Microwave Engineering
Ulm University

universität ulm

Joint Pedestrian Gesture Recognition and Orientation Estimation from Multistatic Radar Data

Nicolai Kern, Ahmed Badr, Timo Grebner, Pirmin Schoeder, Christian Waldschmidt

European Microwave Week 2022 Conference Proceedings.
DOI: 10.23919/EuRAD54643.2022.9924874

Joint Pedestrian Gesture Recognition and Orientation Estimation from Multistatic Radar Data

Nicolai Kern[1], Ahmed Badr, Timo Grebner[1], Pirmin Schoeder[1], Christian Waldschmidt[1]
Ulm University, Institute of Microwave Engineering, 89081 Ulm, Germany
[1]{author}@uni-ulm.de

***Abstract*— Radar-based detection of traffic gestures can play an important role in the interaction between autonomous vehicles (AVs) and vulnerable road users (VRUs). However, it is error-prone without information about the orientation of the VRU, from which the recipient of the gesture's message can be inferred. Hence, this paper proposes a neural network with two output branches for joint gesture classification and orientation estimation. As input serve radar target lists derived from data measured by an incoherent radar sensor network. By training the whole model on a combined loss, the neural network simultaneously produces a gesture prediction and an orientation estimate. The proposed method is validated on a comprehensive dataset containing radar data over a broad range of orientations from 35 participants. On this dataset, it achieves a gesture recognition accuracy of $92.43\,\%$ and a mean orientation estimation error of $8.02°$.**

***Keywords*— automotive radar, gesture recognition, neural networks, orientation estimation, radar sensor networks**

I. Introduction

A particular challenge for autonomous vehicles (AVs) navigating urban areas is the interaction with vulnerable road users (VRUs) like pedestrians in scenarios necessitating communication. One way to resolve these situations is to make AVs understand the gestures performed by pedestrians. However, as illustrated in Fig. 1, the correct distinction between different gestures is not sufficient to guarantee their correct interpretation: In complex scenarios involving multiple vehicles, it is also crucial to have knowledge about the orientation of the gesture performing pedestrian in order to identify the recipients.

While the orientation information is readily available when doing camera-based gesture recognition e.g. from 3D skeletal poses [1], it is harder to obtain when using other sensor types such as radar sensors. Radars are robust with respect to environmental conditions and have been shown to be able to distinguish different activities [2] and gestures [3] at distances in the meter range, but they do not inherently provide orientation information. So far, the majority of work on radar-based orientation estimation is concerned with general objects [4] and cars [5], [6], whose ground truth position can be described by rigid bounding boxes. However, bounding boxes might be hard to apply to VRUs performing traffic gestures, since their appearance in the radar image strongly depends on their pose, which in turn depends on the gesture itself. Therefore, this paper proposes a neural network that directly estimates the orientation of gesturing pedestrians,

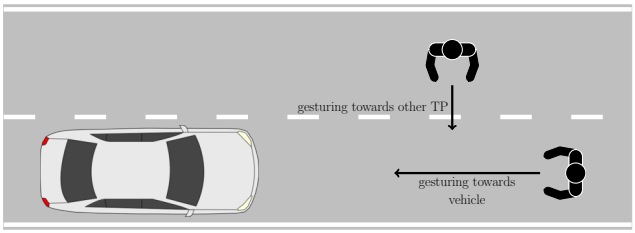

Fig. 1. Illustration of the orientation challenge: The orientation of the gesturing pedestrian determines whether a detected gesture is intended for the AV or an other traffic participant (TP).

