Human tracking and identification through a millimeter wave radar

Citation for published version:

Digital Object Identifier (DOI):
10.1016/j.adhoc.2021.102475

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
Ad Hoc Networks

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.
Human Tracking and Identification through a Millimeter Wave Radar

Peijun Zhao\textsuperscript{a}, Chris Xiaoxuan Lu\textsuperscript{a}, Jianan Wang\textsuperscript{b}, Changhao Chen\textsuperscript{a}, Wei Wang\textsuperscript{a}, Niki Trigoni\textsuperscript{a}, Andrew Markham\textsuperscript{a}

\textsuperscript{a}Department of Computer Science, University of Oxford  
\textsuperscript{b}Deepmind, London

Abstract

The key to offering personalised services in smart spaces is knowing where a particular person is with a high degree of accuracy. Visual tracking is one such solution, but concerns arise around the potential leakage of raw video information and many people are not comfortable accepting cameras in their homes or workplaces. We propose a human tracking and identification system (mID) based on millimeter wave radar which has a high tracking accuracy, without being visually compromising. Using a low-cost, commercial, off-the-shelf radar, we first obtain sparse point clouds and form temporally associated trajectories. With the aid of a deep recurrent network, we identify individual users and show how to detect intruders. We evaluate and demonstrate our system, showing median position errors of 0.16 m, identification accuracy of 89\% and intruder detection accuracy of 73\% for 12 insiders. By increasing observation time from 2 s to 7 s, identification accuracy rises to 99\%.

Keywords: millimeter wave radar, tracking, identification, intruder detection

1. INTRODUCTION

Knowing ‘who is where’ is a key requirement for emerging applications and services in smart spaces, such as personalized heating and cooling, security management, efficiency monitoring, natural light adjustment, background music selection, etc \cite{1}. For these and other pervasive services to be truly seamless,
tracking and identification need to be performed with high accuracy and without active human effort.

Currently, most identification methods in smart spaces are device-based. Via the carried token, such as ID/swipe cards, active badge, smartphone, smartwatch, these methods identify users by the unique identifier of their personal devices. However, an implicit assumption made with these techniques is that the user and their identifying device are inseparable, which is not always the case.

In order to cope with more general scenarios, device-free methods have been proposed and are increasingly being adopted for human identification. Vision based techniques (e.g., cameras) are widely used methods in this category and have good performance when given a clear, frontal view of the face. However, cameras are intrusive and have a low user acceptance in domestic and commercial settings [2]. In contrast, radio frequency based methods are less intrusive and have also been utilized for device-free identification. For instance, it has been found that the variations in ambient WiFi signals can be used to recognize people while they walk [3, 4]. Unfortunately, such methods require a separate transmitter and receiver, and are limited to cases when users walk between the transmitter and receiver. The mmWave radar is a transceiver, so only requires a single device for tracking and identification. Point cloud based sensors, such as LiDAR and depth cameras are also able to identify and track people [5]. However, LiDAR is too expensive for home use while depth cameras only have a limited tracking range and accuracy. As an optics based sensor, they have similar user acceptance concerns as with conventional cameras.

In this paper, we introduce mID, a system that identifies people by their unique characteristics as sensed by a millimeter-wave (mmWave) radar. MMWave radar provides highly precise ranging by analyzing the reflection from obstacles in the environment, such as humans. It has a number of interesting properties. For example, a mmWave radar can be concealed behind furniture, as it is able to penetrate thin layers of different kinds of material [6], unlike optical based sensors. This property makes mmWave radar significantly more unobtrusive by
being able to be installed inside furniture or walls. The unobtrusive nature of mmWave radar also means domestic users are more likely to accept it, similarly to how Amazon Echo has been more widely accepted by users than web based cameras [7].

Exploiting these characteristics, we developed our human tracking and identification pipeline based on a commercial-off-the-shelf mmWave Radar. Our device is based on a single chip solution and operates in the $77 − 81$ GHz band. To the best of our knowledge, this is the first work to use the point cloud generated by a mmWave radar to track and identify people while they are walking.

The contributions of this work are as follows:

- We designed and implemented a human tracking and identification system using mmWave radar, which is capable of providing highly accurate tracking and identification.

- The proposed system is the first to identify people from mmWave radar point clouds, using deep recurrent neural networks.

- We trained the deep recurrent neural network with softmax loss and center loss, and used Random Forest Classifier for intruder detection.

- We evaluated the tracking and identification ability of our system, demonstrating median tracking accuracy of $0.16m$ and identification accuracy of $89\%$ for 12 people with intruder detection accuracy of $73\%$. By increasing observation time from 2 s to 7 s, identification accuracy rises to $99\%$.

