Continuous People Crowd Monitoring defined as a Regression Problem using Radar Networks

Guendel, Ronny; Ullmann, Ingrid ; Fioranelli, Francesco ; Yarovoy, Alexander

DOI
10.23919/EuRAD58043.2023.10289622

Publication date
2023

Document Version
Final published version

Published in
Proceedings of the 2023 20th European Radar Conference (EuRAD)

Continuous People Crowd Monitoring defined as a Regression Problem using Radar Networks

Ronny G. Guendel[#], Ingrid Ullmann[*], Francesco Fioranelli[#], Alexander Yarovoy[#]

[#]Microwave Sensing, Signals and Systems (MS3) Group, Delft University of Technology, The Netherlands

[*]Institute of Microwaves and Photonics, Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany

{r.guendel, f.fioranelli, a.yarovoy}@tudelft.nl, ingrid.ullmann@fau.de

***Abstract* — Radar-based human activity recognition in crowded environments using regression approaches is addressed. Whereas previous research has focused on single activities and subjects, the problem of continuous activity recognition involving up to five individuals moving in arbitrary directions in an indoor area is introduced. To treat the problem, a regression-based approach is used, which offers innovative insights into creating robust and accurate systems for monitoring human activities.**

Novel approaches utilizing LSTM or CNN regression techniques with Linear Regression and Support Vector Machine regressor are compared on extracted features from radar data through the Histogram of Oriented Gradients and Principal Component Analysis. These approaches are rigorously evaluated by a Leave-One-Group-Out method, with performance assessed using common regression metrics such as the RMSE. The most promising outcomes were observed for crowds of three and five individuals, with respective RMSE of approximately 0.4 and 0.6. These results were primarily achieved by utilizing the micro-Doppler (μD) Spectrogram or range-Doppler data domain.

***Keywords* — Radar Signal Processing, Multiple People Monitoring, Distributed Radar, Machine Learning, Deep Learning, Histogram of Oriented Gradients, Principal Component Analysis, Regression, LSTM, CNN.**

I. Introduction

Human Activity Recognition (HAR) has emerged as a crucial research area, not only for enabling vulnerable individuals to maintain an independent lifestyle, but also for ensuring safety in self-determined living environments. A range of technologies, from contactless sensors such as radio frequency (RF) based products to wearable sensors in the form of smartwatches and other devices, have shown the capability to measure various vital metrics, including location, pulse rate, body temperature, blood pressure, and motion characteristics [1]. However, when it comes to monitoring multiple individuals simultaneously, wearable sensors and contactless video-based approaches such as cameras and lidar sensors have limitations in terms of usability and privacy concerns. As a result, radar has arisen as a promising alternative due to its ability to overcome these restrictions [2].

The treatment of crowd monitoring as a discretized classification problem has been discussed in the literature, and Bendali-Braham et al. [3] have provided further insight into the complexities of crowd monitoring. For instance, a slight deviation in the classifier's prediction may result in significant classification accuracy errors, as illustrated in the case where the ground truth provides 10 walking individuals, but a classifier predicts only 9. In such cases, the accuracy becomes 0 %, similar to the accuracy obtained if the classifier predicts no walking. Furthermore, this issue is compounded by the possibility of people starting or stopping walking within the sliding window used for classification, leading to similar errors. Therefore, to address these limitations, this study aims to treat the problem of predicting the number of walking people as a regression problem rather than a classification problem. The contributions can be summarized as follows:

- A regression problem was defined to predict the number of people walking in the scene instead of a more conventional discretized classification problem.
- A variety of regressors including deep learning methods, such as the Long-Short Term Memory (LSTM) network was applied on features extracted from continuous radar recordings, and their results were evaluated using relevant metrics such as the Root Mean Squared Error (RMSE). Other regressors included a Convolutional Neural Network (CNN) operating directly on the image domain.
- The proposed approach was validated with data collected with a radar network of five nodes, synthetically combining the signatures of up to five people walking and stopping to simulate crowd movements in an indoor area.

The rest of this paper is organized as follows. Section II presents the data collection, the dataset, and the proposed regression approach, with experimental results presented in Section III. Finally, Section IV concludes the paper.

