda = 0.001$.

A. Inherent Multistatic Gain

First, we train the classifier with Doppler data from different sensor combinations concatenated at the input. We refer to this as inherent multistatic gain, as no further optimization is performed and the performance boost stems from the availability of multistatic data only. The cross-user validation accuracy and the enhancement is summarized in Table 2, and classification on all subsets strongly benefits from

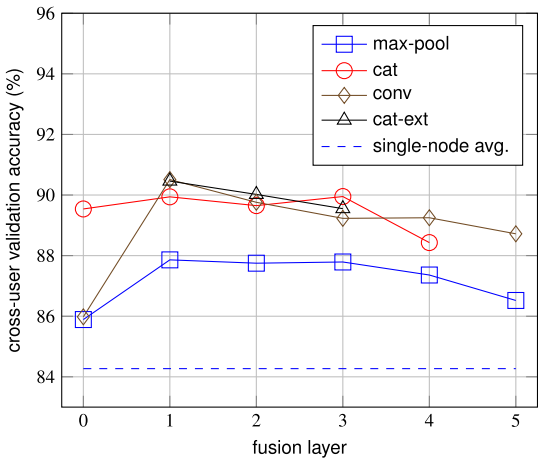

Fig. 5. Cross-user validation accuracy for subset S_0 with different fusion strategies and positions.

multistatic information. For S_0, using data from all sensors increases accuracy by 5.7% compared to sensor 0 only. For the most complex subset S_1 accuracy enhancement doubles using sensors (0,2) that span a multistatic angle of $\alpha_{\mathrm{MS}} = 20.9^\circ$ with respect to the pair (0,1) with $\alpha_{\mathrm{MS}} = 10.5^\circ$. Relative to the single-node average of 79.7% for S_1, the improvement amounts to 6.3%. The largest boost of 8.7% compared to node 0 only experiences S_2, as the sensor network mitigates the particularly low Doppler frequencies, thereby increasing classification robustness. Note that the main error contribution comes from confusion of gestures "Stop" and "Thank you" due to their similar motion patterns.

In addition, the comparison of different sensor combinations also provides further insights into the most valuable multistatic features: Using a side-looking node like sensor 2 alone results in no enhancement, despite this sensor captures more of the tangential velocity components. However, using multiple sensors leads to a significant improvement. This indicates that classification mainly benefits from differential features, i.e., the changes in how motion components are perceived across the sensors according to (1). By identifying and comparing patterns in the spectrograms of different nodes, the classifier is assumed to infer information about the direction of the underlying motions. Hardly surprising, this effect is heightened by wider multistatic angles α_{MS}.

B. Optimal Fusion Strategy

Classification can further be enhanced by incorporating fusion within the network. Therefore, we apply the fusion strategies introduced above at different positions within the CNN to find the optimal fusion strategy. Fig. 5 shows the accuracies achieved for subset S_0, in comparison to the single-node average of 84.3%. The best results are obtained for low-level feature fusion with *conv* or *cat-ext* after the first convolutional module with an accuracy of 90.5%. Both schemes enhance accuracy by 0.7% compared to input fusion and by 6.2% compared to the single-node average. For the most complex subset S_1, the *cat-ext* approach leads to an overall accuracy of 87.1%, enhancing the single-node average of 79.7% by 7.4%. Furthermore, *cat-ext* outperforms *conv* by 0.9% on S_1. We assume this to be a consequence of the larger intraclass variations in S_1, resulting in more diverse features before fusion that can lead to information loss in the channel downsampling step.

V. CONCLUSION

In this letter, we demonstrated the potential of incoherent radar sensor networks for robust and more accurate radar-based gesture recognition in complex scenarios such as urban traffic. We showed that utilization of Doppler information obtained by three time-synchronized FMCW radar sensors which are separated horizontally enables a classifier to deal with orientation-induced large intraclass variations by exploiting differential features. Moreover, the angular diversity helps mitigating inaccuracies related to low Doppler frequencies. By further optimizing data fusion within the deployed CNN, we boost cross-user validation accuracy on our complex dataset by 7.4% compared to the single-sensor case.

ACKNOWLEDGMENT

This work was supported by the Ministerium für Wissenschaft, Forschung und Kunst (MWK) Baden-Württemberg within the Project INTUITIVER.

REFERENCES

[1] J. Hasch, E. Topak, R. Schnabel, T. Zwick, R. Weigel, and C. Waldschmidt, "Millimeter-wave technology for automotive radar sensors in the 77 GHz frequency band," *IEEE Trans. Microw. Theory Techn.*, vol. 60, no. 3, pp. 845–860, Mar. 2012.

[2] J. Dickmann*et al.*, "Automotive radar the key technology for autonomous driving: From detection and ranging to environmental understanding," in *Proc. IEEE Radar Conf.*, 2016, pp. 1–6.

[3] Y. Kim and B. Toomajian, "Hand gesture recognition using micro-Doppler signatures with convolutional neural network," *IEEE Access*, vol. 4, pp. 7125–7130, 2016.

[4] B. Dekker, S. Jacobs, A. S. Kossen, M. C. Kruithof, A. G. Huizing, and M. Geurts, "Gesture recognition with a low power FMCW radar and a deep convolutional neural network," in *Proc. Eur. Radar Conf.*, 2017, pp. 163–166.

[5] S. Wang, J. Song, J. Lien, I. Poupyrev, and O. Hilliges, "Interacting with Soli," in *Proc. 29th Annu. Symp. User Interface Softw. Technol.*, 2016, pp. 851–860.

[6] Z. Zhang, Z. Tian, and M. Zhou, "Latern: Dynamic continuous hand gesture recognition using FMCW radar sensor," *IEEE Sensors J.*, vol. 18, no. 8, pp. 3278–3289, Apr. 2018.

[7] S. Hazra and A. Santra, "Short-range radar-based gesture recognition system using 3D CNN with triplet loss," *IEEE Access*, vol. 7, pp. 125 623–125 633, 2019.

[8] Le Yang and G. Li, "Sparsity aware dynamic gesture recognition using radar sensors with angular diversity," *IET Radar, Sonar Navigation*, vol. 12, no. 10, pp. 1114–1120, 2018.

[9] Z. Chen, G. Li, F. Fioranelli, and H. Griffiths, "Dynamic hand gesture classification based on multistatic radar micro-Doppler signatures using convolutional neural network," in *Proc. IEEE Radar Conf.*, 2019, pp. 1–5.

[10] A. Haimovich, R. Blum, and L. Cimini, "MIMO radar with widely separated antennas," *IEEE Signal Process. Mag.*, vol. 25, no. 1, pp. 116–129, Jan. 2008.

[11] A. Frischen, J. Hasch, and C. Waldschmidt, "A cooperative MIMO radar network using highly integrated FMCW radar sensors," *IEEE Trans. Microw. Theory Techn.*, vol. 65, no. 4, pp. 1355–1366, Apr. 2017.

[12] V. Winkler, "Range Doppler detection for automotive FMCW radars," in *Proc. Eur. Radar Conf.*, 2007, pp. 166–169.

[13] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014. [Online]. Available: http://arxiv.org/pdf/1409.1556v6