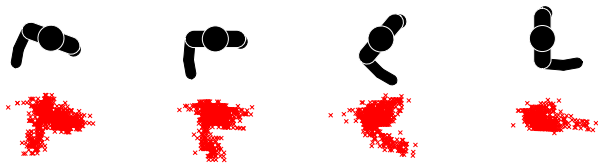

Fig. 2. Positions of the detected targets for a pedestrian signaling "Stop", measured under different orientations and accumulated over 30 frames.

while simultaneously performing gesture classification, and thus implicitly leveraging the gesture information. The neural network consists of two parts: A PointNet [7] for per-frame feature extraction, and a subsequent long short-term memory (LSTM), from whose outputs gesture predictions and orientation estimates are inferred in separate branches. The radar signal processing is explained in Sec. II, while the neural network is introduced in detail in Sec. III. Finally, experimental results on a comprehensive measured multistatic radar dataset are presented in Sec. IV.

II. Generation of Radar Target Lists

The proposed neural network operates on multistatic radar target lists recorded by an incoherent radar sensor network. The fusion of the radar data is done within the neural network itself. Therefore, the extraction of target lists and their preprocessing is done independently for each sensor. First, range-Doppler maps are extracted for each chirp sequence frame by computing the 2D fast Fourier transform along the samples and chirps [8]. In order to obtain a compressed representation of the range-Doppler maps that preserves the relevant information, targets are extracted by means of the ordered statistics constant false alarm rate (CFAR) algorithm [9]. Besides the target's

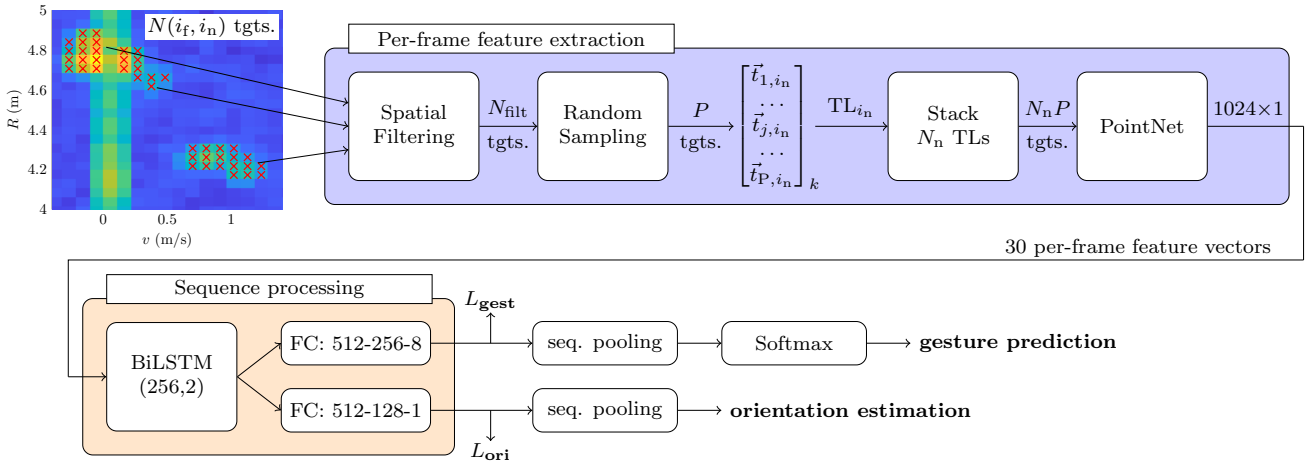

Fig. 3. Illustration of the joint gesture classification and orientation estimation. An LSTM learns on sequences of extracted per-frame feature vectors, and the model is trained on the joint loss $L_\mathrm{gest}+L_\mathrm{ori}$.

range R_n, velocity v_n, and signal-to-noise ratio SNR_n, the azimuth angle θ_n is computed for each target as well as its positions $[x_n, y_n]$. The positions of the accumulated targets of 30 consecutive frames are shown for a "Stop" gesture under different orientations in Fig. 2. Each detection's position is referenced to the pedestrian's estimated position obtained from a regression model to obtain its normalized position described by x_n^norm and y_n^norm. All target parameters are bundled in a radar feature vector $\vec{t}_n = [R_n, v_n, \theta_n, x_n^\mathrm{norm}, y_n^\mathrm{norm}, \mathrm{SNR}_n, i_\mathrm{n}]$ describing the target, and the node index $i_\mathrm{n} = 0, ..., N_\mathrm{n}-1$ determines to which of the N_n radar sensor nodes the detection belongs. The feature vectors of the $N(i_\mathrm{f}, i_\mathrm{n})$ CFAR detections form the radar target list for the radar node i_n and the frame i_f.

