Distributed Radar-based Human Activity Recognition using Vision Transformer and CNNs

Zhao, Yubin; Guendel, Ronny Gerhard ; Yarovoy, Alexander; Fioranelli, Francesco

DOI
10.23919/EuRAD50154.2022.9784575

Publication date
2022

Document Version
Final published version

Published in
Proceedings of the 18th European Radar Conference

Distributed Radar-based Human Activity Recognition using Vision Transformer and CNNs

Yubin Zhao, Ronny Gerhard Guendel, Alexander Yarovoy, Francesco Fioranelli
MS3 Group, Department of Microelectronics, TU Delft, The Netherlands
{Y.Zhao-31@student., R.Gundel@, A.Yarovoy@, F.Fioranelli@}tudelft.nl

***Abstract* — The feasibility of classifying human activities measured by a distributed ultra-wideband (UWB) radar system using Range-Doppler (RD) images as the input to classifiers is investigated. Kinematic characteristics of different human activities are expected to be captured in high-resolution range-Doppler images measured by UWB radars. To construct the dataset, 5 distributed monostatic Humatics P410 radars are used to record 15 participants performing 9 activities in arbitrary directions along a designated trajectory. For the first time a convolution-free neural network based on the novel multi-head attention mechanism (the Vision Transformer architecture) is adopted as the classifier, attaining an accuracy of $76.5\,\%$. A comparison between Vision Transformer and more conventional CNN-based architectures, such as ResNet and AlexNet, is also conducted. The robustness of Vision Transformer and the other networks against unseen participants is also validated by testing via Leave One Participant Out validation.**

***Keywords* — Activities of Daily Living, Deep Learning, Distributed Radar, Human Activity Recognition, Vision Transformer.**

I. Introduction

Human activity Recognition (HAR) is an important step to solve the healthcare challenges brought by the ageing population worldwide, as it enables contactless monitoring and timely warning and potentially delivery of key medical service. Numerous technologies have been proposed to address this problem, predominantly through visual aids [1] or wearable sensors [2]. Recent work, however, reveals that indoor radar is seen as a powerful solution thanks to the inherent advantages of functionality in any light condition, comfort for users, and respect of privacy [3] compared with alternative approaches. Initial approaches for radar-based HAR date back to the work proposed by Kim and Ling [4] in 2009, which used handcrafted features from human spectrograms and Support Vector Machine classifiers. Driven by the development in the field of ML (machine learning) and DL (deep learning), radar-based HAR has seen a very significant growth in interest and research work in recent years.

These can broadly be divided into three categories in terms of the input radar data representations and classifiers:

- Handcrafted features as a latent space-like input to supervised learning algorithms such as K-Nearest Neighbor [5] and Support Vector Machine [6].
- 2D radar data representations treated as an image-like input to Convolution Neural Networks (CNN) [7][8], or arranged as a video-like input processed by CNN-RNN (Recurrent Neural Network) [9].
- Representations that treat radar data as a temporal sequence of samples processed by RNNs, for example Bidirectional LSTM (Long-Short Term Memory) [10].

Based on literature, radar-based HAR can draw great inspiration and methodologies from the research in image (or video) classification, and even use images and videos as a suitable source of data to complement the (typically limited) radar datasets via transfer learning [11]. Hence, there is an interest in evaluating recent methods from the image processing and computer vision community for radar-based recognition problems.

An example of this can be the Vision Transformer (ViT) architecture, which is routinely used in natural language processing tasks and increasingly also in image processing. ViT [12] abandons the conventional convolutional layers used in CNN, and instead adopts attention mechanisms. Experiments on optical image classification tasks indicate that ViT can achieve state-of-art performance for mainstream classification datasets, such as ImageNet. Furthermore, ViT exhibits promising transferable capabilities, i.e. it can obtain excellent results when pre-trained at sufficient scale and transferred to tasks with fewer data points [12].

In this paper, initial results are presented on the investigation of ViT for radar-based HAR tasks. To the best of our knowledge, these are among the first results applying ViT to radar data of human activities and comparing the results to more conventional CNNs such as ResNet and Alexnet. While not outperforming the CNNs, ViT provides promising results in the order of $76\,\%$ on a 9-class problem.

The rest of this paper is organised as follows. Section-II presents the data collection and experimental dataset. Section-III briefly explains the working mechanism of ViT, and the experimental results are presented in Section-IV. Section-V finally concludes the paper.

