Point Transformer-Based Human Activity Recognition Using High-Dimensional Radar Point Clouds

Guo, Zhongyuan ; Guendel, Ronny G.; Yarovoy, Alexander; Fioranelli, Francesco

DOI
10.1109/RadarConf2351548.2023.10149679

Publication date
2023

Document Version
Final published version

Published in
Proceedings of the 2023 IEEE Radar Conference (RadarConf23)

Point transformer-based human activity recognition using high-dimensional radar point clouds

Zhongyuan Guo*, Ronny G. Guendel†, Alexander Yarovoy†, Francesco Fioranelli†
*†Microwave Sensing, Signals and Systems (MS3), Delft University of Technology, Delft, Netherlands
*gary0037@163.com, †{r.guendel, a.yarovoy, f.fioranelli}@tudelft.nl

***Abstract*—Radar-based Human Activity Recognition (HAR) is considered by using snapshots of point clouds. Such point clouds interpret 2D images generated by an mm-wave FMCW MIMO radar enriched by including Doppler and temporal information. We use the similarity between such radar data representation and the core of the self-attention concept in artificial intelligence. Three self-attention models (Point Transformer) are investigated to classify Activities of Daily Living (ADL). An experimental dataset collected at TU Delft is used to explore the best combination of different input features, the effect of a proposed Adaptive Clutter Cancellation (ACC) method, and the robustness in a leave-one-subject-out scenario. Results with a macro F1 score in the order of 90% are demonstrated with the proposed method, including activities that are static postures with little associated Doppler.**

***Index Terms*—Human Activity Recognition, Imaging Radar, Deep Learning, Point Transformer, Activities of Daily Living.**

I. Introduction

Every country experiences growth in both the absolute size and the relative proportion of older people in their population [1]. This aging population creates a strong demand for novel healthcare provision approaches beyond traditional hospital-based systems. For example, technologies for remotely monitoring patients in their homes can provide early diagnosis of changes in behavioral patterns and detect critical events such as falls in a timely manner. These home-centric healthcare technologies can thus enhance the life quality of older and vulnerable people and, at the same time, minimize the disruption to their usual routine and lifestyle [2].

Initially, automated Human Activity Recognition (HAR) was implemented through video-based technologies [3], or wearable sensors [4], [5]. These sensors have inherent limitations [6] that radar-based sensing may complement. Compared with these sensors, radars cannot capture optical-like images or videos, which preserves people's privacy in their home environments. Besides, radars do not require the user to wear, carry, or interact with any sensors or wires, as it is a fully contactless sensing modality. These advantages can potentially help with compliance and acceptance from end-users and confirm the increasing appeal of radar technologies for HAR.

In recent studies on radar-based HAR, the typical processing pipelines are: (1) process the complex signals from radar systems to generate the desired data representations; (2) forward these generated data to the classifiers, such as deep neural networks or conventional classifiers, i.e., the Support Vector Machines (SVM), to identify the different human activities.

The most common 2D matrix *data representations* includes the typical micro-Doppler spectrogram [7], range-Doppler heatmap [8], range-time heatmap [9], amplitude spectrum [10], or range-angle heatmap [11]. These image-like data representations are hardly sensitive to static activities. In this case, the salient features of static postures are often overseen by commonly used radar data domains, e.g., the spectrogram, where no range or shape information is represented. Similarly, range profiles are incapable of representing well movements over time, represented by the Doppler information. In contrast, as many current research studies show, a radar point cloud can contain the target's shape information with either point clouds processed as 2D images [12], or via the spatial coordinates forwarded to the classifier [13], [14]. Currently, to the best of our knowledge, no research in HAR systematically explores the advantages and disadvantages of combining the coordinates of the point cloud with additional features such as Doppler and point intensity.