II. Proposed Methodology

A. Radar Data and Domains

The publicly available data [4] contains single human subjects; hence, multiple recordings of different individuals were coherently fused to generate synthetic data with multiple people, such that: $\bar{R} = \frac{1}{K}\sum_{i=1}^{K} R_i$, with K the maximum number of people, R a complex range-Time (RT) sequence and $\bar{R}$ the fused result. The generated output in logarithmic scale can be visually examined in the RT maps on the left side of Figure 1, with five subjects performing walking and standing in an unconstrained trajectory. Further information on the layout of the radars is provided in [5].

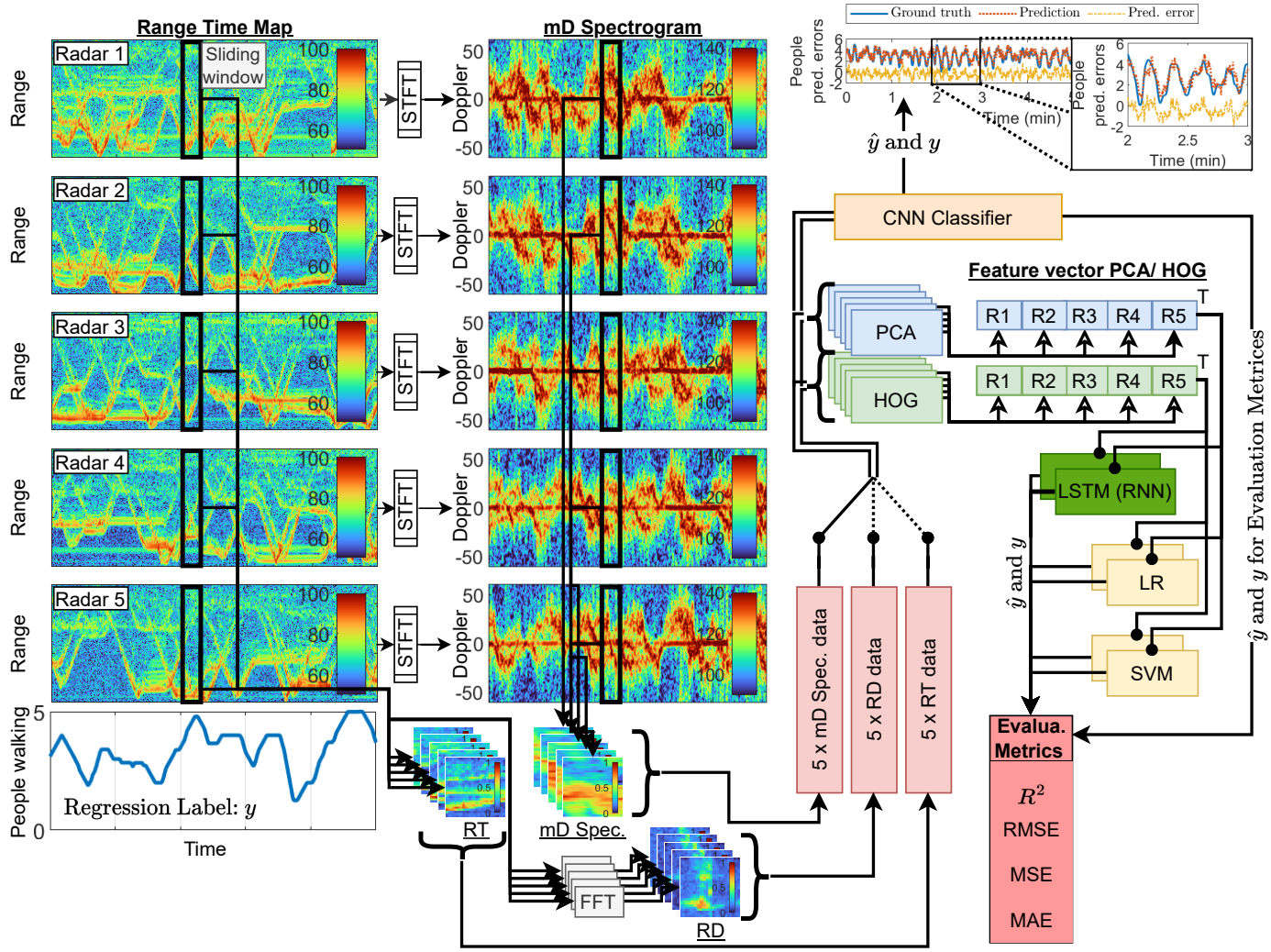


Fig. 1. The flowchart of the proposed approach shows the synthetically merged data providing the RT maps for a group of five people recorded with a radar network consisting of five nodes. In terms of signal processing, three data domains are extracted, namely the RT–, RD–, and µD Spectrogram domain, followed by the feature extraction chains of PCA and HOG, and the ML/DL-based regressors. Lastly, a *Leave-One-Group-Out* (LOGO) test is shown with its prediction error plot (top-right corner) and the four applied evaluation metrics (bottom-right corner).

