Indoor Person Identification Using a Low-Power FMCW Radar

Baptist Vandersmissen*, Nicolas Knudde†, Azarakhsh Jalalvand*,
Ivo Couckuyt†, André Bourdoux§, Wesley De Neve*‡, Tom Dhaene†
*Department of Electronics and Information Systems, Ghent University – imec, Belgium
†Department of Information Technology, Ghent University – imec, Belgium
‡Center for Biotech Data Science, Ghent University Global Campus, Korea
§imec, Belgium

***Abstract*—Contemporary surveillance systems mainly use video cameras as their primary sensor. However, video cameras possess fundamental deficiencies such as the inability to handle low-light environments, poor weather conditions, and concealing clothing. In contrast, radar devices are able to sense in pitch-dark environments and to see through walls. In this paper, we investigate the use of micro-Doppler signatures retrieved from a low-power radar device to identify a set of persons based on their gait characteristics. To that end, we propose a robust feature learning approach based on deep convolutional neural networks. Given that we aim at providing a solution for a real-world problem, people are allowed to walk around freely in two different rooms. In this setting, the IDRad dataset is constructed and published, consisting of 150 minutes of annotated micro-Doppler data equally spread over five targets. Through experiments, we investigate the effectiveness of both the Doppler and time dimension, showing that our approach achieves a classification error rate of 24.70% on the validation set and 21.54% on the test set for the five targets used. When experimenting with larger time windows, we are able to further lower the error rate.**

***Index Terms*—convolutional neural network, feature learning, gait classification, indoor sensing, low-power radar, micro-Doppler, person identification**

I. Introduction

Automatic awareness and smart sensing of the environment is a crucial property of future surveillance systems. Modern systems widely use video cameras to collect information from their surroundings. However, despite the significant advances in picture quality and a sharp price drop in recent years, cameras possess fundamental deficiencies such as being unable to handle sudden light flashes or to record in low-light scenarios or poor weather conditions. In addition, the unrestrained use of cameras is subject to controversy when operating in privacy-sensitive areas. In contrast, a radar device preserves visual privacy while being unaffected by weather or lighting conditions. Moreover, it allows for through-the-wall sensing and it can deal with face-concealing clothes. Therefore, radar technology seems to become an indispensable alternative or complementary sensor for a large set of applications.

A radar device transmits an electromagnetic signal over a certain line of sight (LOS). The reflection of the targets moving in the LOS contains information about their speed as a result of the Doppler effect. In addition, separately moving parts are characterized by their own Doppler signal. Most often, the superposition of all these Doppler signals is summarized in a so-called micro-Doppler (MD) signature [1].

The rich structure of an MD signature is used as input for complex radar-based solutions in a wide array of studies. These can range from differentiating among pedestrians, cyclists, and cars to recognizing the specific action a person is performing [2], [3]. In this study, we go one step beyond action recognition by proposing a novel approach for indoor identification of individual humans based on their gait characteristics. To that end, we leverage state-of-the-art tools in the area of low-power radar technology and deep machine learning.

This study entails a number of aspects that significantly increase the complexity and novelty compared to existing sparse state-of-the-art literature on person identification [4], [5]. First and foremost, a target is allowed to walk around in a free and spontaneous way, solely limited by the boundaries of the room. This is in stark contrast with existing studies that only allow walking directly towards or away from the radar, or limit the walking behaviour by using a treadmill. As a result, the models require robustness against differences in walking direction, short stops, and turns, and are thus better suited for deployment in real-life scenarios. Secondly, we prioritize on the use of power-efficient and compact devices that are tailored to use in a smart home environment. Therefore, a low-power Frequency Modulated Continuous Wave (FMCW) radar is used, resulting in a Signal-to-Noise Ratio (SNR) of $8\,\mathrm{dB}$ on average. The combination of such radar with the low radar cross section of approximately $0.5\,\mathrm{m}^2$ [6] of a human and a highly reflective indoor recording environment results in noisy MD signatures and adds to the complexity of the challenge. Finally, a feature learning approach is employed based on deep machine learning that allows for the creation of robust models that are invariant to the exact radar placement and room setup.

Due to the lack of a large, publicly available and realistic indoor data set recorded with a low-power radar, we constructed the IDRad data set (IDentification with Radar data). This dataset was used to train and evaluate the proposed Convolutional Neural Network (CNN)-based models. To summarize, the main contributions of our research effort are as follows:

- we propose robust classification models that are independent of radar placement and room setup while allowing for spontaneous walking, closely mimicking realistic

identification scenarios,
- by employing deep convolutional neural networks, we consider automatic learning of valuable features based on the collected data, as opposed to a limited number of context-specific and hand-engineered features,
- an extensive and intuitive in-depth analysis is performed on the proposed processing pipeline with respect to the model accuracy,
- we release the IDRad data set in order to facilitate future research and benchmarking.