In terms of *classifiers*, the majority of radar-based HAR research uses Convolutional Neural Networks (CNNs) [7], Recurrent Neural Networks (RNNs) [15], [16], or mixed hybrid models combining these two architectures [17], [18]. CNNs treat the input radar data as images, mainly considering pixel-related features, while different implementations of RNNs treat radar data as temporal sequences focusing on temporal relations. Hybrid models take advantage of these two architectures combined. Although Transformers based on self-attention mechanisms [19] showed superiority to CNN and RNN architectures in many machine learning tasks and applications, minimal research has investigated the performance of these self-attention models on radar-based activity recognition [20]. Specifically, feature selection based on radar point clouds has been hardly used, and almost no conclusions were drawn determining the most relevant features of a point cloud in the context of HAR. The same lack of systematic study applies to the architecture of a classifier to exploit the point cloud's information effectively.

The specific contribution of this paper is to develop and test a pipeline that solely utilizes radar point clouds as input data to train attention-based deep-learning models for classification, specifically Point Transformers (PTs); this pipeline can classify both motions and static postures successfully. Three different architectures of state-of-the-art PTs are considered,

2023 IEEE Radar Conference (RadarConf23) | 978-1-6654-3669-4/23/$31.00 ©2023 IEEE | DOI: 10.1109/RADARCONF2351548.2023.10149679

with point clouds obtained from a Multiple Input Multiple Output (MIMO) imaging radars. Additionally, an effective Adaptive Clutter Cancellation (ACC) scheme is also proposed to pre-process the point clouds used as inputs to the proposed classification pipeline based on PTs.

The rest of the paper is organized as follows. Section II describes the proposed approach. Section III presents the measured dataset for validating the performance of the proposed method. Section IV discusses the attained results for the proposed method and its comparison among three self-attention implementations of PTs. Finally, conclusions are drawn in Section V.

II. Proposed Approach

MIMO mm-wave Frequency-Modulated Continuous-Wave (FMCW) radar can generate six intrinsic features of the target: range, azimuth angle, elevation angle, Doppler, intensity, and temporal relations. The first three features can be represented in cartesian coordinates and form the spatial aspect of the derived radar point clouds. For the last three features, the most common data representations apart from point clouds are range-Doppler heatmaps and spectrograms, as mentioned in the introduction. In the majority of the previous research works, these representations are separately used to recognize human activities. The pipeline proposed in this paper aims instead to integrate Doppler, intensity, and temporal information together with spatial information of point clouds to address HAR. This information-richer representation is exploited in conjunction with the self-attention models described in this section, which will be used as classifiers. An overview of the proposed pipeline is given in Figure 1.

A. Data Pre-Processing: Point Cloud Generation

The *first stage* in the pipeline is responsible for converting the raw radar data cubes containing complex signals to 6D point clouds including 3D spatial coordinates, Doppler (velocity), intensity (related to the Signal to Noise Ratio, SNR), and time. In this module, 2D Fast Fourier Transform (FFT) is first applied to 2D discrete signals to generate the range-Doppler map. Then, 2D OS-CFAR (Ordered Statistics Constant False Alarm Rate) is used on the map to detect the cells occupied by subjects, and the coordinates of the detected cells are the range and Doppler information of the point, while the values of the detected cells correspond to their intensity. Last but not least, an FFT is applied along the channel axis to estimate the azimuth angle and elevation angle. With the angle and range information, the 3D Cartesian coordinates can be derived.

B. Further Data Pre-Processing: Adaptive Clutter Cancellation

In practice, a lot of points in the point clouds from the previous pre-processing stage are not related to the target and can be considered as clutter. Furthermore, the number of points after the point cloud generation stage does not necessarily match the required input size of the Point Transformer (PT) network. For these two reasons, we need the second stage of data pre-processing described in this section. In this *second stage*, a method for removing the clutter is proposed with the following two steps. First, calculate the spatial centroid of the point cloud for each frame with the intensity values of the points as weights. Then, filter out the points with the distance to the centroid higher than $1m$, which is assumed to be a reasonable number for an average human body size. After removing clutter on a frame-by-frame basis, a defined number of frames with the highest Doppler components are selected from the temporal sequence of frames returned by the MIMO radar. This is done because points across multiple frames can better represent the characteristics of the motion. Lastly, we apply down-sampling or up-sampling on the remaining points of the cloud to match the input size of the PT network. The specific resampling algorithms are described in detail in section III [21].