III. Joint Gesture Classification and Orientation Estimation Algorithm

The joint gesture classification and orientation estimation algorithm consists of two main parts, as shown in Fig. 3. First, for each frame the target list is computed out of the detections in the range-Doppler map, which is then converted into a fixed-length feature vector describing the input. From this sequence of vectors, gesture predictions and orientation estimates are inferred by making use of the temporal information encoded in the vector sequence. In this work, the algorithm is optimized for sequences of 30 chirp sequence frames at a frame rate of 15 fps, so each prediction incorporates the previous two seconds of observation.

A. Spatial Filtering

Since measurements in complex environments contain reflections from a multitude of targets, the target lists are cleaned for each frame in a spatial filtering step. Based on a pedestrian position estimate $[\tilde{x}_p, \tilde{y}_p]$, e.g. out of a pedestrian detection step, target list entries are removed if they don't fulfill the inequation

$$\sqrt{(x_n - \tilde{x}_\mathrm{p})^2 + (y_n - \tilde{y}_\mathrm{p})^2} \leq d_\mathrm{filt}\,, \tag{1}$$

i.e. if they lie outside of a circular filter area with radius d_filt, which is set to $d_\mathrm{filt} = 2\,\mathrm{m}$.

B. Random Target Sampling and Input Fusion

While it is advantageous for the training process to have a constant number of targets per frame, the actual number of targets after spatial filtering, $N_\mathrm{filt}(i_\mathrm{f}, i_\mathrm{n})$, varies strongly depending on the gesture, pedestrian position, and physiology. Therefore, the target lists are adjusted to equal lengths to achieve a constant number of P targets per frame and node. In case that $N_\mathrm{filt}(i_\mathrm{f}, i_\mathrm{n}) < P$, the target list is zero-padded by adding empty target vectors. In contrast, if $N_\mathrm{filt}(i_\mathrm{f}, i_\mathrm{n}) > P$ then P targets are sampled randomly from the target list. Finally, the target lists from all radar nodes are fused into a single target list of length $N_\mathrm{n}P$, where the information about which node detected a target is still kept by the node index i_n.

C. Per-frame Feature Vector Extraction

The final step in the per-frame feature extraction is the generation of a fixed-length feature vector out of the input target list using PointNet [7]. By this, the sparse, multi-dimensional input point clouds are transformed into a structured representation suitable for the subsequent sequence processing. The PointNet approximates a function f operating on the input target list $\{\vec{t}_1, ..., \vec{t}_{N_\mathrm{n}P}\}$ by a symmetric function g that summarizes the outputs of a point-wise function h applied to the targets $\vec{t}_n$ individually, i.e.

$$f(\{\vec{t}_1, ..., \vec{t}_\mathrm{N}\}) \approx g(h(\vec{t}_1), ..., h(\vec{t}_\mathrm{N}))\,. \tag{2}$$

The model used for feature extraction computes a 1024-element vector $h(\vec{t}_n)$ for each radar feature vector, and the vectors of all $N_\mathrm{n}P$ targets are then fused into a single feature vector of shape 1024×1 describing the input target list by taking the element-wise maximum over the targets. Note that since the radar target features are either linked through non-linear relationships or not linked at all, the input transform used in [7] is omitted.