II. Experimental Setup and Dataset Description

A distributed system consisting of 5 simultaneously-recording monostatic Humatics P410 pulsed radars is used, with the radars equally spaced on a semi-circle with radius $4.38\,\mathrm{m}$ (see [13]). The radars are placed with a height of approximately 1m above ground to ensure the total illumination of human body within the measuring space. The bandwidth of the radar is $2.2\,\mathrm{GHz}$, with Pulse Repetition Interval (PRI) of $8.2\,\mathrm{ms}$, providing an unambiguous Doppler interval of $\pm 122\,\mathrm{Hz}$.

To simulate Activities of Daily Living, nine different activities (i.e., classes) are included in the dataset, namely: 1) Walking, 2) Standing stationary, 3) Sitting down, 4) Standing up from sitting, 5) Bending from sitting, 6) Bending from standing, 7) Falling from walking, 8) Standing up after falling, and 9) Falling from standing while stationary. These activities were performed in arbitrary directions with respect to the line of sight of the radar sensors.

Additional details on the dataset used in this paper include:

- Overall 15 volunteers participated in the experiment (11 more than the dataset used for the initial results presented in [13]).
- *Training data* consist of 120s-long measurements where the participants performed a continuous, repeated combination of only two or three activities.
- *Testing data* are different 120s-long sequences where each participant performed a continuous, therefore more realistic, sequence of all 9 activities without repetitions.
- Each data segment used to generate Range-Doppler (RD) images has 200 slow-time bins (1.64s). The window to calculate the next RD is moved by 100 slow-time bins (0.82s). The duration and step of the window are chosen via empirical verification to maximise the classification accuracy.
- Overall, there are 143,285 RD images included in the dataset. Not surprisingly, the dataset is imbalanced with the major class (walking) being about 43.7 %, whereas the smallest class (falling from walking) only 0.7 %.

III. Vision Transformer Description

ViT was originally inspired by a breakthrough in natural language processing. In that work [14] the encoder-decoder structure was utilised to construct the so-called Transformer architecture, and more importantly, a multi-head attention mechanism was introduced to relate different positions of a sequence (shown in Fig. 1). The Multi-head Attention layer consists of multiple Dot-Product Attention layers running in parallel, where each Dot-Product Attention layer computes the relations between the input with itself and other positions, and the output probabilities of each Dot-product attention layer are added as the final output vector. Theoretically, in this way, the output/prediction of every single input might pay attention to multiple positions within the sequence. Moreover, it was experimentally proved that multi-head attention layer was better at capturing global dependencies between input and output [14].

ViT [12] inherits the encoder structure of the Transformer [14] to perform image classification. ViT is implemented by transforming one input image into multiple equal-size patches, flattening them into a sequence (analog to a sentence as in natural language processing), and feeding them into the hidden layers through a linear projection layer as shown in Fig. 2, where the structure of hidden layers is identical to Fig. 1.

The ViT used in this paper consists of 8 parallel Dot-product attention layers per multi-head attention layer; the input RD image size and patch size are 224x224 and 16x16, respectively. The rationale of using ViT for radar-based HAR lies in its capability to capture global spatial relationships within each RD image. The performances obtained by the ViT are also compared to AlexNet [15] and ResNet [16].

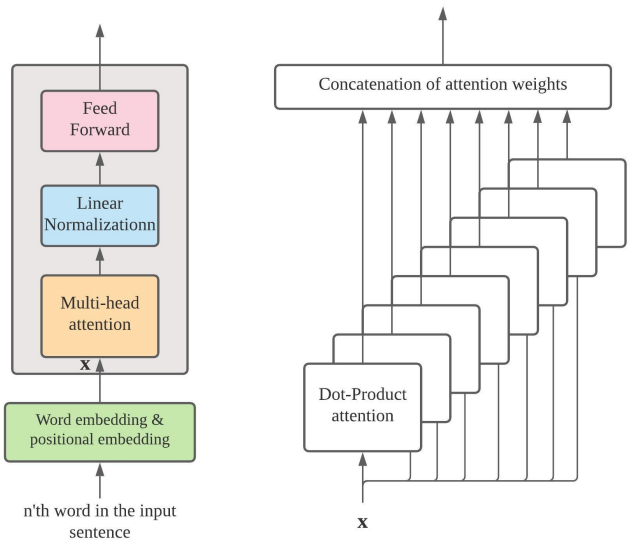

Fig. 1. Visualization of the encoder of Transformer architecture (left); and multi-head attention layer (right).