The rest of the paper is organized as follows: Section II briefly lists related work in the area of radar signal processing for classification purposes. Section III and Section IV describe the principles of micro-Doppler and convolutional neural networks, respectively. In Section V, we subsequently explain the proposed approach. Section VI consists of a description of the experimental setup used to validate the proposed approach and Section VII contains an in-depth discussion of our experimental results. Finally, we conclude the paper and suggest some directions for future research in Section VIII.

II. Related Work

The employment of radar as a sensor has been extensively investigated in the signal processing domain. In this section, we provide a concise discussion of a number of related studies, mainly targeting action classification and person identification.

A large number of radar studies focus on the automatic recognition of multiple actions performed by humans. In these cases, a set of distinct actions is listed and a model is built that attempts to recognize these actions. Use cases range from security applications trying to detect violent intents [7], [8] to elderly monitoring applications that attempt to detect walking behavior or falling people [9]–[11]. In [12], manual feature engineering in combination with a support vector machine (SVM) is applied, achieving over 90% test accuracy on seven different actions. The actions under consideration consist of typical human practices such as *walking*, *running*, and *sitting*, but also practices that hint at violent behavior such as *boxing* and *walking while holding a stick*.

The authors of [12] also applied a CNN-based deep learning approach to the same data set, achieving a similar test accuracy [3], which demonstrates the potential of a feature learning approach. In [13], transfer learning is used to classify human aquatic activities. In particular, the authors started from a CNN pretrained on the ImageNet data set and subsequently fine-tuned the weights based on MD data, concluding that a pretrained CNN performs considerably better than a CNN trained from scratch. The authors of [2], [14] utilize an auto-encoder to automatically learn features from MD signatures. In [2], the authors apply an extreme learning machine (ELM) to differentiate among *pedestrians* and *cyclists*, while in [14], a softmax regression classifier is used to make a distinction between four actions, including *falling* and *bending*.

A significant amount of work has been done in the domain of identifying individual persons based on their rhythmical motion of walking, with the main focus on video images as input. In [4], the authors consider MD signatures to identify individual persons. Thirteen subjects, seven males and six females, walk on a treadmill positioned in front of the radar. Based on k-means clustering and k-nearest neighbors (k-NN) classification, an accuracy of 100% is achieved on identifying the individual humans. They also report an accuracy of 92.4% for the task of gender classification. Tahmoush *et al.* [5] report results for recognizing eight persons using a k-NN classifier and two hand-engineered features, namely the stride and torso line of the subjects. In [15], Gaussian mixture models (GMM) are used to identify individual persons and to differentiate among male and female subjects based on hand-engineered features. A total of 20 recordings of 30 subjects were used to train and test the models developed. Similarly, in [16], eight individuals are identified using GMMs, obtaining over 90% accuracy.

This work aims at improving upon the existing state-of-the-art in person identification by introducing a novel approach that focuses on an uncontrolled scenario, allowing targets to freely walk around. This results in models that are more robust against changes in environmental conditions.

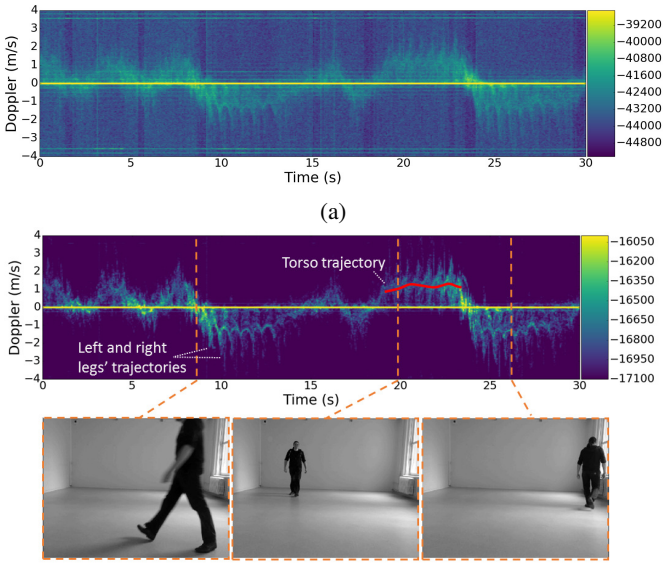
III. Micro-Doppler

The bulk motion of a radar target moving at constant speed induces a constant Doppler frequency shift. However, in addition to the core translation of the target, multiple smaller moving parts result in micro-motion dynamics. These dynamics induce Doppler modulations on the echoed signal, referred to as the micro-Doppler effect. The different moving parts might induce a frequency modulation on the returned signal that results in sidebands around the Doppler frequency shift of the target [1]. The micro-Doppler map can be seen as the power reflected as a function of the speed of the reflector.