D. Sequence Processing for Joint Classification and Regression

To infer predictions based on multiple consecutive frames, a bidirectional LSTM (BiLSTM) is optimized to learn the temporal dynamics of feature vector sequences. The resulting BiLSTM has two recurrent layers for both forward and backward direction, and each layer has 256 neurons. The bidirectional model enhances per-frame predictions particularly at the beginning of the observation interval, i.e. the frames most distant to the recently measured one, which enhances the per-sequence results. Thus, the BiLSTM output at each frame encapsulates information from previous and following frames. From the BiLSTM output, per-frame predictions inferred: For the per-frame gesture prediction, the BiLSTM output is passed through two fully-connected (FC) layers producing a score value for each of the eight gestures, based on which the classification loss L_{gest} is computed. In order to simultaneously estimate the orientation under which the gesture is performed, the output of the BiLSTM is passed through a separate stack of FC layers with a single output neuron representing the predicted orientation. From this prediction, the orientation loss L_{ori} is computed. Then, the whole model is trained on the per-frame loss

$$L = L_{\text{gest}} + L_{\text{ori}} \tag{3}$$

$$= -\log\left(\hat{y}_{\text{gest}}\right) + \left(\hat{y}_{\text{ori}} - y'_{\text{ori}}\right)^2 , \tag{4}$$

which consists of the cross-entropy loss L_{gest} with the gesture prediction $\hat{y}_{\text{gest}}$ and the mean squared error loss L_{ori} with the orientation prediction and label, $\hat{y}_{\text{ori}}$ and y'_{ori}, respectively. Per-sequence gesture predictions are obtained from the per-frame gesture scores in a sequence pooling step by averaging the scores over the frames, resulting in a single prediction for the recently observed time interval. Similarly, per-frame orientation estimates are pooled into a per-sequence orientation prediction by averaging.

The PointNet+LSTM model is trained end-to-end for 100 epochs using stochastic gradient decent with momentum to minimize L, with a weight decay of 0.001. The initial learning rate is set to 0.004 and reduced by a factor of 0.3 after 50 and 60 epochs, respectively. $N_{\text{filt}} = 100$ targets are used as input per frame and node, and all radar target features are normalized to the unambiguous ranges of the sensors. Note that constant accuracy levels are achieved over a wide range of N_{filt}. The orientation labels are rescaled by $y'_{\text{ori}} = y_{\text{ori}}/90$, with the label in degree, y_{ori}, to balance the initial values of L_{gest} and L_{ori}.

IV. Experimental Results

A. Experimental Setup

The dataset is recorded with an incoherent radar sensor network comprising three multiple-input multiple-output radar sensors with chirp sequence modulation, transmitting at 77 GHz. Each sensor has a range resolution of $\Delta R = 4.5\,\text{cm}$, a velocity resolution of $\Delta v = 10.7\,\text{cm}\,\text{s}^{-1}$, and an array with 12 virtual antenna elements used for azimuth beamforming. As shown in the top-view sketch in Fig. 4, the maximum

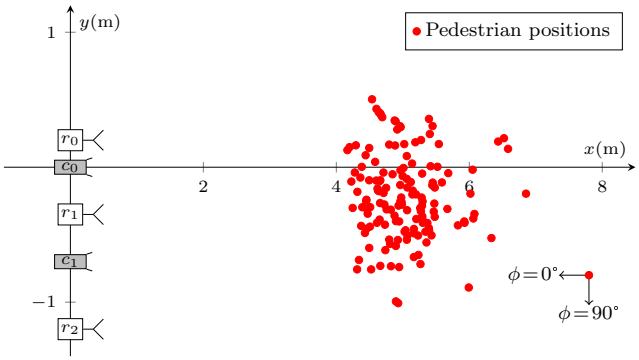

Fig. 4. Top-view sketch of the experimental setup consisting of an incoherent radar sensor network (r_0, r_1, r_2) and a stereo video system (c_0, c_1) for ground truth capture. In addition, the pedestrian positions during the measurements as derived by the stereo system are shown.

