Grouped Target Tracking and Seamless People Counting With a 24-GHz MIMO FMCW

Wang, D.; Yuan, S.; Yarovoy, A.; Fioranelli, F.

DOI
10.1109/TRS.2025.3609436

Publication date
2025

Document Version
Final published version

Published in
IEEE Transactions on Radar Systems

Grouped Target Tracking and Seamless People Counting With a 24-GHz MIMO FMCW

Dingyang Wang, *Member, IEEE*, Sen Yuan, *Member, IEEE*, Alexander Yarovoy, *Fellow, IEEE*, and Francesco Fioranelli, *Senior Member, IEEE*

***Abstract*—The problem of radar-based tracking of groups of people moving together and counting their numbers in indoor environments is considered here. A novel processing pipeline to track groups of people moving together and count their numbers is proposed and validated. The pipeline is specifically designed to deal with frequent changes of direction and stop-and-go movements typical of indoor activities. The proposed approach combines a tracker with a classifier to count the number of grouped people; this uses both spatial features extracted from range-azimuth (RA) maps and Doppler frequency features extracted with wavelet decomposition. Thus, the pipeline outputs over time both the location and the number of people present. The proposed approach is verified with experimental data collected with a 24-GHz frequency-modulated continuous-wave (FMCW) radar. It is shown that the proposed method achieves 93.15% accuracy in terms of counting the number of people and a tracking metric optimal subpattern assignment (OSPA) of 0.335. Furthermore, the performance is analyzed as a function of different relevant variables such as feature combinations and scenarios.**

***Index Terms*—Frequency-modulated continuous-wave (FMCW) radar, grouped target, human counting, human monitoring, multitarget tracking.**

I. Introduction

RADAR-BASED human tracking is a very active research field, leveraging on the advantages of using radar sensors. First, they operate contactless and do not require additional equipment to be attached or worn by the users. Second, they are expected to be more respectful of personal privacy than vision-based sensors and are insensitive to ambient light conditions or glaring. In combination with compact multiple-input–multiple-output (MIMO) frequency-modulated continuous-wave (FMCW) systems, these features make radar a very attractive sensor for indoor observation and tracking of humans.

In the literature, there are different approaches for indoor human tracking leveraging on multiple domain information obtainable by radar sensors [1]. These are typically based on the range-azimuth (RA) domain [2] or the range-Doppler (RD) domain [3], [4], [5], as the starting points for detecting the presence of targets. Two processing pipelines starting from detections on the RA versus RD maps were formulated and compared in our previous work in [6]. Furthermore, an additional comparison on the processing steps of data association and tracking methods was presented in [7]. The optimal subpattern assignment metric (OSPA) metric was used in that work for performance assessment.

While promising, these initial results showed that tracking and counting the number of people moving as a group remained challenging. A "grouped target" can be defined as a cluster of people moving together while remaining close to each other, for example, shoulder by shoulder or following each other. This behavior is very common in daily life, e.g., a couple of colleagues walking together in a corridor to the office or a family with a kid crossing a road section. From the point of view of radar, the challenge comes from the insufficient spatial resolution to separate people in the group and the mixing of their (micro)-Doppler signatures. Additionally, in indoor environments, there are very frequent changes of direction and velocity (e.g., sort of stop-and-go movement pattern) due to the limited physical space for people to move and occlusion by furniture. This makes the Doppler signatures of grouped people even more confusing to analyze to estimate how many people are present in the group.

In order to address the above challenge, in this work, the problem is formulated as "grouped target" tracking instead of tracking individual subjects. Simultaneously, the proposed processing pipeline combines a classifier to count the number of people present in each tracked group. Unlike the previous initial work in [6] and [7], the proposed method takes into account the grouped nature of the targets. Additionally, unlike the work in [5] where performances were tested in outdoor scenarios with relatively low clutter and wide spaces between participants or groups, thus generating little multipath, here a cluttered indoor space is considered. This generates additional multipath components from walls and furniture and constrains spaces for the people to move, leading to frequent acceleration/deceleration as well as changes of direction. To deal with the more challenging radar signatures resulting from these phenomena, features based on wavelet decomposition in the Doppler domain are formulated in this work and used to count the number of people. These outperform the cadence

Received 26 March 2025; revised 21 July 2025 and 8 September 2025; accepted 9 September 2025. Date of publication 12 September 2025; date of current version 25 September 2025. This work was supported in part by the Huawei Sweden Gothenburg Research Center. *(Corresponding author: Sen Yuan.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by TU Delft HREC.

The authors are with the Microwave Sensing Signals and Systems (MS3) Group, Department of Microelectronics, Delft University of Technology, 2628 CD Delft, The Netherlands (e-mail: D.Wang-6@tudelft.nl; S.Yuan-3@tudelft.nl; A.Yarovoy@tudelft.nl; F.Fioranelli@tudelft.nl).

Digital Object Identifier 10.1109/TRS.2025.3609436

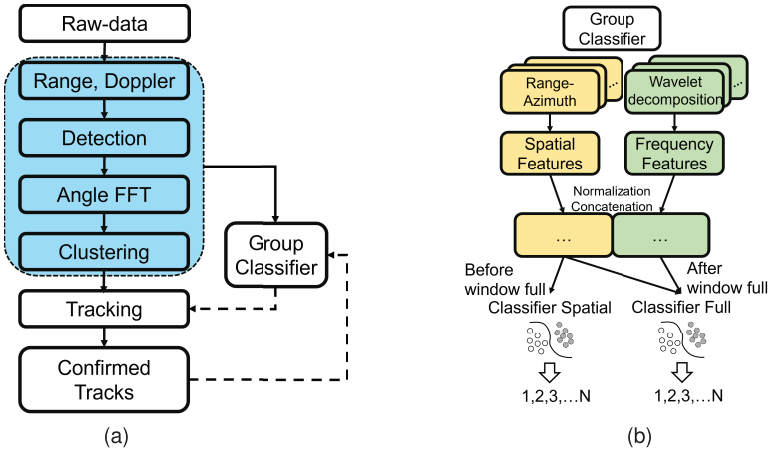


Fig. 1. Pipeline of the proposed approach for grouped target tracking with integrated classifier. (a) Tracking pipeline. (b) Classifier pipeline.

velocity diagram (CVD)-based features used in [5], which, being based on simple Fourier transform, are less suitable to deal with less stationary Doppler signatures as a result of acceleration/deceleration. In general, it is shown that the accuracy of the proposed approach improved by 13.95% compared to that method.

Summarizing, the main contributions of this work are as follows.

1) The problem of tracking and counting grouped targets in an indoor environment is addressed. A dedicated processing pipeline is proposed to utilize the information from an MIMO FMCW radar, combining tracking and classification.
2) The pipeline specifically combines tracking with extended Kalman filter and a classifier to estimate the number of people in the group area. The classifier is trained with spatial features from RA maps and Doppler frequency features derived from wavelet decomposition.
3) The proposed method is experimentally tested with a 24-GHz MIMO FMCW radar in a laboratory room, which is an environment with heavy clutter and multipath. The proposed method achieves 93.15% accuracy in counting the number of people and shows an average error of 0.335 OSPA.

The rest of this article is organized as follows. In Section II, the proposed pipeline for simultaneously tracking and counting people is presented, along with the utilized features. In Section III, the experimental setup and details of the data collection are discussed. The performance analysis and the effect of the different variables are provided in Section IV. Finally, conclusions are given in Section V.