The roadmap of this paper is: the basic principles behind mmWave radar are introduced in Section 2; Section 3 describes mID, which is our proposed mmWave gait identification system; Section 4 introduces the implementation details of mID and the setup of our testbed; Section 5 evaluates the system and Section 6 discusses the scalability and robustness, as well as limitations of mID. Related work is given in Section 7 and Section 8 concludes the paper.
2. Introduction to mmWave Radar Background

mmWave radar is based on the principle of frequency modulated continuous wave (FMCW) radar. FMCW radar has the ability to simultaneously measure the range and relative radial speed of the target. Detailed principles of FMCW radar are briefly introduced below.

2.1. Range Fourier Transform (range-FFT)

FMCW radar uses a linear ‘chirp’ or swept frequency transmission. The chirp is characterized by a start frequency $f_c$, bandwidth $B$ and duration $T_c$. When receiving the reflected signal, the radar front-end computes the frequency difference between the transmitter and the receiver with a mixer, which produces an Intermediate Frequency (IF) signal, from which the distance between the object and the radar can be calculated as:

$$ d = \frac{f_{IF} c}{2S} $$  \hspace{1cm} (1)

where $c$ represents the light speed $3 \times 10^8 m/s$, $f_{IF}$ is the frequency of the IF signal, and $S$ is the frequency slope of the chirp, which is calculated by $B/T_c$. To detect objects at different ranges, we perform an FFT on the IF signal, and each peak represents an obstacle at a corresponding distance. This is called ‘range-FFT’.

2.2. Clutter Removal

As we are interested in identifying people moving in the scene, the background, corresponding to stationary objects, needs to be removed before performing Doppler FFT. This is performed by subtracting a mean for each range bin per antenna across the chirps in a frame. With this step in the processing pipeline, the millimeter wave radar is able to generate a point cloud which does not contain static obstacles. However, this does not guarantee that the point

1Please refer to https://training.ti.com/mmwave-training-series for more information.
cloud does not contain noise. While the users move in the scene, parts of the
background objects are occluded and the reflections from these areas changes
over time, leading to noise in the radar point cloud.

2.3. Doppler Fourier Transform (Doppler-FFT)

A small change in the distance of the object leads to a large shift in the IF
signal phase, so we can obtain the relative velocity of the detected object by
transmitting two chirps with an interval of $T_c$ and measuring the phase difference $\omega$

$$v = \frac{\lambda \omega}{4\pi T_c}$$ (2)

where $\lambda$ is the wave length.

Using this technique, objects moving a different velocities at the same dis-
tance can be distinguished from one another.

2.4. Angle Estimation

Transmitters emit chirps with the same initial phase. With simultaneous
sampling from multiple receiver antennas, we can estimate the Angle of Arrival
(AoA), due to slight differences in phase of the received signals from each receiver
antenna. For two antennas, the AoA can be calculated with

$$\theta = \sin^{-1}\left(\frac{\lambda \omega}{2\pi d}\right)$$ (3)

For a linear receiver antenna array, the AoA can be calculated with FFT, which
is called 'angle-FFT'. Other AoA estimation algorithms, like MUltiple SIgnal
Classification (MUSIC) [8], can also be used in this step. The estimation is most
accurate at $\theta = 0$ and decreases with $|\theta|$.

3. System Design

mID is a tracking and identification system that exploits the unique prop-
ties of millimeter wave radar. It operates by transmitting an RF signal and
recording its reflections off objects. By analyzing the point cloud generated,
it then infers the people’s trajectories and identifies them from a database of known users. The mID system consists of four modules that operate in a pipelined fashion, as shown in Fig. 1:

1. **Point Cloud Generation.** In this module, a FMCW radar transmits millimeter waves and records the reflections from the scene. Reflection from clutters is removed after performing range-FFT on raw data, and then point clouds are generated after Doppler-FFT and Angle of arrival estimation.

2. **Point Cloud Clustering.** In this module, potential human objects are detected by merging individual points into clusters.

3. **Tracking.** In this module, mID associates the same human object in consecutive frames and uses a multiple object tracking algorithm to maintain trajectories of different people.
4. **Identification.** In this module, a deep recurrent neural network is used to recognize user identities from the sequential data of each user.

5. **Intruder Detection.** In this module, the classifier is trained with co-supervision of softmax-loss and center-loss. The distances to each center are used as features to decide whether the input sample belongs to one of the human subjects in the training set, or it is a sample from an unknown person.