In this work, a $77\,\mathrm{GHz}$ Frequency Modulated Continuous Wave radar is used. FMCW radars have the advantage that they can be produced at a low cost, while at the same time being relatively power efficient. Unfortunately, this power efficiency usually comes at the cost of having a low SNR of on average $8\,\mathrm{dB}$, which is one of the challenges faced in this study. An example of such MD signature is shown in Fig. 1a, captured on $30\,\mathrm{s}$ of one person walking in a room. The y-axis in this figure represents the Doppler dimension, also referred to as Doppler channels throughout this paper, while the x-axis represents the time dimension. The zero Doppler channels contain the reflections of all static objects in the room and thus result in high reflected power. In Fig. 1b, the same MD signal is processed so to better demonstrate the structure of the MD signature. To that end, the informative signal is strengthened by removing noise in the range-Doppler domain. This is done by thresholding values below $-45\,\mathrm{dB}$ (cf. Section V-A). Furthermore, the reflected power of the static objects is decreased to better expose the characteristics of a signature caused by a walking person. In this case, the different distinguishable signals represent the body, the arms, and the legs swinging.

IV. Deep Neural Networks

Deep learning or hierarchical learning is a subfield of machine learning that aims at the automatic construction of



(a)

(b)

Fig. 1: MD signature of one person spontaneously walking in a room: (a) the raw MD signature and (b) the thresholded MD signal together with a link to three visual snapshots of a person turning, walking towards, and walking away from the radar, respectively. The color scale shows the accumulated power levels (in dB) after summing over each range-Doppler map.

tailored features based on a stack of nonlinear operations. In particular, algorithms in the field of deep learning aim at automatically creating feature hierarchies, typically through the use of multi-layered feed-forward neural networks (FFNN). Such a FFNN consists of a chain of functions that allow the learning of increasingly complex concepts by stacking many simpler functions:

$$f(\boldsymbol{x}) = f^L(f^{L-1}(\dots f^1(\boldsymbol{x}))), \tag{1}$$

$$f^\ell(\boldsymbol{x}) = \sigma(\mathbf{W}^\ell \boldsymbol{x} + \mathbf{b}^\ell), \qquad \forall \ell \in \{1..L\}, \tag{2}$$

where $\boldsymbol{x}$ represents an input vector, L denotes the number of layers in the network, σ represents a piece-wise nonlinear function, and $\mathbf{W}^\ell$ and $\mathbf{b}^\ell$ describe the layer-specific weights and biases, respectively.

The piece-wise nonlinear operation σ is commonly chosen to be the rectifier linear unit (ReLU) [17], and where this function is defined as follows: $\text{ReLU}(x) = \max(0, x)$.

Even though the concept of deep neural networks already exists for several decades, it has regained considerable attention since a widely published breakthrough in the ImageNet Large-Scale Visual Recognition Competition (ILSVRC) in 2012 [18]. Since then, further empirical evidence has shown the excellent performance of deep learning in several application domains, including image classification, speech recognition, and natural language processing. The recent success of deep learning can be attributed to the current availability of large data sets and cheap computational power, as well as a number of algorithmic advances and a culture of open innovation.

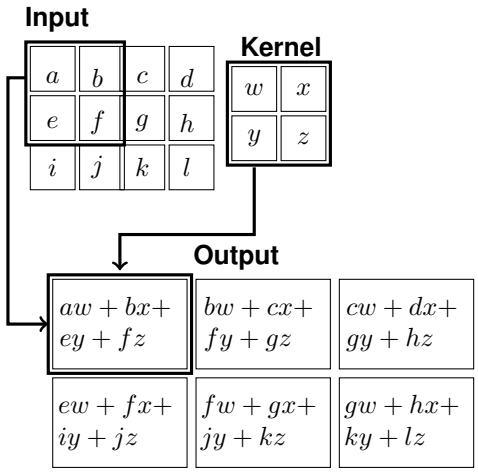


Fig. 2: Example of a two-dimensional convolutional operation. A 2×2-sized kernel is convolved over a 3×4-sized input with zero padding. The operation of each element is exactly described in the resulting output feature map.