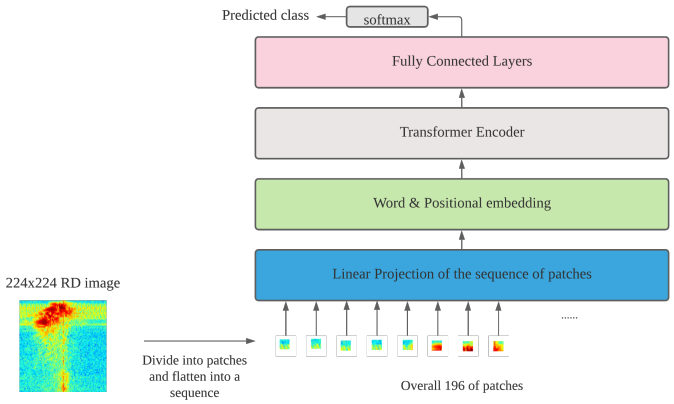

Fig. 2. Model overview of Vision Transformer, ViT

IV. Classification Approaches and Results

This section describes the classification approach and the main results obtained using ViT, ResNet and AlexNet.

A. Classification Approach and Results

All classifiers are trained for 50 epochs with an early stop function, and validated by 5-fold cross validation. Up-sampling the dataset is employed to tackle the problem of imbalanced dataset, as it effectively makes use of all collected data. Grid search on the hyperparameter learning rate is firstly applied as a means to optimize the classification performance. With the optimized learning rate selected, results with data from only 1 node out of 5 are compared with those obtained by combining data from all 5 radar nodes. Furthermore, results

Table 1. Validation results of ViT classifier in comparison with classic CNNs. Datasets include: *Dataset-1*: data from 15 participants and 5 radar nodes; *Dataset-2 to Dataset-6*: data from 15 participants from only node-1 to -5 respectively; *Dataset-7*: randomly selected 20 % of Dataset-1 (hence, the size of Dataset-2 to -7 are approximately the same). Networks include: *ViT*: random initialization; *ResNet-1*: random initialization; *ResNet-2*: pre-trained on ImageNet; *AlexNet-1*: random initialization; *AlexNet-2*: pre-trained on ImageNet. The optimized learning rates are $5 \cdot 10^{-4}$ for ViT, ResNet-1 and AlexNet-1, $5 \cdot 10^{-5}$ for ResNet-2, and $1 \cdot 10^{-4}$ for AlexNet-2.

Train Data	Learning Rate	ViT	ResNet-1	ResNet-2	AlexNet-1	AlexNet-2
Dataset-1	$5 \cdot 10^{-4}$	76.5 %	83.1 %	82.1 %	82.6 %	78.5 %
	$1 \cdot 10^{-4}$	75.7 %	77.1 %	83.7 %	80.1 %	80.3 %
	$5 \cdot 10^{-5}$	67.4 %	74.1 %	84.1 %	81.3 %	77.3 %
Dataset-2	optimized	72.5 %	76.2 %	71.5 %	71.3 %	70.0 %
Dataset-3	optimized	72.8 %	81.3 %	75.1 %	72.6 %	75.1 %
Dataset-4	optimized	74.9 %	67.7 %	78.2 %	77.1 %	76.4 %
Dataset-5	optimized	71.4 %	78.8 %	75.1 %	72.1 %	73.7 %
Dataset-6	optimized	71.1 %	79.2 %	75.7 %	73.8 %	74.4 %
Dataset-7	optimized	54.8 %	53.4 %	65.6 %	60.5 %	63.3 %