C. Classifier: Self-Atention Point Transformer (PT) Models

The *third stage* of the pipeline is the classifier. In this paper, we investigated three different attention-based networks for processing point clouds. They are the Point Transformer (PT) of Hengshuang Zhao et al. [22], the PT of Menghao Guo et al. [23], and the PT of Nico Engel et al. [24]. For the first two networks, the authors employ a hierarchical architecture to extract the features of the input point cloud with an attention mechanism and fully connected layers to present the classification results. For Nico's model, local and global features are related by cross-multi-head attention after being extracted separately. Similar to the other two models, fully connected layers are deployed to provide the classification results. Additional information on the architectures of these networks can be retrieved in the thesis in [21], and in the original references for each proposed network.

III. Dataset Description

A dataset collected in [25] is used to validate the performance of the proposed pipeline. This section also provides more details about the second stage of data pre-processing.

A. Measurements

The radar used is an mm-wave MIMO FMCW radar developed by Texas Instruments (cascaded AWR2243 radar). Specifically, this radar has 12 TXs and 16 RXs operating at 79 GHz. The related radar parameters are given as follows: frame period of 100 ms, range resolution of 52.8 mm, velocity resolution of ± 0.0286 m/s, azimuth resolution of 1.4 degrees, elevation resolution of 18 degrees.

This dataset was collected in a lecture room of TU Delft with tables, chairs, and cabinets so that a real-life indoor environment is simulated. The radar was placed at 0.75m height from the ground to illuminate the whole human body in the field-of-view. A chair was placed approximately 2.7m away from the radar in the Y-axis direction and participants performed activities around it. Six activities were included in the dataset, consisting of 4 most common daily motions, and 2 static postures that can be viewed as the transitional

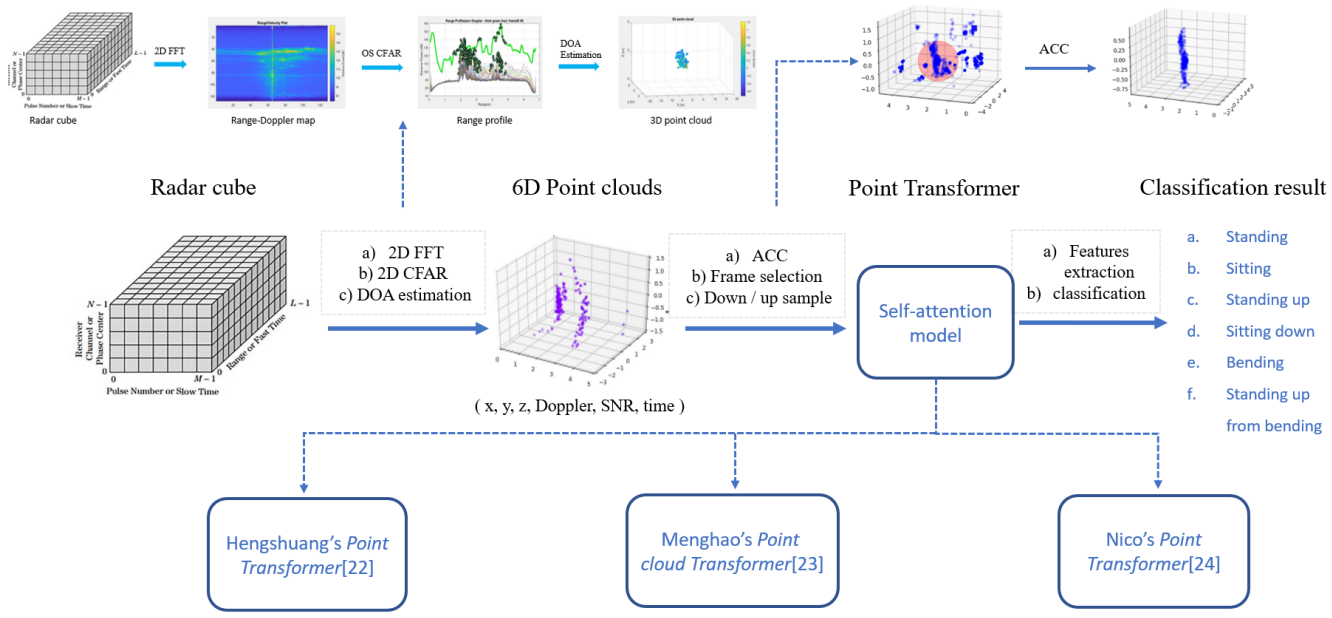

Fig. 1: Overview of the proposed Human Activity Recognition (HAR) pipeline, where the main contributions are the usage of different models of Point Transformer (PT) networks to recognize human activities exploiting radar point clouds, and the Adaptive Clutter Cancellation (ACC) at the pre-processing stage.