In the following subsections, we describe each of these components in detail.

### 3.1. Point Cloud Clustering

The generated sparse point clouds are dispersed and not informative enough to detect distinct objects. Moreover, although static objects are discarded through clutter removal, the remaining points are not necessarily all reflected by moving people. As shown in Fig. 1, this noise can be significant and lead to confusion with points from nearby people. To determine which points in the scene are caused by reflections from people, mID first merges points into clusters using DBScan, a density-aware clustering method that separates cloud points based on their pairwise distance in the 3D space. A major advantage is that it does not require the number of clusters to be specified *a priori*, as in our case people walk in and fade out of the monitored scene at arbitrary times. Additionally, DBScan can automatically mark outliers to cope with noise.
However, in a real-world measurement study, we observed that points of the same person are coherent in the horizontal (x-y) plane, but more scattered and difficult to merge along the vertical (z) axis. Fig. 2 illustrates an example. We hence modify the Euclidean distance to place less weight on the contribution from the vertical z-axis in clustering:

$$D(p^i, p^j) = (p^i_x - p^j_x)^2 + (p^i_y - p^j_y)^2 + \alpha \ast (p^i_z - p^j_z)^2$$  \hspace{1cm} (4)

where $p^i$ and $p^j$ are two different points and the parameter $\alpha$ regulates the contribution of vertical distance.

3.2. Moving Object Tracking

To capture continuous individual point clouds to track and identify a person, we require an effective temporal association of detections as well as correction and prediction of sensor noise. Fig. 3 illustrates the flow of our multi-object tracker. We essentially create and maintain tracks for object detections from each frame. A new track is created for each object detection which is either from the first incoming frame or one which cannot be associated with an existing track. Inter-frame object association is based on the Hungarian Algorithm. If a track object is undetected for $D$ continuous frames, we mark the track as inactive and exclude it from successive associations. Finally, we apply a Kalman Filter to predict and correct tracks. These two components are discussed in more detail below.
3.2.1. Detection and Association

We use the Hungarian Algorithm\cite{10} which is an effective combinatorial optimization algorithm. Our objective is to create an association between each object detection and maintained track objects so that the combined distance loss is minimized. Here we are facing a many-to-many assignment problem where the cost matrix can be non-square because the number of active tracks $K_1$ and the number of object detection at the current timestamp $K_2$ can be different. Given $K$ as the greater of $K_1$ and $K_2$, we essentially augment the true cost matrix with dummy entries to construct a $K \times K$ matrix $M$ where $M_{i,j}$ represents the distance of centers between track object $i$ and object detection $j$ in current frame. If $M_{i,j}$ exceeds step size threshold $\theta$ we set the cost to be a large number $L$ to avoid association given the intuition that $j$ should be a new previously unseen person. If a detection is mapped to an augmented dimension, or if it is mapped to a correspondence with cost $L$, we ignore such mappings and create a new track for this detection. Similarly, if a track object is mapped to an augmented dimension or a correspondence with cost $L$, we treat the track object as undetected. This method enables us to successfully maintain tracks of detection.

3.2.2. Track prediction and correction

We use a Kalman Filter\cite{11} to correct for sensor noise and to offer predictive guidance in scenarios where tracked objects are undetected due to occlusion or temporary loss from the sensing region. For each track we maintain a state which consists of location and velocity along the x and y axes. For each track the initial state consists of the first detection location and velocity. At each successive time step, the Kalman Filter updates the current state variables with transition matrix along with corresponding uncertainties. Based on the current position and velocity, it estimates a new position/velocity as well as new covariance. The Kalman Filter produces estimates that tend to be more accurate than those based on a single sensor measurement, especially in our case with occasional undetected track objects.
3.3. User Identification

After the points corresponding to human objects are determined, we can use tracklets to recognize their identities. Specifically, from each frame in the trajectory, we use a fixed-size bounding box to enclose the points of potential human objects, and voxelize it to form an occupancy grid. Note that the occupancy grids inherently encapsulate body shape information. For instance, tall people tend to have higher center of mass. By feeding the sequential occupancy grids to a classifier, the ID of a tracklet is recognized based on both movement characteristics, i.e., gait, and body shape information. The tracklet used in mID is segmented with a sliding window method. A window contains consecutive occupancy grids for 2 seconds, with a 75% overlap ratio with the previous window. Extracting useful features directly from the occupancy grids is difficult, as most feature engineering methods are not effective for point cloud classification tasks [12]. The Long-short Term Memory (LSTM) network is an established recurrent neural network architecture suited for sequential data classification which is able to learn the features automatically through network training. We therefore propose using it as the identity classifier in mID. The 3-D data is first flattened and then each frame is converted into a feature vector. This is then passed into a bi-directional LSTM network followed by a dense layer. Lastly, a softmax layer is used to output the final classification result (see Fig. 4).