Three radar data domains of range-Time (RT)–, range-Doppler (RD) map, and micro-Doppler (µD) Spectrogram, are obtained using a sliding window on continuously recorded data, as illustrated in Figure 1. As seen, the network consists of five radar nodes, enabling unconstrained Human Activity Recognition (HAR). A sliding window of 1 sec with an hop-size of 0.25 sec is applied. Furthermore, the row-wise Fast Fourier Transform (FFT) was used over the same window to obtain the RD map.

Before using the same sliding window approach over the µD Spectrogram, the Short-Time Fourier Transform (STFT) was applied on the RT signals, with a window size and hop-size of 64 and 63 samples, respectively. Finally, downsampling to 28x28 pixels was applied on all gathered domains of five nodes, with a few examples shown to the bottom of Figure 1.

B. *Feature Extraction*

On the generated images for the three data domains, Principal Component Analysis (PCA) was applied, selecting the five strongest principal components associated with the five strongest singular values; this resulted in a feature vector size of 140 samples for the given input images of 28x28 pixels. Similarly, as a comparative method to PCA, the Histogram of Oriented Gradients (HOG) is used with a cellsize of 8x8, a blocksize of 2x2, 9 bins, and 50% block overlap, resulting in a feature vector size of 144 samples.

Before forwarding the extracted features to the regression algorithm, feature fusion was employed to combine the information seen by all five radar nodes resulting in a total feature vector length of 700 and 720 for PCA and HOG, respectively.

C. Regression Approaches

This section describes the conventional regressors and the deep learning networks used in this research.

1) Conventional Regressors

Several regression methods were tested with the best compromise between computational load and performance provided by the Linear Regression (LR) and the Support Vector Machine (SVM) regression. These two regressors are used throughout this study on the concatenated fused features described in Section II-B.

2) Deep Learning Regressors

Furthermore, the following two deep learning based regressors were used. First, a modified Convolutional Neural Network (CNN) for regression, as proposed in [6], which operates on image tensors with the dimension 28x28x5 directly, and no feature fusion via PCA or HOG required. Then, the Recurrent Neural Network (RNN), proposed in [7] with its version of the Long Short-Term Memory (LSTM) network was modified to fulfil regression problems, simply by changing its last layers (*Softmax Layer*, the *Classification Layer*) to a *Regression Layer*. Furthermore, we decreased the network's depth to 400 hidden units instead of the originally proposed 1500.

The ADAM optimizer was used to train both DL networks with 50 epochs and an initial learning rate of 10^{-3}. It should be noted that additional hyperparameter– and network tuning may further improve performances, but this is left for future work beyond the scope of this paper.

D. Evaluation Metrics

The main evaluation metric throughout this work is the Root Mean Squared Error (RMSE), by default the most popular metric when evaluating regression problems and defined as: $\epsilon_{RMSE}=\sqrt{\frac{1}{m}\sum_{i=1}^{m}(\hat{y}_i - y_i)^2}$ [8], with $\hat{y}_i$, y_i the prediction and the ground truth, respectively, and m the samples in the LOGO (*Leave-One-Group-Out*) test set. The results are reported in Table 1 and in Figure 2, with additional metrics such as the Mean Squared Error Mean (MSE), the Mean Absolute Error (MAE), and the R^2 score.

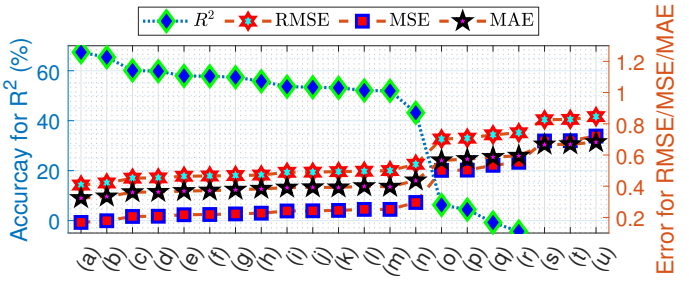
III. Experimental Results

The reported results considered DL models, namely a CNN and a LSTM, with the latter applied to the feature domains extracted from PCA and HOG, respectively. Similarly, those features were tested by the following conventional regressors, the LR and the SVM regression model. The CNN regression model, by nature an image processing regression model [6], was modified for input data size 28x28x5, where five is the number of radars in the network, and thus no prior feature extraction is required.