states between such motions (see the details in Table I). The measurements included independent records of the subjects' postures and movements. Specifically, a complete time interval for each measurement was 2 minutes. During this time, the human subjects were asked to perform either one static posture, such as sitting still on the chair, or a motion pair, such as sitting down and standing up from sitting for a period of 2 seconds for each individual motion. Additionally, the participants were asked to repeat the activities at four additional aspect angles of 45, 90, 135, and 180 degrees.

TABLE I: List of motions, motion pairs, and postures for the measured dataset. [25]

Motion pair	Motion	Posture	
a	1. Sitting down	c	5. Sitting still
	2. Standing up from sitting	d	6. Standing still
b	3. Bending over		
	4. Standing up from bending		

To summarize, data from 7 human subjects were captured for the initial verification of the proposed approach. Each activity was recorded for 2 minutes and each motion lasted for 2 seconds, so for each aspect angle, there are $420 = 120/2 \times 7$ samples for each activity, and $2520 = 420 \times 6$ samples in total.

B. Adaptive Clutter Cancellation

After the data pre-processing pipeline in section II.A, the point cloud data contain a lot of clutter since there are tables, cabinet,s and chairs around the human subjects. We proposed a method for ACC [21] to remove these clutter contributions. The general idea of this method is to locate the spatial centroid of the human body by calculating the mean of x, y, z coordinates from all the points weighted by their corresponding intensity values, as in the equations shown in the upper part of Figure 2. This weighted calculation allows to offset the effect of highly reflective but very localized clutter points. Then a sphere with a radius equal to $1m$ and centered in the centroid point is created to separate the target-related points within, and remove the clutter-related points outside the sphere, as illustrated in Figure 2.

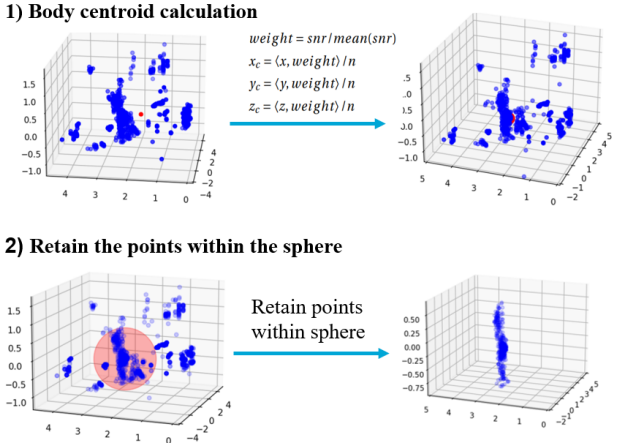

Fig. 2: An example of removing the clutter contributions from a point cloud of '*standing*' activity with the proposed Adaptive Clutter Cancellation (ACC).

Although not tested in this paper, a more complex shape such as an ellipsoid can also be defined as a better represen-

tation of a standing human subject. The position of the sphere is calculated for each frame to make it adaptive to the specific position and motion of the subject over time.

C. Frames Selection

Each motion is recorded for 20 frames (corresponding to 2s), but the period of actual movement is far less than 2s. Therefore, much information contained in the data of 20 frames is redundant to recognize a single motion, and only part of the 20 frames can contain the salient features. To filter the most valuable frames, we deploy a sliding window on the 20 frames with a step of 1 frame to calculate the average Doppler, and the window with the highest Doppler value is selected to extract the new data to represent the sample. The window lengths considered in this paper are 3, 5, and 10 samples, corresponding to 0.3, 0.5, and 1 second, respectively.
Since the detection algorithm applied on range-Doppler maps is OS-CFAR, the number of detected points per frame is not a constant. To ensure that the number of points per sample after ACC can fit the input size of the chosen PT model, we apply the Farthest Point Sampling algorithm to re-sample the point clouds [21]. The relation between the number of points and frames is shown in Table II.