Figure 4: Classification Network Structure (see Sec. 3.3). T represents the number of data frames used for identification and K represents the number of people to distinguish. Numbers in bracket represent layer sizes.
3.4. Intruder Detection

Softmax layers tend to be very confident in their predictions, even when it could be a wrong prediction. This can also happen when the testing sample is actually not from any of the classes in the training set i.e. is out-of-set. As the goal of person identification is to classify into a 1-of-N category, an out-of-set sample e.g. an intruder, will most likely be simply regarded as one of the classes. In addition, the features learned via the softmax loss tend to scatter across the feature space, leading to a sample from an out-of-set intruder being unable to be clearly distinguished. To address this issue, it is necessary to have some form of clustering or metric embedding. This is achieved by training the classifier with a center loss in addition to the softmax loss. This has been proven
to have better performance and authentication ability with CNN networks for
face recognition [13].

The core idea of the center loss is to minimize the intra-class distance of
features extracted by the deep neural network from each sample, so that the
features of the same class samples tend to gather together in the space [14]. For
a sample from an intruder, the features extracted should be far away from the
feature centers of the classes in the training set, and thus we can tell that the
sample is from an outsider, rather than having the sample wrongly identified as
one of the users in the training set. We first define the loss as:

$$L = L_{softmax} + \lambda L_{center}$$  \hspace{1cm} (5)

where the total loss is a combination of softmax loss and center loss. We use a
parameter $\lambda$ to weigh the two losses. The center loss is defined as:

$$L_{center} = \sum_{i=1}^{m} ||f_i - c_{y_i}||^2_2$$  \hspace{1cm} (6)

where $f_i$ represents the $i^{th}$ feature and $c_{y_i}$ represents the center of $i^{th}$ feature of
label $y$. The center loss is the squared Euclidean distance between the learned
features of a sample and the center of the features in the class the sample belongs
to.

We implemented this modified network as shown in Fig. [5] Though the
output of Bi-directional LSTM layer can directly be used as a feature, it does
not converge when trained with center loss. We use a dense layer after the
Bi-directional LSTM to represent the features. We use an Embedding layer to
map the identity label to the center of the features. A custom layer is used to
calculate the squared Euclidean distance, and the output of the layer is used as
center loss.

Features from intruders have different spatial distributions, which leads to
different distances to each center of the features from each class. We use the
distances to different class centers as labels for intruder detection. We use a
Random Forest Classifier, which has been proven to work very well as a shallow
classifier, and outputs a binary prediction of whether the sample comes from an intruder or not.

4. IMPLEMENTATION

4.1. Testbed Setup

mID is developed on top of a commercial, off-the-shelf millimeter wave radar, IWR1443boost. The system was tested in a room with a Vicon optical tracking system which is able to provide ground-truth position of each marker with an error within 1 cm. Ground truth identities of the training and testing samples are manually labeled. The system consists of two parts, the radar and the backend. The radar senses data and generate 3D point cloud, which is transferred to the backend computer for further processing, as shown in Fig. 6. We implemented the deep neural network classifier with the Keras library and a Tensorflow backend.

4.2. Dataset Collection and Breakdown

We collected data from 28 participants, each of them walking randomly

\[\text{http://www.ti.com/tool/IWR1443BOOST}\]
in the designated area for 10 minutes. The study has received ethical approval SSD/CUREC1A CS_C1A_18_024. The ages of the participants range from 18 to 35, and 13 of the participants are female. The heights of the participants range from 155cm to 188cm, and the weights of the participants range from 48kg to 80kg. The body shape of each participant varies which mimics domestic settings. Data from 12 participants were used for Identification Evaluation in Section 5.3. For Intruder Detection in Section 5.4, the data from the other 16 participants were used as intruders. The data from the first 12 participants, which were used for identification evaluation, was split into training and testing set with a ratio of 11:1. For the data from the other 16 participants, data from 12 participants were used for training the Random Forest binary classifier, and the remaining data from 4 participants were used for the evaluation of intruder detection. This dataset breakdown remains the same for all evaluations in Section 4.