In this work, we focus on deep convolutional neural networks (DCNNs). These artificial neural networks make use of neurons that are only locally connected and that share weights. This means that convolutional filters work on small local receptive fields of input data in a sliding-window fashion. This specialized kind of neural network has a grid-like topology. Different filters evolve to become specific feature detectors, for instance ranging from low-level color and edge detectors in early layers to high-level object detectors in later layers. The essential difference with a standard feed-forward neural network is the use of convolutions instead of plain matrix multiplications.

In Fig. 2, an example of a convolution is shown with a kernel of size 2×2 and stride 1 [1]. The mathematical operation of such a convolution is defined as follows:

$$\mathbf{S}_{ij} = (\mathbf{X} * \mathbf{K})_{ij} \tag{3}$$

$$= \sum_m \sum_n \mathbf{X}_{i+m,j+n} \mathbf{K}_{mn}, \tag{4}$$

with $\mathbf{S}$ denoting the resulting feature map, $\mathbf{X}$ a two-dimensional input, and $\mathbf{K}$ a kernel $\in \mathbb{R}^{m \times n}$. Compared to a regular FFNN, Equation 2, is modified as:

$$f_j^\ell(\mathbf{X}) = \sigma(\mathbf{X} * \mathbf{W}_j^\ell + \mathbf{b}_j^\ell), \forall \ell \in \{1..L\}, \tag{5}$$

with f_j^ℓ depicting the j-th feature map of layer ℓ.

Besides weight sharing, a dimension reduction technique known as pooling is applied to effectively mitigate the number of parameters and the data size. By averaging or maximizing the response of $n \times m$ cells, essential information is preserved, while the data size is reduced. In Fig. 3, a max pooling operation of size three and stride two is conceptually displayed. It can be noted that a pooling operation results in translation invariance. Indeed, the essential information is preserved, regardless of its exact cell location.

[1] From a strict point-of-view, we are dealing with a cross-correlation, as the kernel is not flipped.

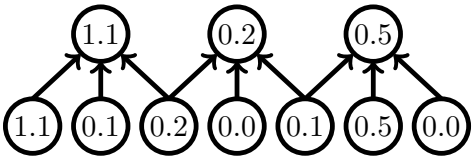


Fig. 3: Example of a one-dimensional max pooling operation with size three and stride two. A seven-dimensional input vector is reduced to a three-dimensional feature vector by selecting the maximum value over a window of three neurons and subsequently shifting by two neurons.

V. Proposed Approach

The goal of this research is to identify people based on the MD signatures provided by a low-power FMCW radar. The key research question we try to answer is whether such micro-Doppler features allow characterizing individual humans in a realistic scenario. The scenario under consideration is defined as an indoor living space, in which people are allowed to freely walk around in every direction possible. In Fig. 4, a schematic overview depicting the proposed approach is given.

In this section, we discuss the different preprocessing steps and machine learning algorithms used to address the aforementioned question.

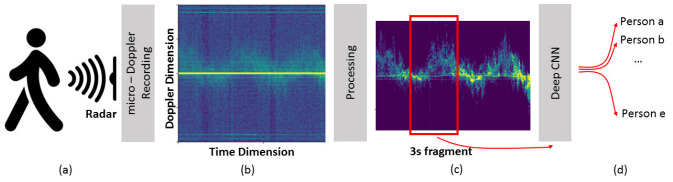


Fig. 4: Schematic overview of the proposed approach: (a) a single target is captured by a low-power radar while walking in a room, (b) the recorded raw signal is computed into an MD signature, (c) the MD signature is processed to reduce noise and retain only essential information, and (d) at each time step, a $3\,\mathrm{s}$ MD fragment is fed to a CNN which predicts probabilities for each target.

A. Preprocessing

In this work, an FMCW radar device produced by INRAS [19] is used. This millimeter-wave radar allows working with a significant amount of bandwidth of $1.5\,\mathrm{GHz}$ and a high frequency of $77\,\mathrm{GHz}$, resulting in an excellent range and velocity resolution of $10\,\mathrm{cm}$ and $2\,\mathrm{cm/s}$, respectively. The device is set up in Single Input Single Output (SISO) mode and the recording parameters are given in Table I. According to our experiments, the SNR of the data provided by this radar varies from $10\,\mathrm{dB}$ for targets within a range of $1\,\mathrm{m}$ to $7\,\mathrm{dB}$ for targets located at around $8\,\mathrm{m}$.

The MD signature is calculated by first determining the range-Doppler map using a two-dimensional Fourier transform. Subsequently, the absolute value of the signal is converted to decibels (dB) and summed over the range dimension. This MD signature is referred to as a raw signal throughout the remainder of this paper.