II. Proposed Method

In this section, the processing pipeline proposed is presented in two parts as shown in Fig. 1. The first part on the left-hand side is about the tracking method based on detections from the RD maps. On the right-hand side, the seamless group classifier for counting the number of people is introduced in the second part of the pipeline. In order to get enough data to extract Doppler frequency features, an observation time of suitable duration is required. Before this time window is complete, the classifier only uses spatial features based on RA maps for initial counting. Additional details for the two parts of the pipeline are presented in Section II.

A. Tracking

The left part of Fig. 1 shows essentially a conventional tracking pipeline. Detections from the RD maps are used as the starting point of the processing, leveraging on the relatively finer Doppler resolution compared to the angular one. More details on performance differences in using detections from RD maps rather than RA maps were discussed in our previous work [6]. After detection, fast Fourier transform (FFT) is performed to estimate azimuth angles of the detected targets and radar cubes are obtained. Then, a clustering method is applied; specifically, the density-based spatial clustering of applications with noise (DBSCAN) algorithm is used in this work [8]. DBSCAN is particularly well-suited for extended targets due to its ability to handle irregularly shaped clusters and automatically identify outlier points (e.g., noise) within the data. This step can help differentiate between multiple targets and reduce false alarms when dealing with point clouds. If multiple targets are close to each other and within the resolution cell, there will be one cluster according to the DBSCAN hyperparameters. The center point of the cluster will represent the cluster position for further tracking. The detailed parameters of the FFT and DBSCAN algorithms are listed in Table I.

Following clustering, the data association step is used to find an optimal match between new detections and current tracks. The process steps are generally based on the extended Kalman filter framework [9] and denoted by the "Tracking" block in Fig. 1. We previously investigated three different data association algorithms [7], which led to the selection of the global nearest neighbor (GNN) method for this study [10]. The GNN method considers all possible pairings between new detections and current tracks. The weight is calculated according to a distance-based cost function. The most likely association is made from the weight. Compared with the Joint Probabilities Data Association (JPDA) [11], GNN provides a

TABLE I
JOBY (FORMER INRAS) FMCW RADAR AND ALGORITHM PARAMETERS

FMCW radar model	RadarBook2 (RBK2)
Operating frequency	24 GHz
Sweep bandwidth	250 MHz
ADC sampling rate	120 ksps
ADC samples per chirp	56
Up chirp duration	467 μs
Chirp repetition interval	483 μs
Number of chirps in a frame	90
Slow-time sampling frequency	10 Hz
Number of TX & RX channels	2 x 8
Antenna horizontal 3 dB beamwidth	76.5°
FFT size (Range, Doppler, Azimuth)	256,256,256
DBSCAN (ε, minimum points)	0.7, 5

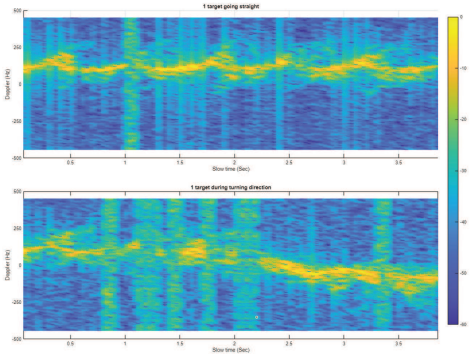


Fig. 2. Comparison of spectrograms between the case of "normal" walking in one direction and walking with direction changing.

hard rather than a soft assignment, which is able to directly find the detection contribution to tracking.

B. Feature Extraction

In the previous study, features were extracted from the RA maps and CVD maps [5]. The CVD $S_C(\epsilon, k)$ is obtained by taking the FFT of the Doppler spectrum across the time axis as

$$S_C(\epsilon,k) = \sum_{l=0}^{N_w-1} \left|\hat{S}(l,k)\right| w(l) \exp\left(-j2\pi \frac{\epsilon l}{N_w}\right) \tag{1}$$

where $S(l,k)$ is the spectrogram, $w(l)$ is a window function with length equal to N_h, and N_w is the total number of windows when performing the FFT. $\epsilon = 0, 1, 2, \ldots, N_{h-1}$ and $k = 0, 1, 2, \ldots, N_{w-1}$ denote the cadence frequency index and Doppler index, respectively. In order to estimate the proper frequency from the CVD map with sufficiently fine resolution, a longer observation window is required. If the observation of people is performed in an outdoor environment, the long observation window may not be a problem as typical people move along longer and more regular trajectories. However, when considering an indoor environment with walls and furniture in normal-sized rooms, due to the limited moving space, people can change their direction or stop and go again within the ideal observation window time. As shown in Fig. 2, the Doppler pattern changes completely due to the direction turning within the observation window. Therefore, in these cases, frequency features directly extracted from the CVD may not be reliable.

To extract reliable frequency features related to the Doppler domain (and therefore target velocity), wavelet-based methods are a good approach to decompose the original signal into different levels. Furthermore, it is important to see how each decomposition level can contribute to the final accuracy so that levels containing little information for the problem at hand can be discarded and not be used by the classifiers. In this work, the Maximal overlap discrete wavelet transform (MODWT) is used for this purpose [12], [13], [14]. First, the time-series data to be analyzed is denoted as $\mathbf{X} = \{X_t\}$ where $t = 0, \ldots, T-1$, in which T is the total number of frames for the observation window, and X_t represents the range bin values corresponding to the target's location at frame t. Note that X_t is a vector, as each frame consists of multiple chirps. The jth level wavelet and scaling filter are denoted as $\{\tilde{h}_{j,l}\}$ and $\{\tilde{g}_{j,l}\}$, respectively. l is the index of filter coefficients. The scaling and wavelet coefficient can then be expressed as follows:

$$\tilde{V}_{j,t} = \sum_{l=0}^{T-1} \tilde{g}_{j,l} X_{t-l \bmod T} \tag{2}$$

$$\tilde{W}_{j,t} = \sum_{l=0}^{T-1} \tilde{h}_{j,l} X_{t-l \bmod T} \tag{3}$$

where $j = 1, 2, \ldots, J$ is the level of the wavelet decomposition, in this work specifically $J = 4$.

Compared to the normal wavelet method mentioned in [14], the MODWT approach avoids downsampling and prevents the data size to be too low with the increasing number of levels. Additionally, the MODWT preserves time information alignment and provides multiresolution frequency analysis. For statistical feature extraction and subsequent usage into classifiers, having the same data size for each level is important for consistency.

In order to implement this method and extract the relevant MODWT features, a tracking trajectory at time t and tracking ID is needed and denoted as $\text{Trac}_t^{\text{ID}}$. For a given ID, the range bin data X_t^{ID} and angle information are stored in a buffer to extract the features. The spatial features are those extracted from the RA map and are specifically the width occupied by the target in the azimuth axis, the length occupied in the range axis, the mean value of angle bins, the median value of angle bins, the variance of angle bins, the number of different angle bins occupied, the mean value of angle profile, the median value of angle profile, the variance of angle profile, and the number of pixels occupied in the RA map. The spatial features are extracted from each window frame and averaged along the time axis to smooth their values. The frequency features are extracted from the four levels of MODWT. There are eight statistical features for each level, including the variance, standard deviation, mean value, median value, root-mean-square value, skewness value, kurtosis value, and entropy value.