4.3. mID Configuration

4.3.1. Sensor Setup

The IWR1443Boost sensor was configured to use all three transmitter antennas and all four receiver antennas in order to generate 3D point cloud data. Start $f_c$ and end frequencies were set to $77\text{GHz}$ and $81\text{GHz}$ respectively, so the bandwidth $B$ was $4\text{GHz}$. The Chirp Cycle Time $T_c$ was set to $162.14\mu s$ and the Frequency Slope $S$ was set to be $70\text{GHz/ms}$. With such a configuration, mID has a range resolution of $4.4\text{cm}$ and maximum unambiguous range of $5m$. In terms of velocity, it can measure a maximum radial velocity of $2m/s$, with the resolution of $0.26m/s$. The sensor was set to transmit 128 chirps per frame and the number of frames per second was 33.

4.3.2. Classifier Training

Through multiple trials we worked out a set of parameters that have the best performance. Each frame of the input data was first flattened to a vector of dimension 16000. A bi-directional LSTM with 128 hidden units was used. We set the dropout ratio to 0.5 and used the Adam optimizer. We used a
balanced dataset where training/test sample ratio was set to 11:1. To decrease overfitting, we further augmented the training data to 8 times the original size, by shifting the data in X and Y axis respectively for 1 voxel, and rotating each frame by 90°, 180° and 270°. The model was trained for 30 epochs.

4.3.3. Intruder Detection

Based on extensive experiments, we find that 64 is the best size for the dense feature layer, which is used to extract features as shown in Fig. 5. If the feature layer is too large, the model can easily get overfitted, while if the feature layer is too small, it lacks representation ability which will also lead to performance downgrade. We set the loss weight of softmax loss and center loss to 1:0.5, which has the best performance in practice. We set the dropout ratio of both the Bidirectional LSTM layer and the feature extraction layer to 0.5, in order to reduce overfitting. The model is also trained with Adam optimizer for 30 epochs. The parameters of Random Forest Classifier are decided by grid search with 5-fold cross validation.

4.3.4. Parameters of DBScan Algorithm

DBScan has two parameters, namely Eps which indicates the maximum distance of two points in the same cluster and MinPts which indicates the minimum point number in a cluster to cope with noise points. In practice, we choose 0.05 as Eps and 20 as MinPts. α was set to 0.25 in the customized distance function.

5. Evaluation

5.1. Sensitivity Analysis

5.1.1. Non-Line-of-Sight Conditions

In the first experiment, we study the robustness of mmWave radar under occluded conditions. This property is important because optical imaging based tracking and identification methods, such as RGB and depth cameras, cannot cope with obstructions. We evaluate the robustness of millimeter wave radar
Figure 7: Impact of different materials on point cloud density of mmWave radar. The differences in point cloud density are all below 1% for all cases.

with four types of obstructions: foam, plastic wood and aluminium. We used a sheet of each material in turn with a thickness of approximately 3mm and a size of $10^5 \text{ mm}^2$. The obstacles are placed 1cm away from the sensor so that the signals cannot be transmitted in a line-of-sight condition. We let a user walk back and forth in front of the millimeter wave radar while collecting sensor readings. mID uses the generated 3D point cloud for tracking and identification, so we compare the percentage of change in point cloud density for mID. As can be seen in Fig. 7, mmWave is very robust against non-line-of-sight interference, with less than 1% change in point-cloud density. Robustness to thin obstructions could enable mID to work under furniture or concealed within a picture frame, etc., making it less intrusive.

5.1.2. Impact of weighing the vertical axis in DBScan

As introduced in Section 3, it is important to define the weighting parameter to make the DBScan algorithm work better. Therefore, we need to find a suitable value of $\alpha$ to obtain a good clustering result. In practice, we found that $\alpha = 0.25$ results in good clustering performance, as shown in Fig. 8. In contrast, when $\alpha = 0$, (points are effectively projected onto the $x-y$ plane), outliers are merged into the cluster. When $\alpha = 1$, (standard Euclidean distance) points corresponding to a person are split into two clusters.
Figure 8: Clustering results with different $\alpha$. A small $\alpha$ leads to loose clusters containing many noisy points. A large $\alpha$ splits a human object into two clusters. Setting $\alpha = 0.25$ gives the best empirical performance, as shown in the middle one.