TABLE I: Recording parameters of the FMCW radar. The range and velocity resolution of $10\,\mathrm{cm}$ and $2\,\mathrm{cm/s}$, respectively, allow for fine-grained capturing of detailed movements.

Waveform Parameters		Sensing Parameters	
Center freq.	$77\,\mathrm{GHz}$	Range resolution	$10\,\mathrm{cm}$
Chirp bandwidth	$1.5\,\mathrm{GHz}$	Velocity resolution	$2\,\mathrm{cm/s}$
Chirp duration	$256\,\mu\mathrm{s}$	Ambiguous range	$38.4\,\mathrm{km}$
Sampling freq.	$2\,\mathrm{GHz}$	Ambiguous velocity	$13.68\,\mathrm{km/h}$

To mitigate the significant amount of noise (*cf.* Fig. 1a), a thresholded variant of the MD signature is computed and investigated in Section VI. Specifically, a lower threshold in the range-Doppler domain filters out noise after subtraction of the maximum value. This value is derived from Fig. 5, which shows a normalized histogram of a random set of range-Doppler maps in both cases of either an empty or a non-empty room. For both histograms, we have filtered the influence of the reflected power of all static objects by removing the zero Doppler channels. From this figure, it can be derived that the skew-normal distributed noise can be filtered by setting a lower threshold of $-45\,\mathrm{dB}$.

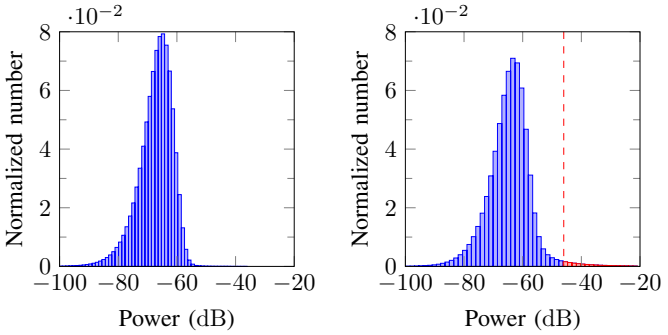


Fig. 5: Normalized histogram of $1,000$ (a) empty and (b) non-empty range-Doppler maps. Figure (a) clearly shows the presence of noise that is skew-normal distributed and with a relative power below $-45\,\mathrm{dB}$. In figure (b), the perceived signal resulting from walking activity is highlighted.

The final MD signatures (both raw and thresholded) contain 256 Doppler channels per time step, representing speeds from $-3.8\,\mathrm{m/s}$ to $3.8\,\mathrm{m/s}$. It was visually observed that three middle Doppler channels represent all non-moving objects. The time dimension is represented by the frequency for which a range-Doppler map is produced by the radar device. In this case, a total of 256 chirps are taken, with each chirp having a duration of $256\,\mu\mathrm{s}$, thus resulting in approximately 15 frames per second (FPS). Throughout this paper, a frame represents one time step in the MD signature and consists of 256 Doppler channels. In Section VI, both the time and Doppler input dimension are extensively investigated.

B. Neural Network Architecture

Two key properties of convolutional neural networks make them appealing for the task of identifying persons based on low-SNR MD signatures: (1) the capability of building models

that are robust against noisy data and (2) the learning of valuable features in an automatic way.

In this study, the MD signatures are represented as two-dimensional spatial structures that are fed to a deep convolutional neural network. We assume that the features necessary to identify multiple persons can be learned from short MD fragments. The need for large amounts of data prevents the use of more modern deep networks to boost performance such as for example the inception networks [20] or residual networks [21].

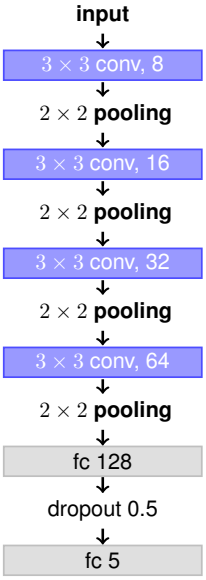


Fig. 6: Schematic diagram of the neural network architecture. The first convolutional layer consists of eight 3×3 filters, followed by a pooling layer with non-overlapping 2×2 cells. This sequence, which is repeated four times with an increasing number of filters, is followed by a fully-connected two-layer network.