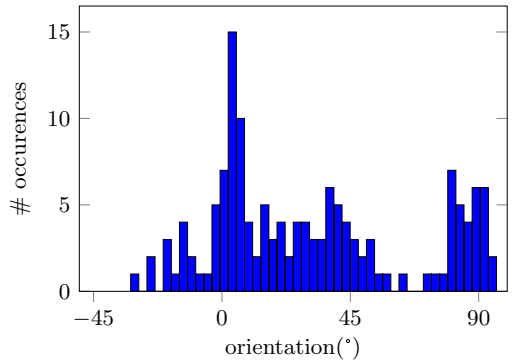

Fig. 5. Histogram of the orientation values in the measured dataset.

sensor distance is 140 cm. In order to avoid interference, all sensors receive a common trigger and transmit chirps with a slight sensor-dependent time delay. The radar sensor network is complemented by two RGB cameras calibrated for stereo depth estimation and synchronized to the radar frame starts. The stereo video data is utilized for the generation of ground truth information about the 3D skeletal pose of the participants. For this purpose, skeletal keypoints are detected with *Detectron2* [10] and their 3D positions are computed by triangulation. From the 3D skeletal information, the orientation labels for the computation of L_{ori} are extracted from the positions of the hip joints, with the definition of the orientation values as shown in Fig. 4.

B. Dataset

Radar and stereo video data are recorded for eight different traffic gestures, namely “Fly”, “Come Closer”, “Slow Down”, “Wave”, “Push Away”, “Wave Through”, “Start”, and “Stop”, which are visualized in [3]. The gestures are recorded for 35 participants, under different orientations, and both indoors and outdoors. In order to properly cover the presumably most common orientations, every participant was recorded under 0° and 90°. In addition, up to three measurements were taken under arbitrary orientations to reflect the variability in real-world scenarios. The distribution of the measurements

Table 1. Cross-validation classification accuracy (Gest. Acc.) and orientation estimation mean error (ME Ori. Est.) results

Training procedure	**Gest. Acc.**	**ME Ori. Est.**
Training on L_{gest}	92.1 %	-
Training on L_{ori}	-	9.9°
Training on $L_{\text{gest}}+L_{\text{ori}}$	92.4 %	8.0°
Training on $L_{\text{gest}}+L_{\text{ori}}$, corr. class.	100.00 %	7.8°
Training on $L_{\text{gest}}+L_{\text{ori}}$, wrong class.	0.00 %	10.9°

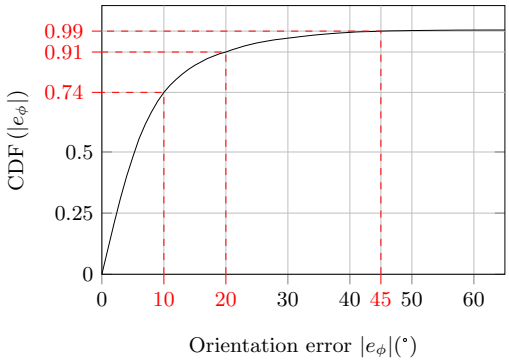

Fig. 6. CDF of the orientation estimation errors.

over the orientation angles as extracted from the stereo video data is shown in Fig. 5. Moreover, recordings are conducted at ranges of 4 m to 7 m, as shown in Fig. 4. After the radar signal processing, measurements are subdivided into snippets of two seconds, resulting in 15,700 samples for training and testing.

C. Classification and Orientation Estimation Accuracy

For the experimental validation of the proposed algorithm, the dataset is split into five subsets, each containing data from seven participants exclusively. Then, the neural network is trained 5-fold with one subset excluded for cross-validation in each fold. The average gesture recognition accuracies and mean orientation estimation errors are reported in Table 1. When training on the combined loss, the model achieves a classification accuracy of 92.43 % and a mean orientation error of 8.02°. Moreover, Fig. 6 shows the cumulative distribution function (CDF) of the error in the orientation estimate, with the share of samples exceeding the application-dependent acceptable error threshold being an important criterion.