TABLE II: Mapping relation between the number of frames and size of resulting input data, where 6 indicates the 'features' of each point of the point cloud (i.e., spatial coordinates, Doppler, intensity, time index).

Number of frames and equivalent time	Size of input per sample
20 (2s)	(1024,6)
10 (1s)	(512, 6)
5 (0.5s)	(256, 6)
3 (0.3s)	(256, 6)

IV. Classification Results of Point Transformer

To evaluate the proposed pipeline, 3 different self-attention-based networks are used as classifiers: the PT from Hengshuang [22], the point cloud transformer from Menghao [23], and the PT from Nico [24]. The classification performances of the proposed pipeline are analyzed in this section. Training and testing of the networks are performed in an Alienware laptop with an NVIDIA GeForce RTX 3070 Laptop GPU, with GPU memory equal to 8 GB.

A. Performances with Different Input Features

This section aims to identify the most suitable input features from the point cloud for recognizing the 6 activities. Specific results (without ACC) for different combinations of input features are shown in Figure 3. The PT network by Hengshuang was used in this initial analysis. Only with the spatial coordinates of the point cloud, the network is able to classify the six classes with an F1 score of roughly 0.741. Several instances of misclassification happen for paired motions such as standing up and sitting down that are symmetrical over time. That is, standing up can be regarded as the inverse motion of sitting down, so the point clouds of these two motions have almost the same spatial distributions, leading to some misclassifications among them. However, when adding independent Doppler and time information to the point clouds, there are notable improvements in F1 scores. The reason is that Doppler features and time information of points can represent the direction of movement and the order of appearance of the points, respectively. This variation in features breaks the spatial symmetry of the motions, leading to an increase in classification performance as shown in Figure 3, with F1 scores improving to above 0.86. However, when adding intensity as an extra feature, the F1 score is reduced to about 0.7. The main reason for this low performance is that the intensity will highly fluctuate while subjects perform motions, so the intensity distribution varies from sample to sample for the same motion. After comparing the F1 scores of different input feature combinations, the best input features are 3D spatial coordinates plus Doppler and time.

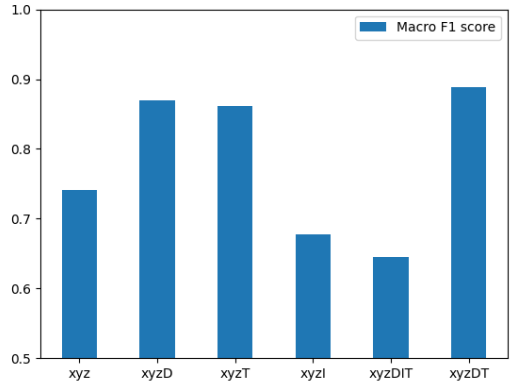

Fig. 3: Macro F1 scores for different input features combinations, where *XYZ* are 3D spatial coordinates, and *D I T* indicate Doppler, intensity, and time, respectively. The PT network by Hengshuang is used for these results.

TABLE III: F1 score of PT by Hengshuang with and without the proposed Adaptive Clutter Cancellation (ACC).

Input features	With ACC	Without ACC	Difference
xyz	0.792	0.741	+0.051 (5%)
xyzD	0.893	0.869	+0.024 (2.4%)
xyzT	0.880	0.861	+0.021 (2.1%)
xyzDT	0.928	0.888	+0.040 (4%)

As described in Section III, many points in the point clouds are actual clutter contributions due to the items around the human subjects, so the proposed ACC is applied to remove the clutter in the experimental scene. The improvements brought by ACC for different features as input are listed in Table III. It is noticeable that after removing the clutter points outside the human movement area, the improvement in the F1 score is significant. For the best case (*XYZDT*, i.e., spatial coordinates plus Doppler plus time as input features), the F1 score improves by about +4% and reaches 92.8%. For other

cases, ACC can also bring improvements, and it benefits most the case with only spatial coordinates as input.