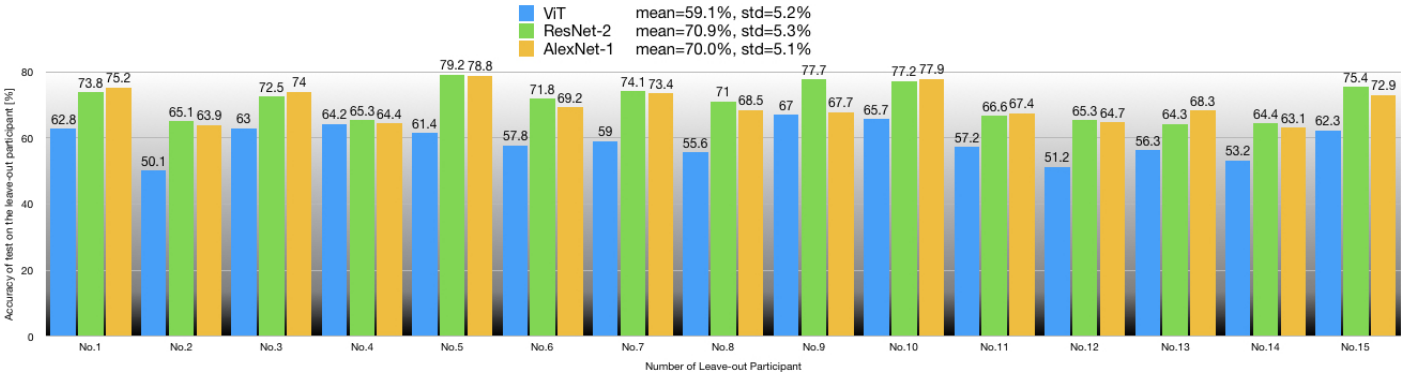

Fig. 3. Leave-one-participant-out test accuracy, evaluated using classifiers with optimized learning rate.

are also generated using Leave One Participant Out validation, i.e. training on 1st to 14th participants and testing on the 15th, and repeating this process for each participant followed by calculating the average and the standard deviation of the results. This is done to analyze the robustness of performances across different participants, and the similarity amongst their movements, as shown in Table 1.

1) *Learning Rate Hyperparameter*

Through the grid search on learning rates, it is established that this hyperparameter plays an important role in the classification performance despite what classifier is applied. Moreover, the optimal learning rate may vary from a network architecture to another as can be seen in Table 1.

2) *Data Fusion of Multiple Radars*

Data fusion is implicitly performed by using data from all the 5 radar nodes together as the input to the same network (here labelled as *Dataset-1* in Table 1). Using the optimized learning rate, the impact of data fusion is examined comparing the results with those obtained with only one single node, labelled as *Dataset 2-6*. From this comparison, it would appear that for a given classifier the combination of data from 5 nodes provides a boost in classification performance thanks to the larger amount of data used for training the network. To assess the effect of amount of data and its spatial diversity from different radar nodes, another dataset (named as *Dataset 7*) is generated by selecting 20 % of *Dataset 1*. Fig. 4 shows the accuracy for different networks as a function of the number of

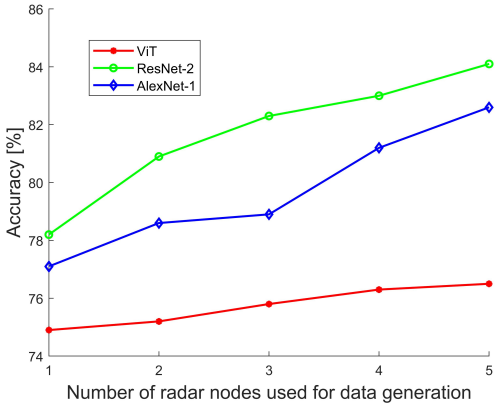

Fig. 4. Accuracy convergence of classifiers using sequential forward selection of radar nodes.

nodes providing data for training and validation. Increasing the number of nodes and the amount of data is beneficial for all networks. Conventional CNNs appear to outperform the ViT in this case.

3) *Transfer Learning from ImageNet*

Transfer learning from optical images, such as those from ImageNet, has been often used as a technique to improve performances for radar-based classification tasks where the size of the radar dataset is too small for proper training of the networks [11], [17]. Here pre-training is also used for the CNNs (*ResNet-2* and *AlexNet-2* in Table 1). However,

this yielded little improvement compared to the networks trained with radar data from scratch, which may be due to the lack of similarity between optical images and RD images. Transfer learning from optical data on the ViT did not yield improvements, but this aspect deserves further investigation (for example using speech data rather than images [18]).

4) Test via Leave One Participant Out

The results obtained with Leave One Participant Out validation are presented in Fig. 3 for each individual participant, along with the mean and standard deviation across participants. The results appear comparable across all the participants. In this case the CNNs provide higher results (approximately +10 %) compared to ViT, which appears to suggest better capabilities to generalise to unseen individuals.