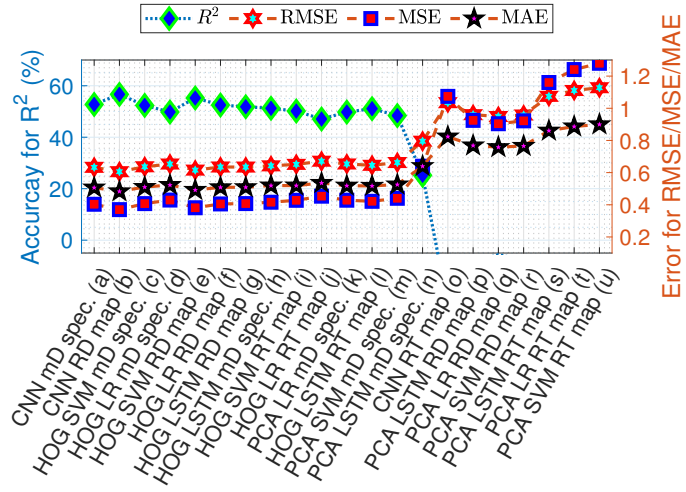
An example of the attained performance can be seen in Figure 1 (top-right corner) by comparing the *Ground Truth* (blue) with the *Prediction* curve (brown), and the *Prediction error* curve (yellow), here demonstrated for a group of five people.

Table 1. Leave three and five people out test results, known as *Leave-One-Group-Out* (LOGO), using a CNN, LSTM, and conventional regressors applied on the µD Spectrogram, the RD map, and the RT map, respectively. The RMSE results using a CNN are also shown in Figure 2.

Regression model	**Group size** Classif./Ev. Metrics	**3 People** RMSE	**5 People** RMSE
DL	CNN µD spec.	0.408	0.633
DL	CNN RT map	0.703	1.035
DL	CNN RD map	0.421	0.606
DL	HOG LSTM µD spec.	0.471	0.642
DL	HOG LSTM RT map	0.496	0.645
DL	HOG LSTM RD map	0.469	0.634
Conventional	HOG LR µD spec.	0.453	0.653
Conventional	HOG SVM µD spec.	0.452	0.636
Conventional	HOG LR RT map	0.490	0.671
Conventional	HOG SVM RT map	0.489	0.651
Conventional	HOG LR RD map	0.466	0.635
Conventional	HOG SVM RD map	0.464	0.616
DL	PCA LSTM µD spec.	0.541	0.795
DL	PCA LSTM RT map	0.826	1.074
DL	PCA LSTM RD map	0.708	0.960
Conventional	PCA LR µD spec.	0.494	0.654
Conventional	PCA SVM µD spec.	0.500	0.663
Conventional	PCA LR RT map	0.828	1.111
Conventional	PCA SVM RT map	0.846	1.128
Conventional	PCA LR RD map	0.729	0.949
Conventional	PCA SVM RD map	0.742	0.958



(a) Group of people is 3



(b) Group of people is 5

Fig. 2. Regression results shown for different regression models and for three– and five-people groups in (a) and (b), respectively. The considered evaluation metrics are the R^2 score, RMSE, MSE, and MAE.

A. Group of 3 People

The associated results for a group of three people in the scene are shown in Figure 2a with ordered performance declining from left to right, as well as in the third column of Table 1. The best overall performance for *Leave-One-Group-Out (LOGO)* tests was achieved using the CNN regressor applied on the μD Spectrogram with an RMSE of 0.4, closely followed by the CNN architecture operating on RD maps. Fusing the μD Spectrogram and RD map was not able to boost the results further. Then conventional regression models, such as the LR and the SVM, follow, with an RMSE in the order of 0.45. PCA feature extraction, applied on the μD Spectrogram with five principal component vectors, also achieves satisfactory results with two times 0.50 and 0.54 for LR, SVM, and LSTM regression models, respectively.

Regardless of the chosen regression model, a drastic performance drop can be observed using the RT map in combination with features extracted from the PCA.