Fig. 6 shows the conceptual architecture of our network. The network structure was carefully designed by experimenting with a large number of hyperparameters such as the number of layers (convolutional, pooling, or fully-connected), the number and size of filters, etc. The resulting network consists of four convolutional layers and two fully connected layer, with the number of output neurons dependent on the number of persons present in the data set. A select number of 3×3 filters is used to avoid rapid overfitting. Each convolutional layer and the first fully connected layer make use of an Exponential Linear Unit (ELU) non-linearity operation. This non-linear operation is defined as follows:

$$\mathrm{ELU}(x) = \begin{cases} x, & \text{if } x > 0 \\ \alpha(\exp(x) - 1) & \text{if } x \leq 0 \end{cases}, \tag{6}$$

with $x \in \mathbb{R}$ representing the input and α a predefined parameter greater than zero. Compared to other non-linearities, ELU non-linearities possess improved learning characteristics [22]. Moreover, the negative values that are part of the range of an ELU allow pushing mean unit activations closer to zero. The latter is similar to batch normalization, but coming with a computational complexity that is lower. The last fully-connected layer uses a softmax operation to produce outcome probabilities for each target class.

The input dimension is sequentially reduced by four 2×2 pooling layers. This network in total consists of $286,408$ trainable weights for a default input of 256×45, resembling 256 Doppler channels and 45 time steps (i.e., three seconds) of data.

TABLE II: Physical characteristics of the persons who participated in collecting the data set.

Target ID	Age	Height	Weight
1	23	178 cm	82 kg
2	32	185 cm	99 kg
3	28	180 cm	79 kg
4	24	182 cm	60 kg
5	28	179 cm	71 kg

VI. Experimental Setup

In this section, we describe the characteristics of the IDRad data set constructed.

A. IDRad Data Set

In order to create a realistic data set, we considered different rooms, encouraging people to walk around spontaneously in any possible direction. Each person was recorded individually, hence, no recording contains multiple persons present at the same time. We have captured our data over multiple days and rooms, so to take into account the effect of contextual influences like different moods, clothing, shoes, etc. By not focusing on a single recording per user in a single room, we explicitly aimed at developing a robust system that is capable of dealing with different environments. In this study, we aimed at simulating a challenging household setting by recording five different persons.

In a first stage, we recorded the random walking of five persons in a room for five consecutive minutes, and the same five people were again recorded in the same room for 15 consecutive minutes two weeks later. Table II lists some basic information about each target. All our subjects are males between 23 and 32 years old with comparable postures. Their weights range from $60\,\mathrm{kg}$ to $99\,\mathrm{kg}$ and their heights range from $178\,\mathrm{cm}$ to $185\,\mathrm{cm}$.

The whole training data set contains 20 minutes of MD signatures per person. As mentioned before, each target is recorded in a continuous matter. Therefore, each recording also contains other movements than regular walking, including turns, short stops, and accidental moves. A video camera is also used for simultaneously recording the walking targets, mainly for easing the analysis of the MD signatures. The video data were not used to train our models in this study.

In a second stage, a different room was used to create the validation and test set. Again, the recordings for both the validation and the test set contain five minutes of continuous walking for all targets and were created with two weeks in between. In Fig. 7, an image sequence visualizes a three-second walking fragment. We can observe that the continuous

Fig. 7: Image sequence showing three seconds of walking.

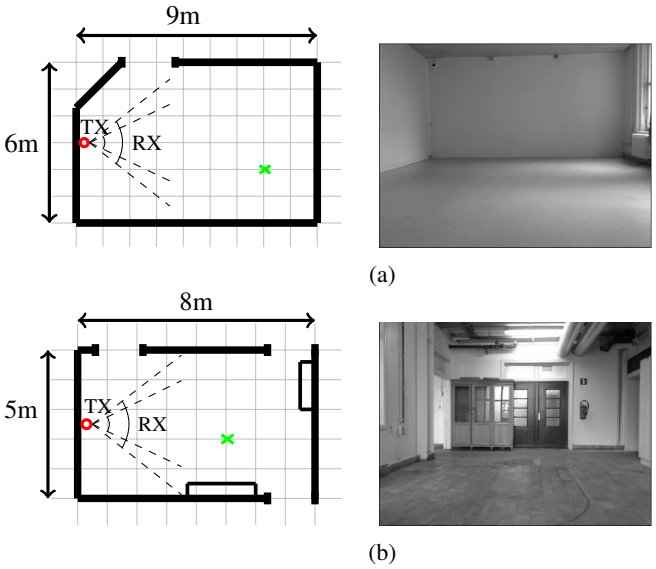


Fig. 8: Conceptual and photographic visualization of (a) the training room and (b) the validation/test room. The radar position is represented by a red circle. The line of sight is indicated by dotted lines with the receiving beamwidth (RX) covering $76.5°$ and the transmitting beamwidth (TX) covering $51°$. The green cross denotes a possible target.

recording possesses a significant amount of variation, including walking parallel to the radar.