By using the best combination of features plus the proposed clutter cancellation, an improvement in overall classification performances of +5.8% is achieved, compared with our previous work on the same TUD dataset that did not use attention-based networks, but a combination of spectrograms and PointNet to process radar point clouds [25].

B. Performance Comparison among the Three Attention-Based Point Transformer Networks

The performances of the three Point Transformer (PT) networks with a decreasing number of frames to extract the input data are displayed in Figure 4. These results are computed by applying a 5-fold cross-validation approach with 80% and 20% of data for training and testing, respectively. There are two noticeable characteristics reflected by the results: (1) F1 scores decline with fewer input frames, but data from 3 frames (equivalent to 0.3 seconds) are still enough for Hengshuang's and Menghao's networks to achieve reliable predictions; (2) Hengshuang and Menghao's networks are performing significantly better than Nico's. The differences can be related to the architectures of these three networks, where the self-attention mechanism is deployed in each hierarchical block to extract features for Hengshuang's and Menghao's architecture, while Nico's network just employs self-attention to relate features. This also shows the good match of radar point cloud representations with self-attention mechanisms.

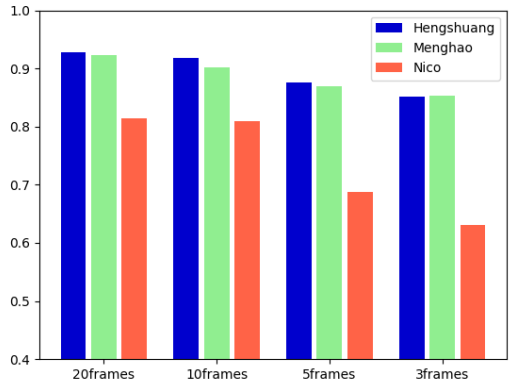

Fig. 4: Classification F1 score for 3 different Point Transformer implementations with decreasing number of frames as input, where the horizontal axis represents the number of frames. The F1 scores are the average from 5-fold cross-validation.

C. Human Activity Recognition with Different Aspect Angle

As human activities can be performed at any aspect angle with respect to the line-of-sight of the radar, we test the potential of the proposed pipeline to realize activity recognition at aspect angles of 0, 45, 90, 135, and 180 degrees. The performances are shown in Figure 5 using Hengshuang's network and the 3D spatial coordinates plus Doppler and time as input features. As Doppler values depend on the aspect angle and, theoretically, Doppler is practically zero if the direction of motion is tangential to the radar line-of-sight, one would expect a large decrease in classification performance for aspect angles close to 90 degrees. Such decrease can be seen in Figure 5, when the aspect angles are not zero, the most noticeable at 90 degrees. However, the performances remain relatively robust, showing the value of the proposed pipeline, which relies not only on Doppler to perform classification, but on the combination of spatial, temporal, and Doppler information together in the point cloud.

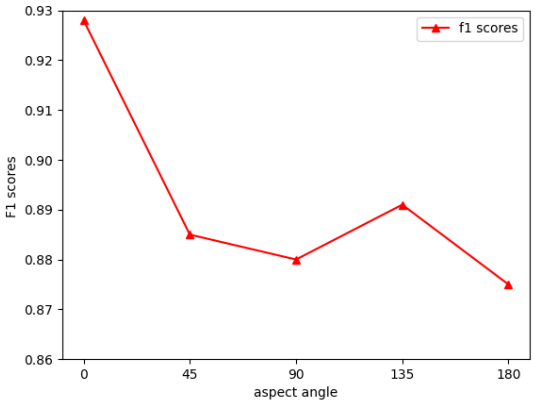

Fig. 5: Macro F1 scores for different aspect angles with Hengshuang's network and 3D spatial coordinates plus Doppler and time as input features.

D. Leave-One-Subject-Out Test

Each participant is expected to have specific kinematic patterns when performing certain activities. In order to train a generally applicable pipeline, it is important to test its generalization capabilities and see how the various kinematic patterns from each individual can affect the classification results. Figure 6 shows the results of a leave-one-subject-out test where data from each of the seven participants was used in turn as the test set. There is a drop of about 10% in the mean F1 score compared to the previous cross-validation results, confirming that more data and specific care are needed to analyse and, in case, enhance the performances for each individual.