5.2. User Tracking

Figure 9: Comparison of tracking performance between mmWave radar and Kinect v2. (a) Kinect v2 can only track people within 4.5m. (b) Inferred trajectory. (GT: Ground Truth by Vicon)

To evaluate the tracking accuracy of mmWave radar, we compared it to a Kinect v2, an RGB-D camera which is widely used in homes for gaming. We installed the mmWave radar and Kinect v2 co-located in a room. The Vicon system was set to track a marker placed on the top of a hat worn by the participant. Time was synchronized through a NTP server and the coordinate transformation matrices of the systems were calibrated. A state-of-the-art open
source tracking algorithm[^1] is used for Kinect v2 people tracking. The tracking error and a comparison of the trajectories is shown in Fig. 9, which shows the median tracking error of mID is \( \sim 0.16m \) whereas the Kinect is 0.9m. Besides a significantly smaller error compared to Kinect v2, the tracking range of mID is also larger than Kinect v2. mID can track people at a distance of more than 5.5m, whereas Kinect has a tracking limitation of 4.5m. This could be further extended, at the cost of a reduction in identification accuracy. This shows that the radar based technique is able to achieve highly accurate tracking over a larger tracking area, making it suitable for increasing levels of automation.

5.3. Identification Evaluation

The point cloud generated by mmWave radar makes it can be very hard, if not impossible, to analyse people’s gait with traditional vision based gait recognition methods [15, 16], because it is too sparse to recognize different body parts of the subject. Deep Neural Networks have the ability to automatically extract relevant features of the data while training the model, but it is not straightforward to tell which neural network architecture best suits the problem. We evaluated the performances of neural networks of different architectures and sizes on our dataset by performing an ablation study, which will be discussed in turn below.

5.3.1. Identification Performance

Overall, mID is able to reach an accuracy of 89% for 12 people. The confusion matrix is shown in Fig. [10]. Note that this performance is non-trivial considering that the original point clouds are very sparse, and demonstrates the utility of using deep-networks for feature extraction.

5.3.2. Impact of number of people

In this experiment, we further explore identification accuracy with varying group sizes. Intuitively, a smaller group size should make the problem easier.

[^1]: https://github.com/mcgi5sr2/kinect2_tracker
Fig. 11 shows the performance trend of mID when varying the number of participants from 4 to 12 with a step of 2. As we can see, mID is able to cope with various scenarios and works extremely well in the cases with \( \leq 6 \) people with 95% accuracy. In contrast, WiWho \cite{[17]}, based on WiFi CSI, achieves an
identification accuracy of 80% with 6 people. Note that, the number of people in domestic settings is generally less than 6, and we envision that mID could greatly benefit many applications in smart homes of the future.

5.3.3. Neural Network Architecture Comparison

To explore the best artificial neural network architecture for the identification problem, we compare 3 different architectures that are based on LSTM. We use variants of LSTM as it has been shown to be performant for end-to-end learning of time sequence data. Each model is trained under the same settings: 30 epochs and 0.5 dropout ratio. The LSTM layer used in all the 3 models has the same size of 256 and 128 hidden units. The CNN used in the CNN+LSTM model has two convolution layers, with a max pooling layer after each convolution layer. The CNN is time distributed, which means the data of each frame is first sent into a two-layer 3D CNN for feature extraction, then the sequence data was sent into LSTM for classification.

The accumulative identification accuracy of the different architectures is shown in Fig. 12(a). The accuracy differences between architectures become negligible after 3 guesses. However, Bi-directional LSTM converges more quickly and significantly outperforms the other two architectures within fewer guesses. This is presumably because Bi-directional LSTM is able to model the rich temporal correlations in a long sequence of frames from both ends. In contrast, a standard LSTM is essentially a feed-forward network that is difficult to encode the information from the beginning of a long sequence [18]. Such information loss degrades the identity inference performance.

5.3.4. Impact of Network Size

Beyond architecture, it is also important to consider the impact of the network size. If the network is too small, it may lack the representation ability, while networks that are too large can suffer from over-fitting problems. We evaluated the Bi-directional LSTM network with 3 different sizes and the comparison of the classification performance is shown in Fig. 12(b).
The Bi-directional LSTM with size of 256 and 128 hidden units significantly outperforms the same network architecture with a larger or smaller size, especially in the first two guesses. This suggests that a Bi-directional LSTM with size of 256, 128 hidden units is a good fit for this classification problem.

5.3.5. Impact of Observation Time

Intuitively, the longer that the sensor senses the walking pattern of a person, the more likely it is that the person can be identified correctly. Recall that the testing data is split into windowed samples with length of 2 seconds and 75% overlap between consecutive samples. By taking $k$ consecutive windows, the
Table 1: Precision, Recall, F1 score and Accuracy of Intruder Detection. Here the hypothesis is the person is intruder. Precision is the ratio of real intruders to all samples classified as intruders. Recall is the percentage of intruders identified. Accuracy is the chance of correctly classifying a sample.