Fig. 8a and 8b show a conceptual and visual representation of the training and validation/test room. We would like to emphasize that the walls of this building are built based on a metallic construction framework, which resulted in reflections and a considerable amount of noise in the recordings. In addition, the presence of a metallic and wooden closet together with an open ceiling with metallic tubes in the test room produced ghost targets coming from multi-path reflections.

In order to facilitate further research on this topic, the IDRad data set is made publicly available [2].

B. Statistical Analysis

Given that the FMCW radar records range-Doppler maps with a speed of around $15\,\mathrm{FPS}$, the training set contains $67,625$ frames, while the validation and test set contain $22,535$ frames each. One frame represents one time step in the MD signature and is depicted by 256 Doppler channels (i.e., the sum over all range channels per Doppler channel of one range-Doppler map). For both the validation and test set, we generate samples by cutting up the MD signal into windows with a length of 45 frames (representing $3\,\mathrm{s}$ of data) with an overlap of $1\,\mathrm{s}$, thus resulting in $1,490$ samples. Throughout this paper, we report the error rate, which is the ratio of wrongly classified samples to a total of $1,490$ samples, as a measure to compare the obtained results.

Fig. 9 shows the average walking speed of each target in all data sets. The speed per target is computed by averaging the Doppler channels linked to the maximum power present in each range-Doppler map after removing the zero-Doppler channels. It can be noted that average speed in itself is potentially a relevant feature as it is dimly linked to the walking behavior of a person. However, it is clear that speed in itself is not sufficient to solve the challenge originally put forward since multiple targets have similar walking speeds in all data sets. Moreover, the speed of a target is naturally varying due to the relatively small rooms and thus the need for turns. This is shown by the large standard deviation of the speed of each target, which can be linked to large variations in walking speed.

Fig. 10 displays the standard deviation of the reflected power per Doppler channel for raw and thresholded MD signatures. It is clear that the zero-Doppler (indices 127 – 129) channels, representing the static objects, contain a lower amount of variation compared to their surrounding channels. Moreover, we can conclude that most information resides in the middle third Doppler channels and that the thresholded signals contain more overall variance. In Section VII-B, the removal of the static and outer Doppler channels is investigated and the effect on the accuracy of the model is reported.

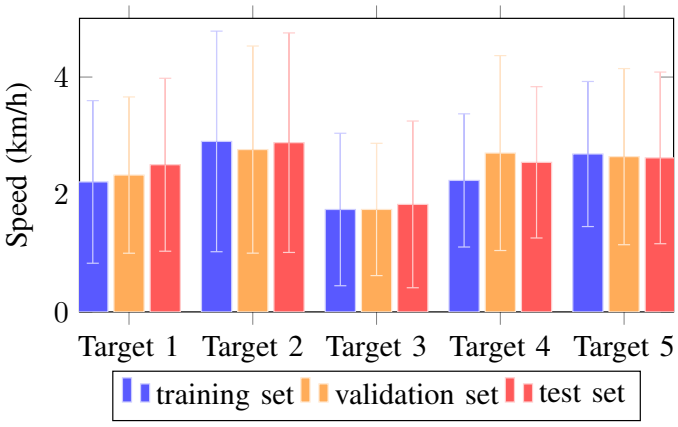


Fig. 9: Average target speed per set.

[2]The data set is publicly available at https://www.imec-int.com/IDRad.

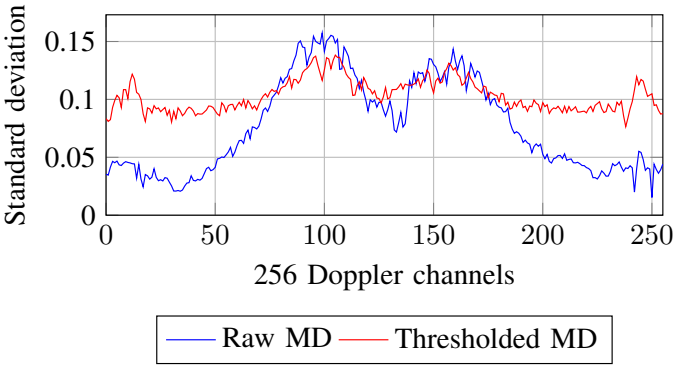


Fig. 10: Standard deviation of raw and thresholded Doppler channels over all frames of the training set.

C. Learning

We trained our models on a GeForce GTX 980 graphics card using the Theano[3] and Lasagne libraries[4]. Mini-batches of size 64 were used and the parameters were learned based on the Adam optimizer, using a learning rate of 0.0001 and 0.0005 L_2 regularization. This learning rate was adaptively lowered based on the cross-entropy loss of the validation. The model was trained for around 300 epochs and required 15 minutes on average to converge.