V. Conclusion

This paper proposed a Human Activity Recognition (HAR) pipeline with high-dimensional radar point clouds as input to self-attention Point Transformer (PT) networks used as classifiers. The proposed pipeline achieves an F1 score of 92.8% for the problem of classifying 4 motions and 2 postures, bringing a +5.8% improvement compared with our previous work on the same dataset, that did not use attention-based

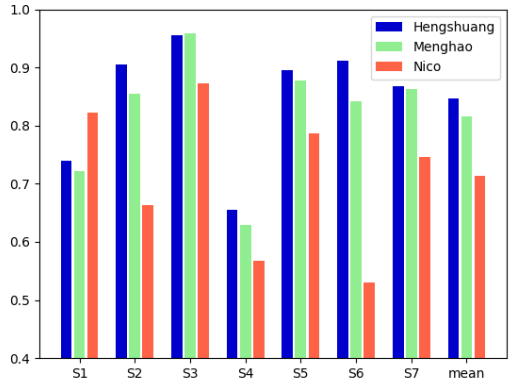

Fig. 6: Classification F1 score in leave-one-subject-out test when using 20 frames, where the horizontal axis represents the index of the left-out subject, and S denotes specific subjects. Here Hengshuang's network is used with 3D spatial coordinates plus Doppler and time as input features.

approaches [25]. The best feature combination for recognizing daily activities is the spatial 3D coordinates plus Doppler and time information. In addition, the proposed Adaptive Clutter Cancellation (ACC) method is proved to be a crucial contribution to the pipeline. It improves the accuracy of about +2% to +5%, depending on the different input features.

The comparison of the three PT networks shows that the PT from Hengshuang performs best: it obtains the highest classification F1 score while consuming the least training time. These results confirm that the self-attention mechanism matches well with the information encoded within radar point clouds. Further work will aim to apply this method to larger datasets, including sequences of unconstrained activities in free-form trajectories.

ACKNOWLEDGMENT

The authors are grateful to the volunteers who participated in the data collection, mainly the former student Y. Zhao for contributing to this work. Support from the Dutch Research Council NWO via the NWO-KLEIN project *RAD-ART* is also acknowledged.

REFERENCES

[1] "Who report ageing and health," https://www.who.int/news-room/fact-sheets/detail/ageing-and-health, accessed: 2020-08-23.

[2] S. A. Shah and F. Fioranelli, "Rf sensing technologies for assisted daily living in healthcare: A comprehensive review," *IEEE Aerospace and Electronic Systems Magazine*, vol. 34, no. 11, pp. 26–44, 2019.

[3] N. Lu, Y. Wu, L. Feng, and J. Song, "Deep learning for fall detection: Three-dimensional CNN combined with LSTM on video kinematic data," *IEEE journal of biomedical and health informatics*, vol. 23, no. 1, pp. 314–323, 2018.

[4] S. C. Mukhopadhyay, "Wearable sensors for human activity monitoring: A review," *IEEE sensors journal*, vol. 15, no. 3, pp. 1321–1330, 2014.

[5] T. R. Bennett, J. Wu, N. Kehtarnavaz, and R. Jafari, "Inertial measurement unit-based wearable computers for assisted living applications: A signal processing perspective," *IEEE Signal Processing Magazine*, vol. 33, no. 2, pp. 28–35, 2016.

[6] C. Debes, A. Merentitis, S. Sukhanov, M. Niessen, N. Frangiadakis, and A. Bauer, "Monitoring activities of daily living in smart homes: Understanding human behavior," *IEEE Signal Processing Magazine*, vol. 33, no. 2, pp. 81–94, 2016.

[7] Y. Lang, C. Hou, Y. Yang, D. Huang, and Y. He, "Convolutional neural network for human micro-Doppler classification," in *Proc. Eur. Microw. Conf.*, 2017, pp. 1–4.

[8] H. Sadreazami, M. Bolic, and S. Rajan, "On the use of ultra wideband radar and stacked lstm-rnn for at home fall detection," in *2018 IEEE Life Sciences Conference (LSC)*, 2018, pp. 255–258.