An example of MODWT decomposition is shown in Fig. 3 with its time axis aligned to the upper subplot of Fig. 2. In the case of missed detections while performing the tracking process, there will not be related data stored in the buffer for a few time bins. In these cases, the mean value of the remaining existing data will be used to fill the missing samples

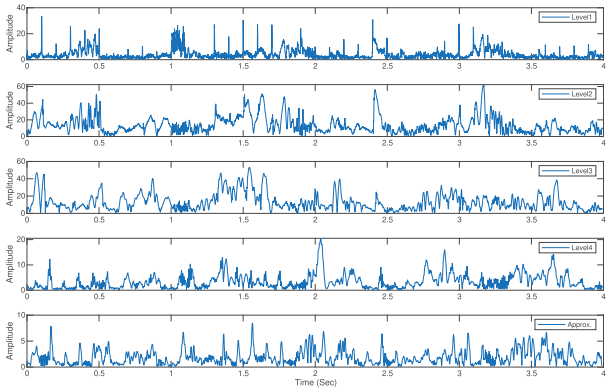


Fig. 3. Example of MODWT decomposition with different levels, from data for the case of a single target walking along a straight trajectory.

in that observation window. With the chosen four layers for decomposition, the resulting frequency bands are level 1 from 225 to 450 Hz, level 2 from 109 to 233 Hz, level 3 from 54 to 116 Hz, level 4 from 27 to 58 Hz, and "Approx" level from 0 to 28 Hz.

It should be noted that artifacts such as those in the Doppler spectrum around 1 s in Fig. 2 are filtered and categorized as belonging to the high frequency level in Fig. 3 after MODWT decomposition. In the later results section, the effect of selecting features from different frequency bands will be analyzed. This ability to separate spectral artifacts from useful signals is one of the advantages of using a wavelet-based method for extracting features.

To better understand how the number of people affects the radar cross section (RCS) and frequency characteristics, refer to Fig. 4, which illustrates representative features from both the spatial and frequency domains. A clearer separation between each target's distributions enhances the classification accuracy. The width of the intensity in the RA domain, measured in angle bins, varies with the overall RCS of the group as the number of people changes. Additionally, the amplitude at Level 3 increases as the number of people increases. Level 3 has a frequency range spanning from 54 to 116 Hz, which corresponds to the band of Doppler frequencies mostly occupied by human walking activities.

C. Classifier

The classifier assigns a label to a given set of input features. In this study, the assigned label corresponds to the number of individuals in a group. Given the relatively limited size of the available datasets in this context, classification methods that do not rely on neural networks (NNs) are considered. Various classification algorithms exist, each exhibiting different performance characteristics depending on the extracted features and the specific classification task. To present comprehensive results, four common non-NN classification algorithms are selected and compared. They are: 1) k-nearest neighbors (KNN) [15]; 2) Naïve Bayes [16]; 3) support vector machine (SVM) [17]; and 4) random forest [18]. The SVM is a powerful and versatile supervised learning algorithm used for classification tasks. SVM aims to find the optimal hyperplane

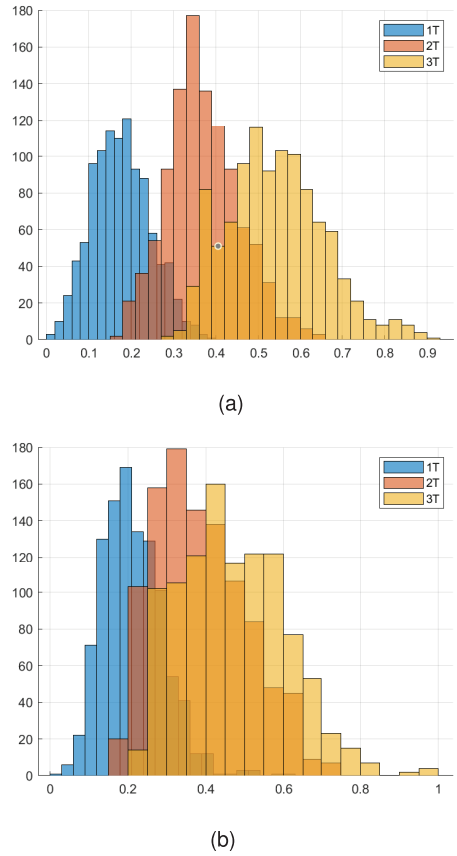


Fig. 4. Histograms with distribution of feature values for the cases of 1, 2, and 3 people. (a) Comparison of number of people versus angle bin width. (b) Comparison of number of people versus mean value of Level 3.

that maximizes the margin between different classes. The margin is the distance between the hyperplane and the nearest data points from each class, known as support vectors. The SVM is calculated by minimizing

$$\left[\frac{1}{n}\sum_{i=1}^{n}\max\left(0, 1 - N_i\left(w^T x_i - b\right)\right)\right] + \lambda\|w\|^2 \tag{4}$$

where $\max(0, 1 - N_i(w^T x_i - b))$ is the Hinge loss used for training the SVM classifiers, w and b are used to calculate the estimated labels, and N_i is the ith data point. $\lambda\|w\|^2$ is a term introduced to limit the margin size and make x_i assigned on the correct side. SVMs can be rather sensitive to the choice of the kernel and its parameters. In this case, the SVM classifier was created by using a Gaussian kernel, which appeared to be the most suitable choice for this work. The data are divided into 70% for training and 30% for testing.

After obtaining the output of the classifier for each time/observation window, there is a median filter applied to make the output smoother and remove temporary prediction errors.

D. Seamless Counting

To extract suitable Doppler frequency features using the MODWT approach, a longer observation window is required. However, this window exceeds the time window used in the

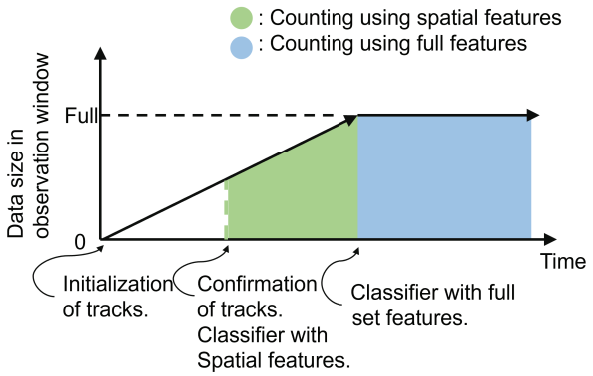


Fig. 5. Diagram of the proposed tracking approach with integrated seamless counting over time. There is an interval, while the classifier can only utilize a subset of features while the data are still being collected.

tracking process for data association and track confirmation. The difference in processing speeds between these two stages creates a time gap between trajectory confirmation and the availability of sufficient data for accurate Doppler frequency feature extraction. In contrast, spatial features derived from RA maps are available immediately for each frame. These features can be used for an initial counting estimation, ensuring continuous, "seamless" results, while more data are being acquired.

Fig. 5 presents an illustration of the seamless counting process. The green zone represents the time interval during which only spatial features are available; however, the data are insufficient to reliably compute Doppler frequency-based features. The blue zone indicates the period when the classifier can perform people counting using the full set of features.

In general, once a track is initialized, the amount of data available for extracting spatial features gradually increases. The counting classifier, based on spatial features, becomes active upon track confirmation and continues operating until the observation time window is complete. After this point, the classifier incorporates both spatial and frequency features for improved accuracy.

For a summarizing overview of the proposed tracking combined with the group classifier, it is noted that the method processes the current data input from the RA-Doppler domains along with previously extracted features from confirmed tracks with the same tracking ID. After making a prediction, the classifier passes the predicted count of the number of people to the tracking block in a sort of feedback connection, which in turn helps manage the number of active tracks.