<table>
<thead>
<tr>
<th>Insiders</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>0.83</td>
<td>0.95</td>
<td>0.89</td>
<td>0.87</td>
</tr>
<tr>
<td>6</td>
<td>0.77</td>
<td>0.88</td>
<td>0.82</td>
<td>0.81</td>
</tr>
<tr>
<td>8</td>
<td>0.75</td>
<td>0.88</td>
<td>0.81</td>
<td>0.80</td>
</tr>
<tr>
<td>10</td>
<td>0.73</td>
<td>0.88</td>
<td>0.80</td>
<td>0.78</td>
</tr>
<tr>
<td>12</td>
<td>0.79</td>
<td>0.70</td>
<td>0.74</td>
<td>0.73</td>
</tr>
</tbody>
</table>

Total duration would be equal to:

\[
duration = (k - 1) * 0.25 * t_{sample} + t_{sample}
\]  

(7)

where \(t_{sample}\) equals 2 seconds. We used consecutive windows and applied them as normal to the identification network. The final prediction for the longer sequence is simply decided with a majority voting policy on the prediction result of its samples. We evaluated the identification accuracy against different sequence duration for the identification results of 12 people scenario. The results are shown in Fig. 13. The percentage of correct prediction reaches 95% when the observation time is longer than 5 seconds, and further reaches around 99% at 6 seconds. Results show that with longer observation time, mID can have an excellent identification performance, even with more than 10 people in the dataset.

5.4. Intruder Detection Evaluation

In this section, we examine the performance of the intruder detection classifier through using the center loss in addition to softmax loss. We collected additional data from 16 people, who were not involved in the center loss model training process, who act as out-of-set intruders. We trained the base network using center loss and softmax loss to force intra-cluster feature distances to be large, using a different number of users in the insider dataset. We then used
12 out of the 16 intruders to train the binary random forest classifier together with the insider dataset. The remaining 4 intruders were used as testing data. The testing samples are from insiders and the 4 testing intruders, with a ratio of 1:1. We balanced both the training set and the testing set, i.e., using same number of samples from insiders and intruders, for a better understanding of the intruder detection ability of the proposed model.

Results are shown in Table 1. We can see that when there are fewer users in the dataset, the performance of intruder detection model increases greatly. We find that intruder samples scatter widely across the high dimensional feature space, leading to a higher chance of confusion when the number of users (classes) is high. From the results above, we can see that for intruder detection, mID also works best in places where the number of users is limited, such as smart homes. Note that with access to a larger training set e.g. hundreds of users, it could be possible to improve these results further.

6. Discussion on Scalability and Robustness

Despite the fact that this novel system works well for tracking and identification. We provide a discussion on scalability and robustness, as well as elaborate the limitations.

Scalability

1. Large Number of Non-concurrent Users. We have demonstrated the reliable performance of mID when working with a relatively small group of users (12). However, tracking and identifying a large number of non-concurrent people remains an open problem for two reasons. First, the point cloud generated by mmWave is sparse and sometimes the sparsity could significantly disturb human detection and tracking. Secondly, body shape and human gaits used in this work are weak biometrics, and could become ambiguous with the increasing number of users in the database.
2. *Large Number of Concurrent Users.* When multiple people appear in the scene at the same time, the tracking and identification ability of the system are both affected. Firstly, The IWR1443Boost we use in this work features a 3 transmitter 4 receiver MIMO antenna array. The best azimuth angular resolution of mID processing pipeline based on this device is $15^\circ$, so if two people are too close together, it can be hard for them to be separate apart in the point cloud generated. Also, due to hardware limit, the total points of each frame is limited. Multiple in the scene could share the point cloud, making it even more sparse, and extremely hard to identify people with the model trained on the dataset of single person walking. This work mainly aims to prove the feasibility of identifying people with mmWave radar, and identifying multiple people at the same time robustly will be one of our future works.

3. *Large Buildings* mID uses millimeter wave radar, and the wavelength of the signal cannot penetrate thick walls. As a result, when using the system in large buildings with many rooms, we need devices deployed in each room, and the sensors will not interference each other.

**Robustness**

1. *Monitoring Range.* In our experiment setting, we set the maximum unambiguous range of mID to $5m$, which is roughly $3/4$ the size of monitored room (see Sec. 4). In principle, the range of the mmWave radar can be as large as $30 m$, but this increased range comes at the cost of reduced spatial precision and worse signal-to-noise ratio. If the subject is too far away from the sensor, it is very hard to detect and distinguish them from sensor noise.