B. Group of 5 People

The results for the five people group case are shown in Figure 2b, as well as in the fourth column of Table 1. A similar order of regression performances for different models can be observed, as the one in Section III-A. Higher performance is mainly provided by the μD Spectrogram and RD domain, i.e., the best result with an RMSE of 0.6 for the RD map and the CNN regressor, followed again by the LSTM, SVM, and LR models with features extracted by the HOG. Similarly, the RT map in conjunction with PCA provides poor results.

C. Discussion

The poor regression performance using PCA-based features from 28x28 images of the RT map may be due to its rotation-invariant feature extraction. For RT maps, the feature vector may not convey information about the slope of the dominant signature from a walking human, and the RT domain may not provide direct crucial velocity information, as compared to the μD Spectrogram or RD domain. Similarly, the CNN applied to the RT map also provided poor performance, possibly due to similar concerns. It is also notable that, although the HOG feature descriptor is not inherently rotation-invariant, it has shown superior performance in detecting the slopes of people within an RT map [9]. Finally, while such a small image size is perhaps also not favourable, this choice was made to limit the computational runtime and burden for the regression models.

Therefore, these initial results appear to suggest that it is crucial to either consider domains that include the velocity information or a suitable feature extraction method, such as the HOG, must be used. Although the LSTM does not compete performance-wise with the CNN, it should be noted that the LSTM has much more freedom for hyperparameter and network tuning, and with finer tuning, it may outperform other models. However, it is worth noting that the LSTM and conventional regressors (LR, SVM) operate on the HOG or PCA feature vectors, whereas the CNN operates directly on the sample images, making a fair comparison arguable.

IV. Conclusion

The paper's aim is the estimation of crowd activities using distributed radars addressed as a regression problem to forecast the ratio of walking vs standing individuals in arbitrary directions. Traditional Linear Regression (LR) and Support Vector Machine (SVM) regressor models, as well as deep learning regressors were used. Experimental recordings with 15 single subjects were recorded with synthetic data of groups of three and five people created. *Leave-One-Group-Out* sets evaluate the performance of the trained regressor on unknown data. Three different radar data domains were tested, and two classes of features were extracted from each domain. The CNN regressor applied on the micro-Doppler Spectrogram achieved superior performance with an RMSE of 0.4 for crowds of three people. Similarly, for five people, the range-Doppler map delivered an RMSE of 0.6, with the CNN operating directly on the image domain. On the other hand, the Histogram of Oriented Gradient features provided close performance, with the LSTM, LR, or SVM regressor producing results of approximately 0.45 (crowds of three) and 0.65 (crowds of five). In contrast, PCA often fails to compete with its aforementioned counterparts, especially on range-time data.

Acknowledgment

This work was partly funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - SFB 1483 - Project-ID 442419336, EmpkinS.

References

[1] 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.

[2] J. Liu, H. Liu, Y. Chen, Y. Wang, and C. Wang, "Wireless sensing for human activity: A survey," *IEEE Communications Surveys and Tutorials*, vol. 22, no. 3, pp. 1629–1645, 2020.

[3] M. Bendali-Braham, J. Weber, G. Forestier, L. Idoumghar, and P.-A. Muller, "Recent trends in crowd analysis: A review," *Machine Learning with Applications*, vol. 4, p. 100023, 2021.

[4] R. G. Guendel, M. Unterhorst, F. Fioranelli, and A. Yarovoy. (2021, Nov) Dataset of continuous human activities performed in arbitrary directions collected with a distributed radar network of five nodes. [Online]. Available: http://dx.doi.org/10.4121/16691500.v3

[5] R. G. Guendel, F. Fioranelli, and A. Yarovoy, "Distributed radar fusion and recurrent networks for classification of continuous human activities," *IET Radar, Sonar and Navigation*, vol. 16, no. 7, pp. 1144–1161, 2022.

[6] The MathWorks, Inc. (2021, Oct) Train convolutional neural network for regression. [Online]. Available: https://www.mathworks.com/help/deeplearning/ug/train-a-convolutional-neural-network-for-regression.html

[7] K. Itakura. (2021, Oct) Video classification using lstm with matlab. [Online]. Available: https://github.com/KentaItakura

[8] A. Géron and a. O. M. C. Safari, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, 2nd Edition*. O'Reilly Media, Incorporated, 2019.

[9] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1, 2005, pp. 886–893.