III. Experimental Setup

This section presents the experimental setup, including the radar and reference sensor. Data were collected across six scenarios to validate the proposed method. In addition, the metrics used to quantify the results is introduced.

A. Radar Sensor

A commercial 24-GHz FMCW radar (by Joby Austria, former INRAS) with a relatively low bandwidth of 250 MHz

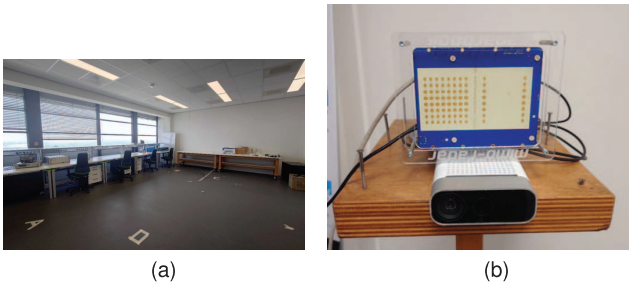


Fig. 6. Experimental environment for data collection in the MS3 radar laboratory at TU Delft. (a) Experimental room. (b) Setup of INRAS radar and Azure Kinect DK camera.

TABLE II
Summary of the Six Different Movement Scenarios Collected for Testing the Proposed Method

#NO.	Movement Scenarios	Duration (min)
①	1 Target walking forward and backward	20
②	1 Target randomly walking	10
③	2 Targets walking forward and backward	20
④	2 Targets following each other while walking	5
⑤	2 Targets randomly walking	10
⑥	3 Targets walking forward and backward	20
⑦	4 Targets walking forward and backward	10

is used to evaluate the performance. The relatively small bandwidth results in a range resolution of approximately 60 cm, which indicates that the target occupies a large area in the range profile while also making it difficult to isolate individual body parts. The detailed parameters used are listed in Table I. The radar is equipped with 15 virtual channels in azimuth, which provide a fine angular resolution (7.63°). Furthermore, an analysis of the performance when using a different number of receiving channels is possible.

B. Environment for Data Collection

For the purpose of this research, a dataset was specifically collected with five individuals, and up to three persons were moving simultaneously in the laboratory room of the MS3 group at TU Delft. Essentially, the number of people simultaneously moving in the environment was 1, 2, or 3. To emulate the cluttered environment of normal office space, pieces of furniture such as tables, chairs, and cabinets were placed in the environment, and a metallic curtain was also present at the window, which contributed to the relatively high level of multipath recorded. As illustrated in Fig. 6(a), a normal laboratory room with open space in the center is presented. The radar was placed at a height of around 1.3 m in the corner of the room, with its line of sight pointing along the diagonal of the room to get wide coverage, as shown in Fig. 6(b). In total, six different movement scenarios are performed and collected. These activities are listed in Table II, where it is shown that they include a mixture of different walking patterns with variable distances between the people in case of groups of them being present. Most of the activities are recorded for around 20 min. This leads to having around 51 000 frames of data collected and used for validation. Since the different experiments were performed in an indoor laboratory room, the radar sensor collects the reflections from the

targets but also the multiple reflections from the wall or other large pieces of furniture. Ghost targets resulting from multipath propagation are in the current approach mitigated using a physical constraint-based method. Specifically, the approach enforces spatial restrictions such that only trajectories confined within the known physical boundaries of the room are retained for subsequent feature extraction. More complex methods that aim to explicitly model or even exploit multipath components [19], [20] are left for future work.

C. Reference Sensor

To assess and compare different tracking algorithms in the proposed pipeline, a reference sensor is needed to act as ground truth. To quantitatively analyze the tracking performance, an auxiliary RGBD camera (Azure Kinect DK) was used to collect such ground truth data at the same time as the radar measurements [21]. The reference method used for positioning analysis is the one in [7]. The RGB data were processed by Detectron 2 for human detection [22], with the output being a bounding box. Its center is considered as the human position. Then, the bounding box position is mapped to a depth map to obtain the 2-D positioning of the human. Notably, due to the maximum range of the camera being limited to approximately 6 m, tracking ID switches can happen in the ground truth if people exceed that range during measurements.

D. Performance Metrics

For a single target scenario, assessing distance errors in tracking (e.g., median error, mean absolute error, and root-mean-square error) is sufficient in this context. To evaluate the accuracy of multitarget tracking systems, the OSPA metric is considered [23]. From the following equation, the OSPA metric has two components: one is the distance error (related to localization performance) and the other is the cardinality error (related to the number of people recognized to be in the room):

$$\text{OSPA} = \left(d_{\text{loc}}^{p} + d_{\text{card}}^{p}\right)^{1/p} \tag{5}$$

where $d_{\text{loc}} = \left\{(1/n)\sum_{i=1}^{m} d_c^p(x_i, y(i))\right\}^{1/p}$ is the localization error, and p denotes the order, which is set to 2 in this case, with $x(i)$ from a list of ground truth data and $y(i)$ from a list of tracks from the same timestamp, which is assigned to $x(i)$. In the formulation, $d_c(x, y) = \min\{d_b(x, y), c\}$, where c is the cutoff-based distance and is set to 1 m in this analysis. $d_{\text{card}} = \{((n + q) - m/(n + q))c^p\}^{1/p}$ is the cardinality error component, where n is the number of considered tracks and m is the number of ground truth tracks. Since the OSPA is the final metric chosen to compare the overall performance, it also includes the difference q between the classifier predictions and the true number of targets present in the scene.

IV. Results

This section presents the results for the proposed method by comparing the four previously mentioned classifiers and assessing the effect of using different subsets of the considered features and levels of the MODWT decomposition. Additionally, the overall performance for tracking, localization, and estimation of number of people via the OSPA metric is discussed. The proposed method is compared to CVD-based counting, as well as conventional tracking methods.

TABLE III

Accuracy Comparison of Different Classification Methods. (Spatial: Features Extracted From RA Map and Frequency: Features Extracted From MODWT Method)

Input features	Classifier methods	Accuracy (%)
Spatial + frequency	KNN	97.2
Spatial + frequency	Naïve Bayes	73.4
Spatial + frequency	SVM	99.2
Spatial + frequency	Random Forest	95.4

TABLE IV

Accuracy Comparison of Different Wavelet Families and Orders

Wavelet family	Order	ACC (%)
Symlets	Sym1	53.66
	Sym2	65.39
	Sym3	73.68
	Sym4	76.36
	sym6	71.57
Daubechies	db4	70.36
Fejér-Korovkin filters	fk4	62.2
Coiflets	coif1	75.77

A. Analysis of Classification Algorithm

Results for the four considered statistical classifiers are compared. The input features are both spatial and Doppler frequency features. The accuracy of each classifier is listed in Table III. From this, the SVM provides the best performance among the four classifiers. The KNN shows better performance than the random forest method, while the Naïve Bayes method performs the worst. The potential reason is the redundancy of information between each MODWT level against its assumption on independence of features. The SVM used here is with a Gaussian kernel and a kernel scale of 4.5. The smaller kernel scale is empirically selected to generate a complex decision boundary, but might eventually cause overfitting.