2. *Flat and Planar Surfaces.* As we found in the experiments, the
reflection profile of mmWave can be affected by flat and planar surfaces, such as windows or mirrors. As a result, noisy ‘mirror’ human objects appear occasionally when encountering these surfaces. Our experimental site is mainly with walls, which are typically not strong reflectors due to their dielectric properties. However, it is worthwhile to consider the impact of disturbing surfaces in real world deployment.

3. Clothes and Injuries Clothes made of different materials could have different reflection profile. And whether clothes can affect the identification accuracy remains an open question and will be one of our future works. Injuries could affect the identification accuracy as it is likely to change the features extracted by the deep neural network.

4. Room Clutter and Obstacles As mentioned in Section 3, clutter removal is applied when generating point cloud, so the reflection from static objects like room clutters would be removed in the point cloud and does not affect the system. However, there may be obstacles in the room like furniture that can block the signal and make the system lost tracking of the subject. There are many techniques and algorithms, including Multi-hypothesis Tracking [19], Joint Probabilistic Data Association [20], Maximum Joint Likelihood Tracking and Association [21], etc., which can be used for robust multi-object tracking. This works focus on proving the feasibility of using mmWave radar for human identification, so dealing with obstacles is out of scope of this work. We will look into this in our future works.
Table 2: Comparison of different identification methods and their relative merits.

<table>
<thead>
<tr>
<th></th>
<th>Floor Sensor</th>
<th>RGB</th>
<th>Depth</th>
<th>WiFi CSI</th>
<th>mID</th>
</tr>
</thead>
<tbody>
<tr>
<td>ID Accuracy</td>
<td>Moderate</td>
<td>Very high</td>
<td>High</td>
<td>Moderate</td>
<td>Moderate</td>
</tr>
<tr>
<td>Multiple People</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Tracking</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Env. Independent</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Privacy Concerns</td>
<td>None</td>
<td>High</td>
<td>Medium</td>
<td>Low</td>
<td>Low</td>
</tr>
<tr>
<td>Ease of Deployment</td>
<td>Very Difficult</td>
<td>Easy</td>
<td>Easy</td>
<td>Difficult</td>
<td>Moderate</td>
</tr>
</tbody>
</table>

7. Related Work

7.1. Device-Free Gait Recognition

There has been a lot of works on device-free gait recognition. Middleton et al. built a floor sensor based gait recognition system and achieved 80% accuracy over 15 people [22], at high deployment cost. Vision-based methods are one of the most established techniques. Some use silhouette analysis approaches like [16], others use model-based methods like in [23]. Besides monocular vision, multiple cameras, stereo cameras and depth cameras are also utilized in gait recognition tasks [24]. However, as long as video data is used, there will be a risk that users’ privacy would be compromised if data leakage occurs.

It has been shown that the Channel State Information (CSI) of WiFi signals, for both 2.4GHz and 5GHz bands, captures human gait information to a certain extent [4, 17]. These gait recognition systems are easy to deploy as WiFi devices are common in daily lives. However, such methods are generally scene-dependent and cannot handle environmental dynamics very well. Furthermore, these methods struggle to cope with identifying multiple people in the same scene.

A full comparison of different device-free identification methods are provided in Table. 2

7.2. Other RF-based Human-centric Applications

RF sensing has been widely used in human-centric applications, such as fall detection [25], occupancy counting [26], breathing monitoring [27] etc. Google
has recently published a touchless gesture interface based on mmWave radar, which recognizes hand gesture in high fidelity [28]. Zhao et al. have recently proposed a human pose reconstruction approach with a customized RF radar. By using the pose extracted from the collocated cameras as supervision signals, their reconstruction network could learn to estimate human skeleton, in both 2D and 3D scenes [29, 30].

8. Conclusion

In this paper, we propose mID, a highly accurate tracking and identification system for smart spaces based on millimeter wave radar. With the aid of a commercial-off-the-shelf millimeter wave FMCW radar, we first obtain sparse point clouds. Then, we extract the point clouds representing human objects and associate them to their historical trajectories. Based on the tracklets, a recurrent neural network is used to recognize their identities, and we trained the network with co-supervision of softmax loss and center loss followed by a Random Forest Classifier for intruder detection. Extensive experimental results show that mID achieves an overall recognition accuracy of 89%, and intruder detection accuracy of 73% among 12 people, which increases when there are fewer people in the dataset. mID also bears approximately 0.16m positioning error. We envision mID as a promising step towards smart home human identification and tracking, for the sparse point clouds adopted in mID are not themselves as privacy sensitive as vision based techniques, and mID can be concealed inside furniture or walls, which can be highly unobtrusive and gain acceptance by smart home users.
References