5) Test on Continuous Activities

The classifiers yielding the best results on *Dataset-1* and with optimised learning rate (see blue shading in Table 1) are tested on the sequences of continuous activities from all 15 participants, collected as described in section II. The accuracy for ViT, ResNet-2 and AlexNet-1 are 45.3 %, 49.6 % and 47.3 %, respectively. This drop in accuracy is thought to be related to the difference in kinematic patterns, and hence in resulting RD plots, between the training sequences (containing repetitions of only 2-3 activities) and the testing ones (containing all activities in a more realistic fashion).

V. Conclusion

This paper has described initial classification results of using ViT classifier for HAR based on distributed radar data. An experimental dataset with 143,285 RD images collected from 15 participants and 5 UWB radar nodes was used to investigate the performance of ViT in comparison to more conventional CNNs such as ResNet and AlexNet.

These initial results show the significance of tuning key hyperparameters such as the learning rate, and the advantage of multiple radars' data fusion as the input to the networks. This "data fusion" capitalises on a larger amount of data to train the networks as well as exploit the spatial diversity of the same actions being seen by multiple radar simultaneously. Further tests on more realistic sequences of continuous activities showed a considerable drop of accuracy, implying that there is a gap in kinematic similarity between simpler measurements used as training sequences (with only 2-3 activities repeated several times) and realistic sequences of activities.

While ViT yielded initial results below those provided by conventional CNNs, directions for improvements can be considered for future work. These include getting to work transfer learning approaches with ViT, as well as an additional optimisation of the architecture of the ViT itself to fit the radar data and their specific characteristics.

Acknowledgment

The authors would like to thank the volunteers for the data collection, and NWO for partially funding this work under grant RAD-ART (Radar-aware Activity Recognition with Innovative Temporal Networks).

References

[1] 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, 2019.

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

[3] B. Çağlıyan and S. Z. Gürbüz, "Micro-Doppler-based human activity classification using the mote-scale bumblebee radar," *IEEE Geoscience and Remote Sensing Letters*, vol. 12, no. 10, pp. 2135–2139, 2015.

[4] Y. Kim and H. Ling, "Human activity classification based on micro-Doppler signatures using a support vector machine," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 47, no. 5, pp. 1328–1337, 2009.

[5] F. Fioranelli, M. Ritchie, and H. Griffiths, "Bistatic human micro-Doppler signatures for classification of indoor activities," in *2017 IEEE Radar Conference (RadarConf)*, 2017, pp. 0610–0615.

[6] M. G. Amin, Y. D. Zhang, F. Ahmad, and K. D. Ho, "Radar signal processing for elderly fall detection: The future for in-home monitoring," *IEEE Signal Processing Magazine*, vol. 33, no. 2, pp. 71–80, 2016.

[7] Y. Kim and T. Moon, "Human detection and activity classification based on micro-Doppler signatures using deep convolutional neural networks," *IEEE geoscience and remote sensing letters*, vol. 13, no. 1, pp. 8–12, 2015.

[8] B. Vandersmissen, N. Knudde, A. Jalalvand, I. Couckuyt, A. Bourdoux, W. De Neve, and T. Dhaene, "Indoor person identification using a low-power FMCW radar," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 7, pp. 3941–3952, 2018.

[9] 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*, 2016, pp. 851–860.

[10] H. Li, A. Shrestha, H. Heidari, J. Le Kernec, and F. Fioranelli, "Bi-LSTM network for multimodal continuous human activity recognition and fall detection," *IEEE Sensors Journal*, vol. 20, no. 3, pp. 1191–1201, 2020.

[11] M. S. Seyfioglu, B. Erol, S. Z. Gurbuz, and M. G. Amin, "DNN transfer learning from diversified micro-Doppler for motion classification," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 55, no. 5, pp. 2164–2180, 2019.

[12] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," 2021.

[13] R. G. Guendel, M. Unterhorst, E. Gambi, F. Fioranelli, and A. Yarovoy, "Continuous human activity recognition for arbitrary directions with distributed radars," in *2021 IEEE Radar Conference (RadarConf21)*, 2021, pp. 1–6.

[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2017.

[15] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, pp. 1097–1105, 2012.

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015.

[17] E. Cippitelli, F. Fioranelli, E. Gambi, and S. Spinsante, "Radar and RGB-depth sensors for fall detection: A review," *IEEE Sensors Journal*, vol. 17, no. 12, pp. 3585–3604, 2017.

[18] Y. Li, K. He, D. Xu, and D. Luo, "A transfer learning method using speech data as the source domain for micro-Doppler classification tasks," *Knowledge-Based Systems*, vol. 209, p. 106449, 2020.