B. Analysis of Wavelet Family for Feature Extraction

The selection of wavelet family can be critical to the quality of feature extraction and consequently affect the classification accuracy. Different wavelet families and their orders exhibit varying time–frequency localization properties, symmetry, and filter lengths, which influence the performance in effectively representing the original signal characteristics. In this work, a few commonly used wavelet families were evaluated in terms of the impact of representing the features from the Doppler spectrum. Table IV shows the averaged testing accuracy metric for different wavelet families and orders. The classifier is trained with frequency features only in this case. The analysis results show that the Symlets wavelet with order 4 and the Coiflets wavelet with order 1 exhibit comparable performance in terms of classification accuracy. Ultimately, Sym4 was selected as the mother wavelet in this work due to

TABLE V
ACCURACY COMPARISON OF THE COMBINATION OF SPATIAL AND FREQUENCY FEATURES

Input Features	Proposed Accuracy (%)	CVD based counting [5] Accuracy (%)
Spatial + frequency	99.78	94.86
Spatial	96.93	95.06
Frequency	98.48	58.02
PCA (80% principal)	94.61	78.12

its slightly higher classification accuracy, offering improved overall performance.

C. Analysis of Spatial and Frequency Domain Features

In this section, the accuracy of the proposed method with a combination of spatial and frequency domain features is analyzed. By using all features together, the highest accuracy is obtained. However, it is noted that using only features from one domain in isolation does not reduce too much the performance, with the spatial features alone offering better results than the frequency ones. Moreover, to reduce the overall feature vector size and its redundancy, principal component analysis (PCA) is performed. The PCA reduces the dimension of the feature vector and helps prevent the model from overfitting. With 80% of the information retained in the PCA, the accuracy is 94.61% as shown in Table V; this is 5.17% lower than the original results with the full feature vector.

For comparison, the CVD-based counting method used in [5] is also implemented. As mentioned in Section II-B, the CVD method could not extract suitable features during the targets' direction changes, which happen frequently in an indoor environment. The accuracy results are shown in the right column of Table V. In particular, while the accuracy using all combined features is only slightly lower than when using MODWT features, the performance significantly declines when relying only on CVD-based frequency domain features. Furthermore, frequency features do not contribute to the accuracy in this case. This trend is also evident in the PCA (80%) scenario, where the accuracy decreases by approximately 16.7%, a larger drop compared to the proposed method.

D. Analysis of Different MODWT Level Features

Unlike other wavelet approaches for signal decomposition, with the MODWT approach, the same data size remains for each level. It is also observed by looking at the Doppler spectrum shown in Fig. 2 that the most significant components of the human signature correspond to level 2 (109–233 Hz) and level 3 (54–116 Hz) of the MODWT decomposition. The higher frequency band corresponds to regions with significant Doppler activity, which contains less information in this work. This is reflected in the accuracy values shown in Table VI to use only a single MODWT level isolated as input features, where Level 2 achieves the highest accuracy. In contrast, Level 1 exhibits lower accuracy due to containing less relevant information. As shown in the right row of Table VI, the highest accuracy is obtained when spatial features are combined with Level 3. This suggests that the low-frequency band likely captures critical information related to changes in movement direction, making it more effective than the high-frequency band.

TABLE VI
ACCURACY COMPARISON OF THE COMBINATION OF DIFFERENT LEVELS OF MODWT FEATURES. APPROX DENOTES THE LOWEST FREQUENCY RANGE FROM 0 TO 28 HZ

MODWT Level	Frequency (Hz)	ACC of frequency features (%)	ACC of Spatial + Frequency (%)
Approx	0-28	88.50	99.20
level4	27-58	87.28	99.03
level3	54-116	94.22	99.71
level2	109-233	94.23	99.65
level1	225-450	88.76	99.55

E. Analysis of Overall Tracking and Counting Performance

In terms of overall people tracking and counting performance, the modified OSPA metric mentioned in Section III-D can be used for quantitative analysis. This includes both position/localization error as well as a cardinality error, which also accounts for possible misclassifications on the number of people by the classifiers. Examples of OSPA-based performance analysis follow.

First, the prediction results given by the classifier for an example of three target scenarios are shown in Fig. 7(a). The trajectories can be interrupted due to missed detections during tracking. The different tracking IDs are distinguished by colors. The upper plot in the figure is computed before applying the median filter. The lower plot shows the prediction after the application of the median filter with 25 samples. This helps remove temporarily wrong predictions and improve the final accuracy. However, the median filter also has a limitation; when the wrong predictions are consistent for more than half the window length, the error is propagated for a long time until the true value accumulates back for a long time. This drawback can be observed from the lower plot of Fig. 7(a) around 30 s. This can also be seen as the limitation of statistical classifiers when applied to time-series data. The features are extracted by shifting a window considering a single block of data; however, the model does not have a feedback loop from the previous predictions to the current frame.

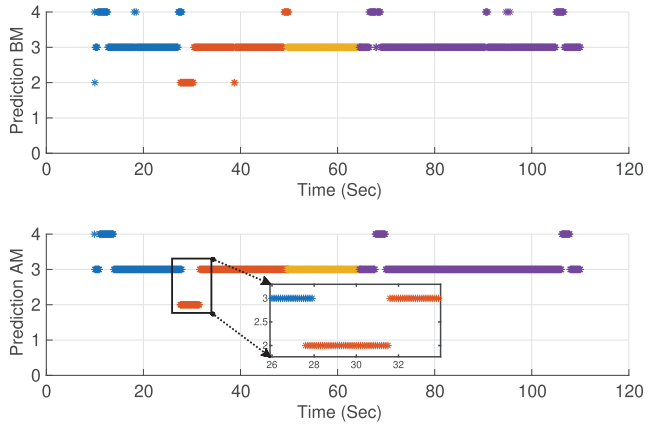
Then, the overall OSPA plot over time is shown in Fig. 7(b) for the same example. At the beginning of the plot, there is no confirmed trajectory. Hence, the OSPA raises its value up to 1, which is the cutoff distance defined in Section III-D. Afterward, the mean value of the OSPA remains around a mean value of 0.32. It should be noted that there is an increase in the cardinality error component at about 30 s, which corresponds to the prediction errors of the classifier, as discussed in the previous paragraph.

An overview of detailed performance metrics for each scenario previously listed in Table II is presented in Table VII. The proposed pipeline is compared with a CVD-based counting method and a conventional tracking method (without classifier). Specifically, the presented metrics include the

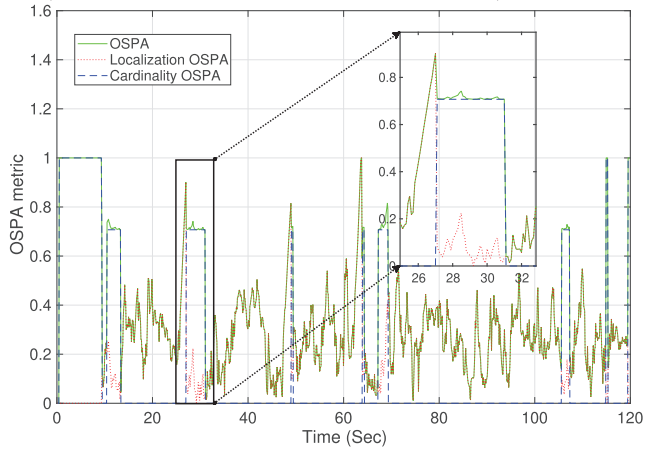
TABLE VII

SUMMARY OF PERFORMANCE METRICS FOR TRACKING WITH CLASSIFIER. (BM: BEFORE MEDIAN FILTER AND AM: AFTER MEDIAN FILTER)