In addition, the impact of faulty gesture detections on the orientation estimation accuracy is investigated. For the correctly classified samples, the orientation estimation is slightly better than the overall value. Contrary, the mean error increases to 10.88° in cases where the classifier branch predicts a wrong gesture, indicating that these samples are inherently less informative, e.g. due to low numbers of detections. Moreover, by comparing the results achieved with training on the combined loss L with the orientation estimation accuracy achieved by training on L_{ori} only, it can be seen that the simultaneous training for gesture recognition and orientation estimation significantly enhances the model's performance, reducing the mean error by 1.86°. This demonstrates that the additional training on the classification task leads to more descriptive LSTM outputs, providing the orientation estimation layers with additional cues. Furthermore, simultaneous training doesn't negatively affect the gesture recognition accuracy compared to training the model with only the gesture branch on L_{gest}, with cross-validation accuracy even slightly increased.

V. Conclusion

In this paper, the problem of ambiguous non-verbal communication in traffic scenarios due to gestures performed under unknown VRU orientations was approached by a radar-based joint gesture recognition and orientation estimation model. A PointNet is used to extract feature vectors out of multistatic radar target lists, and two independent branches following a BiLSTM simultaneously predict the gesture and orientation. Based on a challenging dataset it is shown that the joint approach with training on a combined loss is superior to the standalone orientation estimation, with the mean orientation estimation error at 8.02° being 1.86° lower. Also importantly, the joint model still provides a high gesture recognition accuracy of 92.43 %, even slightly higher than the standalone classifier.

Acknowledgment

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

References

[1] J. Wiederer, A. Bouazizi, U. Kressel, and V. Belagiannis, "Traffic control gesture recognition for autonomous vehicles," in *IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2020, pp. 10 676–10 683.

[2] A. Shrestha, H. Li, J. Le Kernec, and F. Fioranelli, "Continuous human activity classification from FMCW radar with Bi-LSTM networks," *IEEE Sens. J.*, vol. 20, no. 22, pp. 13 607–13 619, 2020.

[3] N. Kern, M. Steiner, R. Lorenzin, and C. Waldschmidt, "Robust Doppler-based gesture recognition with incoherent automotive radar sensor networks," *IEEE Sens. Lett.*, vol. 4, no. 11, pp. 1–4, 2020.

[4] G. Zhang, H. Li, and F. Wenger, "Object detection and 3d estimation via an FMCW radar using a fully convolutional network," in *IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2020, pp. 4487–4491.

[5] F. Roos, D. Kellner, J. Dickmann, and C. Waldschmidt, "Reliable orientation estimation of vehicles in high-resolution radar images," *IEEE Trans. Microw. Theory Techn.*, vol. 64, no. 9, pp. 2986–2993, 2016.

[6] J. Schlichenmaier, N. Selvaraj, M. Stolz, and C. Waldschmidt, "Template matching for radar-based orientation and position estimation in automotive scenarios," in *IEEE MTT-S Int. Conf. Microw. Intell. Mobility*, 2017, pp. 95–98.

[7] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "PointNet: Deep learning on point sets for 3D classification and segmentation," in *Proc. of the IEEE Conf. Comput. Vis. Pattern Recog.*, July 2017.

[8] V. Winkler, "Range Doppler detection for automotive FMCW radars," in *Eur. Radar Conf.* Piscataway, NJ: IEEE, 2007, pp. 166–169.

[9] H. Rohling, "Radar CFAR thresholding in clutter and multiple target situations," *IEEE Trans. Aerosp. Electron. Syst.*, vol. AES-19, no. 4, pp. 608–621, 1983.

[10] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, and R. Girshick, "Detectron2," https://github.com/facebookresearch/detectron2, 2019.