[9] X. Li, Y. He, Y. Yang, Y. Hong, and X. Jing, "Lstm based human activity classification on radar range profile," in *2019 IEEE International Conference on Computational Electromagnetics (ICCEM)*. IEEE, 2019, pp. 1–2.

[10] G. Park, V. K. Chandrasegar, and J. Koh, "Hand gesture recognition using deep learning method," in *2021 IEEE International Symposium on Antennas and Propagation and USNC-URSI Radio Science Meeting (APS/URSI)*. IEEE, 2021, pp. 1347–1348.

[11] L. Zheng, J. Bai, X. Zhu, L. Huang, C. Shan, Q. Wu, and L. Zhang, "Dynamic hand gesture recognition in in-vehicle environment based on fmcw radar and transformer," *Sensors*, vol. 21, no. 19, p. 6368, 2021.

[12] I. Alujaim, I. Park, and Y. Kim, "Human motion detection using planar array FMCW radar through 3d point clouds," in *2020 14th European Conference on Antennas and Propagation (EuCAP)*, 2020, pp. 1–3.

[13] A. D. Singh, S. S. Sandha, L. Garcia, and M. Srivastava, "Radhar: Human activity recognition from point clouds generated through a millimeter-wave radar," in *Proceedings of the 3rd ACM Workshop on Millimeter-wave Networks and Sensing Systems*, 2019, pp. 51–56.

[14] G. Lee and J. Kim, "Improving human activity recognition for sparse radar point clouds: A graph neural network model with pre-trained 3d human-joint coordinates," *Applied Sciences*, vol. 12, no. 4, 2022. [Online]. Available: https://www.mdpi.com/2076-3417/12/4/2168

[15] M. Wang, Y. D. Zhang, and G. Cui, "Human motion recognition exploiting radar with stacked recurrent neural network," *Digital Signal Processing*, vol. 87, pp. 125–131, 2019.

[16] H. Li, A. Mehul, J. Le Kernec, S. Z. Gurbuz, and F. Fioranelli, "Sequential human gait classification with distributed radar sensor fusion," *IEEE Sensors Journal*, vol. 21, no. 6, pp. 7590–7603, 2021.

[17] S. Wang, J. Song, J. Lien, I. Poupyrev, and O. Hilliges, "Interacting with soli: Exploring fine-grained dynamic gesture recognition in the radio-frequency spectrum," in *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*, ser. UIST '16. New York, NY, USA: Association for Computing Machinery, 2016, p. 851–860. [Online]. Available: https://doi.org/10.1145/2984511.2984565

[18] S. Zhu, R. G. Guendel, A. Yarovoy, and F. Fioranelli, "Continuous human activity recognition with distributed radar sensor networks and cnn–rnn architectures," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–15, 2022.

[19] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[20] Y. Zhao, R. G. Guendel, A. Yarovoy, and F. Fioranelli, "Distributed radar-based human activity recognition using vision transformer and cnns," in *2021 18th European Radar Conference (EuRAD)*, 2022, pp. 301–304.

[21] Z. Guo, "Msc thesis: *Point Transformer-Based Human Activity Recognition using high-dimensional Radar Point Clouds*," 2022. [Online]. Available: http://resolver.tudelft.nl/uuid:b3074b25-f0de-49e5-888c-5d8d03606501

[22] H. Zhao, L. Jiang, J. Jia, P. H. Torr, and V. Koltun, "Point transformer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 16 259–16 268.

[23] M.-H. Guo, J.-X. Cai, Z.-N. Liu, T.-J. Mu, R. R. Martin, and S.-M. Hu, "Pct: Point cloud transformer," *Computational Visual Media*, vol. 7, no. 2, pp. 187–199, 2021.

[24] N. Engel, V. Belagiannis, and K. Dietmayer, "Point transformer," *IEEE Access*, vol. 9, pp. 134 826–134 840, 2021.

[25] Y. Zhao, A. Yarovoy, and F. Fioranelli, "Angle-insensitive human motion and posture recognition based on 4d imaging radar and deep learning classifiers," *IEEE Sensors Journal*, vol. 22, no. 12, pp. 12 173–12 182, 2022.