Scenarios	Proposed			CVD based counting [5]			Conventional (tracking)
	Accuracy (%) BM	Accuracy (%) AM	OSPA	Accuracy (%) BM	Accuracy (%) AM	OSPA	OSPA
① 1 target walking	94.3	95.29	0.243	98.01	98.48	0.204	0.203
② 1 target random walking	88.21	92.63	0.318	92.28	98.0	0.274	0.255
③ 2 targets walking	100	100	0.208	91.57	93.62	0.253	0.726
④ 2 targets following	94.42	95.89	0.454	80.82	86.32	0.533	0.619
⑤ 2 targets random walking	96.75	96.64	0.385	74.21	79.42	0.469	0.714
⑥ 3 targets walking	89.8	91.45	0.321	38.55	48.31	0.495	0.826
⑦ 4 targets walking	81.65	80.16	0.413	50.97	53.79	0.544	0.863
Average	92.16	93.15	0.335	75.20	79.70	0.396	0.601



(a) Classification results before the median filter and after median filter (BM: before median filter, AM: after median filter)



(b) OSPA results over time

Fig. 7. Analysis of results for the case of three targets walking forward and backward. (a) Number of counted people and (b) OSPA metric.

classifier accuracy for counting before (BM) and after (AM) the median filter and the overall OSPA. An average performance across all scenarios is also presented for each method. The average accuracy of the proposed method is 92.16% and improved to 93.15% with the median filter to eliminate temporary prediction error. Moreover, the counting accuracy is around 13.45% higher than the CVD-based counting method. In terms of OSPA performance, the proposed method is 0.061 better than CVD-based counting. As the tracking approach is identical in both cases, the performance is improved by the contribution of higher classification accuracy and seamless counting. It has been observed that the CVD-based method from [5] does not achieve the desired results for more than one person as a group. On the other hand, the conventional tracking method is not able to generate a trajectory for each target, as it cannot use the additional information on the number of people from the classifier. This reflects into higher error in the OSPA.

In most cases, the median filter helps improve the accuracy, but in the case of the two-target random walking scenario ⑤ (refer to the 5th scenario from Table II) and the four-target walking scenario, the accuracy after the median filter is slightly less than before. The reason is that the median filter is not able to correct instances of continuous, consistent errors. For the single target scenario, the proposed method and CVD-based counting show a bit higher OSPA than conventional tracking due to the effect of the counting error. This error is still in an acceptable range. Overall, it can be seen that the performance is much improved compared to the conventional tracking method.

To provide additional examples of results, the trajectory of a single target randomly walking and two targets following each other is shown in Figs. 8 and 9, respectively. Due to the multiple overlaps of trajectories over time, the trajectory in these figures is displayed in two time segments spanning in total from 0 to 60 s. As mentioned before, the presence of ghost targets is not considered in this work as they are forcibly removed when present outside the physical boundaries of the room. Hence, only the detections inside the room are selected for tracking and feature extraction. It should be noted that the ground truth is discontinued due to the limited detection range of the used sensor. The trajectory of the single target is relatively well overlapped with the ground truth, with a corresponding OSPA of only 0.318. The scenario with two targets following each other is a more challenging case, both in terms of classification and tracking. With the proposed method, the trajectory follows the ground truth relatively well, with OSPA values reported around 0.454.

To further test the approach performance, an additional dynamic experimental scenario was collected involving three moving targets, which were later split into groups of one and two individuals, respectively. The resulting trajectory is shown in Fig. 10. The average OSPA error in this scenario was 0.587. As shown in the classification results in the top subfigure,

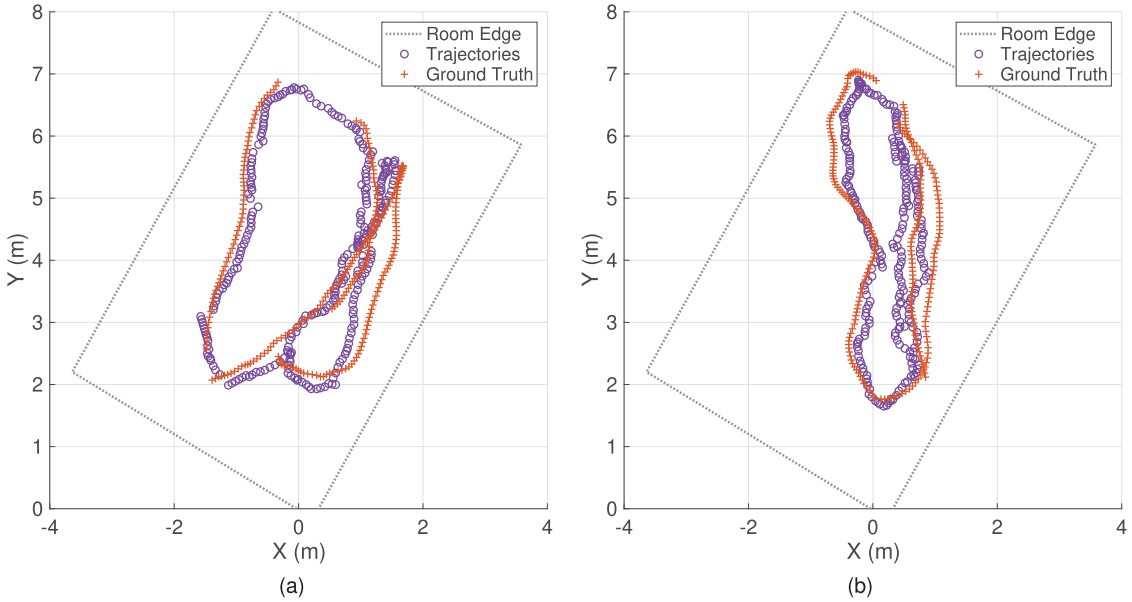


Fig. 8. Example of trajectories results for the scenario of a single target randomly walking. (a) Time duration from 0 to 30 s. (b) Time duration from 31 to 60 s.

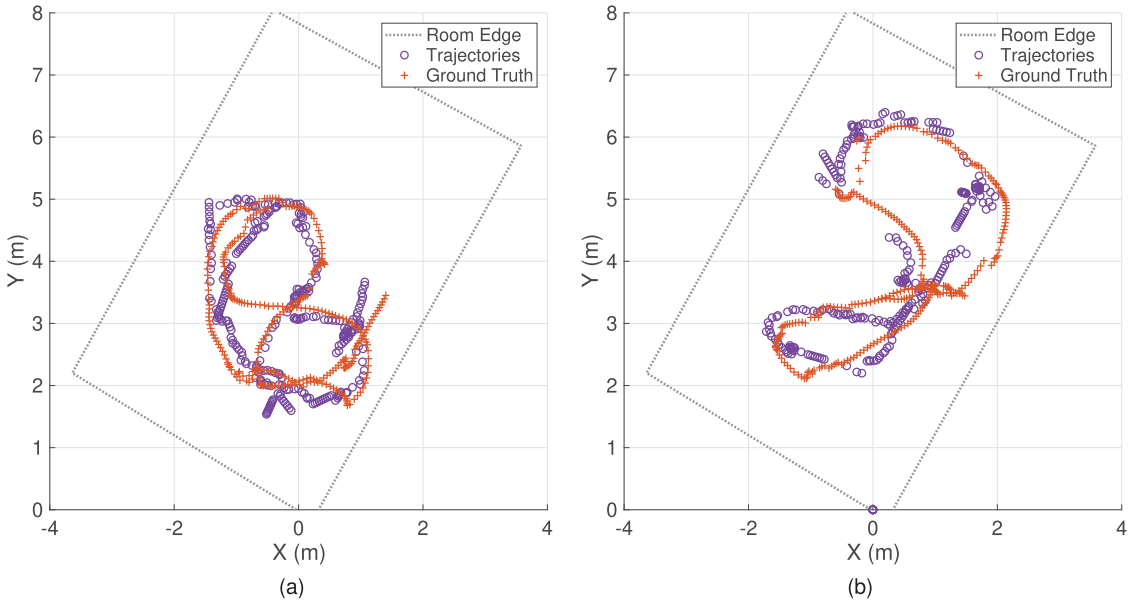


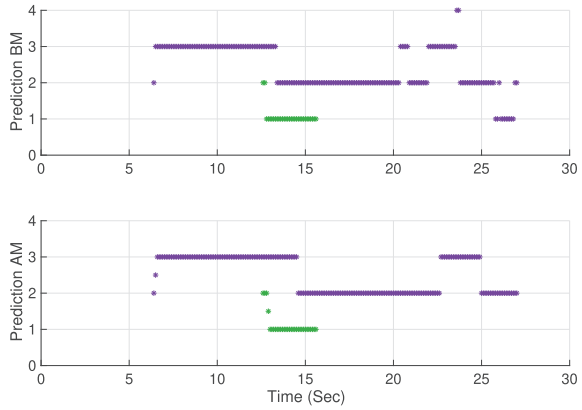
Fig. 9. Example of trajectories results for the scenario of two targets following each other and walking along a random path. (a) Time duration from 0 to 30 s. (b) Time duration from 31 to 60 s.

the total estimated number of people accurately follows the changes in group configuration and the appearance of new trajectories. A median filter was applied to suppress outliers; however, this also introduced a slight delay in reflecting transitions in the number of people detected.

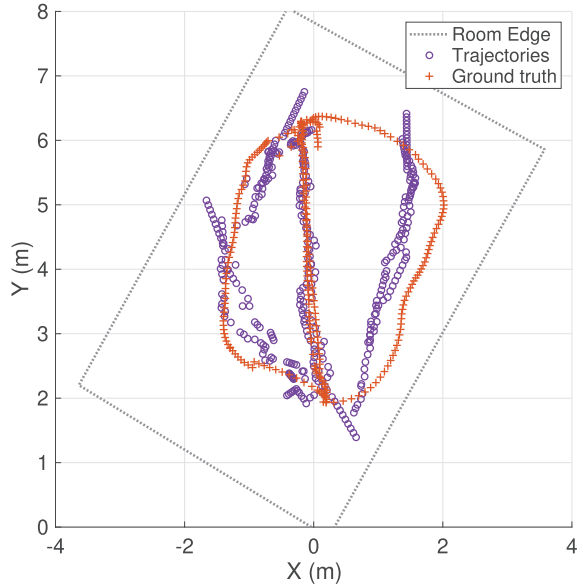
F. Additional Discussion on Results

The problem of counting the number of people has been approached as a classification task, as the target variable to be predicted is inherently discrete. Although this could be framed as a regression task, predicting a continuous value for the number of people would be less practical, requiring rounding to the nearest integer. On the other hand, classification offers a probability distribution over possible outcomes, making the solution more interpretable and practical.

Although the SVM method for classification used in this work shows a significant performance, it works by only considering input features frame by frame independently, without attempting to make connections between the current input and previous output(s). Methods based on NNs have been found good at dealing with time-series or sequence data such as long short-term memory (LSTM) [24] and recurrent NN (RNN) [25]. Furthermore, deep learning networks such as transformers have been studied for classification tasks [26]. For future work, the extracted features of multiple frames



(a) Classification results before the median filter and after median filter (BM: before median filter, AM: after median filter)



(b) Trajectories for three-target split & merge scenario

Fig. 10. Dynamic scenario of three targets splitting into two groups of 1 and 2 individuals and then re-merging (a) classification results and (b) trajectory results.

could be used as input as a sequence and the network can find the connection in a short period and reduce the temporal error.

V. Conclusion

In this article, a processing pipeline is proposed to approach track groups of people moving together and count their numbers in indoor environments. The pipeline is designed to handle the phenomenon of people moving independently of changes in direction or stop and go. The proposed approach combines a tracker based on the extended Kalman filter framework with a classifier to obtain the number of people. The classifier utilizes the spatial features from the RA map and Doppler frequency features with wavelet decomposition. As a result, the pipeline outputs the location and number of people appearing over time.

The approach is tested and validated with experimental data from a 24-GHz FMCW radar. The result shows that the proposed method achieves 93.15% accuracy in counting the number of people and a combined tracking metric OSPA of 0.335.

Acknowledgment

The authors would like to thank the Huawei team, Zhong Chen, Yanming Wu, and Jingjing Chen, for the technical discussions and are also grateful to the volunteers who participated in the data collection. They also thank anonymous reviewers for their constructive comments and suggestions in the review process.

References

[1] S. Yoo, D. Wang, D.-M. Seol, C. Lee, S. Chung, and S. H. Cho, "A multiple target positioning and tracking system behind brick-concrete walls using multiple monostatic IR-UWB radars," *Sensors*, vol. 19, no. 18, p. 4033, Sep. 2019.

[2] Texas Instruments. (2020). *PeopleTrackingandCounting Reference Design Using Radar mmWave Sensor.pdf*. [Online]. Available: https://www.ti.com/lit/ug/tidue71d/tidue71d.pdf

[3] A. Ninos, J. Hasch, M. Heizmann, and T. Zwick, "Radar-based robust people tracking and consumer applications," *IEEE Sensors J.*, vol. 22, no. 4, pp. 3726–3735, Feb. 2022.

[4] D. Wang, J. Park, H.-J. Kim, K. Lee, and S. H. Cho, "Noncontact extraction of biomechanical parameters in gait analysis using a multi-input and multi-output radar sensor," *IEEE Access*, vol. 9, pp. 138496–138508, 2021.

[5] L. Ren, A. G. Yarovoy, and F. Fioranelli, "Grouped people counting using mm-wave FMCW MIMO radar," *IEEE Internet Things J.*, vol. 10, no. 22, pp. 20107–20119, Nov. 2023.

[6] D. Wang, F. Fioranelli, and A. Yarovoy, "Analysis of processing pipelines for indoor human tracking using FMCW radar," in *Proc. IEEE Radar Conf. (RadarConf24)*, May 2024, pp. 1–6.

[7] D. Wang, F. Fioranelli, and A. Yarovoy, "Quantitative assessment of people tracking with FMCW MIMO radar," in *Proc. 21st Eur. Radar Conf. (EuRAD)*, Sep. 2024, pp. 380–383.

[8] M. Ester, H. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proc. 2nd Int. Conf. Knowl. Discovery Data Mining*, 1996, pp. 226–231.

[9] G. Welch, *An Introduction To the Kalman Filter*. Chapel Hill, NC, USA, 2001.

[10] I. J. Cox, "A review of statistical data association techniques for motion correspondence," *Int. J. Comput. Vis.*, vol. 10, no. 1, pp. 53–66, Feb. 1993.

[11] Y. Bar-Shalom, F. Daum, and J. Huang, "The probabilistic data association filter," *IEEE Control Syst. Mag.*, vol. 29, no. 6, pp. 82–100, Dec. 2009.

[12] F. Xiao, T. Lu, M. Wu, and Q. Ai, "Maximal overlap discrete wavelet transform and deep learning for robust denoising and detection of power quality disturbance," *IET Gener., Transmiss. Distribution*, vol. 14, no. 1, pp. 140–147, Jan. 2020.

[13] J. Quilty and J. Adamowski, "A maximal overlap discrete wavelet packet transform integrated approach for rainfall forecasting—A case study in the awash river basin (Ethiopia)," *Environ. Model. Softw.*, vol. 144, Oct. 2021, Art. no. 105119.

[14] V. Alarcon-Aquino and J. Barria, "Change detection in time series using the maximal overlap discrete wavelet transform," *Latin Amer. Appl. Res.*, vol. 39, no. 2, pp. 145–152, Jun. 2009. [Online]. Available: https://www.scielo.org.ar/scielo.php?script=sci_abstract&pid=S0327-0793%2009000200009&lng=es&nrm=iso&tlng=en

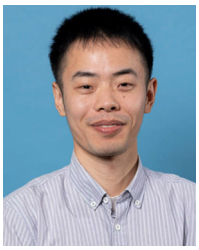
[15] G. Guo, H. Wang, D. Bell, Y. Bi, and K. Greer, "KNN model-based approach in classification," in *Proc. OTM Confederated Int. Conf., Move Meaningful Internet Syst.*, Catania, Italy, Nov. 2003, pp. 986–996.

[16] I. Rish, "An empirical study of the naive Bayes classifier," in *Proc. IJCAI Workshop Empirical Methods Artif. Intell.*, Seattle, WA, USA, vol. 3, 2001, pp. 41–46.

[17] M. A. Hearst, S. Dumais, E. Osuna, J. Platt, and B. Schölkopf, "Support vector machines," *IEEE Intell. Syst. their Appl.*, vol. 13, no. 4, pp. 18–28, Jul. 1998.

[18] M. Belgiu and L. Drăguţ, "Random forest in remote sensing: A review of applications and future directions," *ISPRS J. Photogramm. Remote Sens.*, vol. 114, pp. 24–31, Apr. 2016.

[19] R. G. Guendel, N. C. Kruse, F. Fioranelli, and A. Yarovoy, "Multipath exploitation for human activity recognition using a radar network," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5103013.
[20] R. Feng, E. D. Greef, M. Rykunov, H. Sahli, S. Pollin, and A. Bourdoux, "Multipath ghost recognition for indoor MIMO radar," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5104610.
[21] (Sep. 2025). *Azure Kinect DK-develop AI Models — Microsoft Azure*. [Online]. Available: https://azure.microsoft.com/en-us/products/kinect-dk
[22] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, *Detectron2*. [Online]. Available: https://github.com/facebookresearch/detectron2
[23] D. Schuhmacher, B.-T. Vo, and B.-N. Vo, "A consistent metric for performance evaluation of multi-object filters," *IEEE Trans. Signal Process.*, vol. 56, no. 8, pp. 3447–3457, Aug. 2008.
[24] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.
[25] K. Cho et al., "Learning phrase representations using RNN encoder–decoder for statistical machine translation," 2014, *arXiv:1406.1078*.
[26] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.



Dingyang Wang (Member, IEEE) received the B.S. degree in electronic engineering from Andong National University, Andong-si, Gyeongsangbuk-do, South Korea, in 2015, and the Ph.D. degree in gait analysis with MIMO radar from Hanyang University, Seoul, South Korea, in February 2022, under the supervision of Prof. Sung Ho Cho.

In March 2016, he joined the Radar Computing Laboratory, Seoul, with the joint M.S. and Ph.D. Program. Since April 2023, he has been a Post-Doctoral Researcher at the Microwave Sensing, Signals and Systems (MS3) Group, Faculty of Electrical Engineering, Delft University of Technology, Delft, The Netherlands. His research interests include radar signal processing and multitarget tracking.

Sen Yuan (Member, IEEE) was born in Shanxi, China, in 1998. He received the Ph.D. degree in electrical engineering from Delft University of Technology, Delft, The Netherlands, in 2024.

He is currently a Post-Doctoral Researcher with the Microwave Sensing, Signals and Systems (MS3) Group, Delft University of Technology. His research interests include SAR imaging, signal processing for radar systems, and new scheme of radar system design.

Dr. Yuan is an Associate Member of the IEEE Signal Processing Society Autonomous Systems Initiative (ASI). He was a recipient of European Microwave Association Student Grant from 2021 to 2024. He serves as an Associate Editor for *IEEE Aerospace and Electronic Systems Magazine*.

Alexander Yarovoy (Fellow, IEEE) received the Diploma degree (Hons.) in radiophysics and electronics and the Candidate Physics and Mathematical Sciences and Doctor Physics and Mathematical Sciences degrees in radiophysics from Kharkov State University, Kharkiv, Ukraine, in 1984, 1987, and 1994, respectively.

In 1987, he joined the Department of Radiophysics, Kharkov State University, as a Researcher, where he became a Full Professor in 1997. From September 1994 to 1996, he was with the Technical University of Ilmenau, Ilmenau, Germany, as a Visiting Researcher. Since 1999, he has been with Delft University of Technology, Delft, The Netherlands, where he has been leading the Chair of Microwave Sensing, Systems and Signals since 2009. He has authored and co-authored more than 500 scientific or technical articles, seven patents, and 14 book chapters. His current research interests include high-resolution radar, microwave imaging, and applied electromagnetics (in particular, UWB antennas).

Dr. Yarovoy was a recipient of European Microwave Week Radar Award for the paper that best advances the state-of-the-art in radar technology in 2001 (together with L. P. Ligthart and P. van Genderen) and in 2012 (together with T. Savelyev). In 2010, together with D. Caratelli, he got the Best Paper Award of the Applied Computational Electromagnetic Society (ACES). He served as the General TPC Chair for the 2020 European Microwave Week (EuMW-20), the Chair and the TPC Chair for the Fifth European Radar Conference (EuRAD-08), and the Secretary for the First European Radar Conference (EuRAD-04). He also served as the Co-Chair and the TPC Chair for the Tenth International Conference on GPR (GPR2004). He serves as an Associate Editor for IEEE TRANSACTIONS ON RADAR SYSTEMS. From 2011 to 2018, he served as an Associate Editor for *International Journal of Microwave and Wireless Technologies*. From 2008 to 2017, he served as the Director for European Microwave Association (EuMA).

Francesco Fioranelli (Senior Member, IEEE) received the Laurea (B.Eng.) (cum laude) and Laurea Specialistica (M.Eng.) (cum laude) degrees in telecommunication engineering from the Universitá Politecnica delle Marche, Ancona, Italy, in 2007 and 2010, respectively, and the Ph.D. degree from Durham University, Durham, U.K., in 2014.

He was a Research Associate at University College London, London, U.K., from 2014 to 2016, and an Assistant Professor at the University of Glasgow, Glasgow, U.K., from 2016 to 2019. He is currently an Associate Professor at TU Delft, Delft, The Netherlands. He has authored over 190 peer-reviewed publications and edited the books *Micro-Doppler Radar and Its Applications* and *Radar Countermeasures for Unmanned Aerial Vehicles* (IET-Scitech, 2020). His research interests include the development of radar systems and automatic classification for human signatures analysis in healthcare and security, drones and UAVs detection and classification, automotive radar, wind farm, and sea clutter.

Dr. Fioranelli received the four best paper awards and the IEEE AESS Fred Nathanson Memorial Radar Award in 2024.