A major challenge in our research was to prevent our models from overfitting on the training set. To that end, we augmented our data set by randomly shifting over the time dimension. All samples were locally normalized to the range $[0, 1]$. We experimented with mirroring of the MD signatures, but observed no noticeable improvements.

VII. Experimental Results

A. Analysis of Time Dimension

In the process of learning valuable features from given input data, it is crucial to determine the relevant input dimensions. In what follows, we investigate how many time steps and which Doppler channels optimize the performance of the identification model. Selecting the essential information channels for both dimensions (i.e., time and Doppler) is a determining factor to prevent overfitting of the network. In that regard, we analyzed the effect of ranging the time dimension from 5 to 150 frames, while keeping the Doppler dimension fixed to 256. As mentioned above, frames are recorded at an interval of 15 FPS, with the input window range thus varying from $1/3$ s to 10 s. We repeated this experiment for both raw and thresholded MD signatures.

Our results are depicted in Fig. 11. We can note that each number is the result of averaging the output of the experiment three times. It is clear that the model cannot effectively learn valuable features from raw MD signatures when using all 256 Doppler channels as input. This effect is studied in more detail in Section VII-B. In contrast, the results of using the thresholded MD signatures show that there is a clear benefit of adding more time steps to the input. In this case, the error rate ranges from 61.08% for the shortest window to 21.26% when the input consists of 140 time steps, resembling 9.33 s of information. We notice a sharp decrease in terms of error rate in the early phase, when the length of the window ranges from 5 to 45 frames. The improvement becomes less significant for longer fragments. Considering a trade off between short-term predictions and high model performance, we conclude that the use of 45 time frames — resembling three seconds — is optimal.

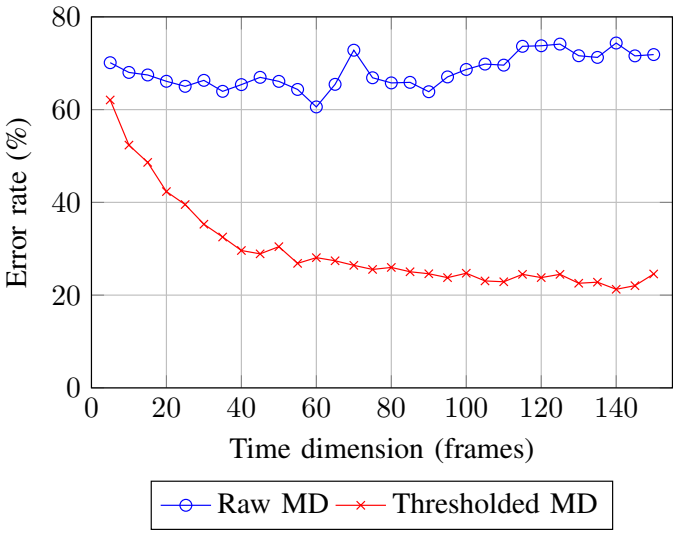


Fig. 11: The error rate as a function of the length of the input window for both the raw and thresholded MD signals.

B. Analysis of Doppler Dimension

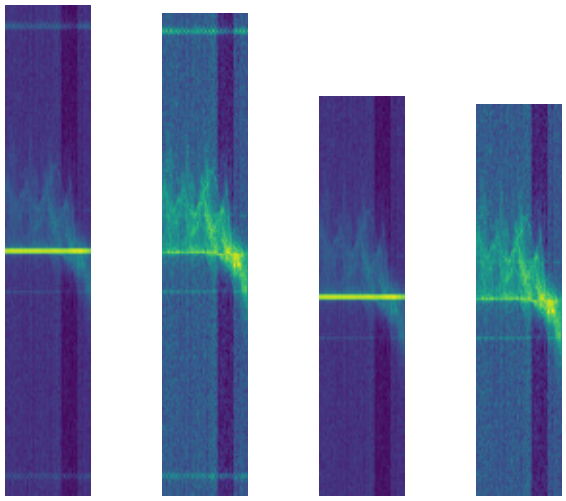
In the previous section, we concluded that a window of 45 frames contains a sufficient amount of valuable information, while still enabling short-term predictions. A second conclusion is related to the fact that the model was unable to effectively learn from the full raw MD signatures. Moreover, in Section VI-B, we statistically observed that some channels contain more variance than others, which can be an indication that they hold more useful information. Therefore, the influence of downsampling the entire Doppler dimension is investigated, as well as removing specific channels. We hypothesize that intelligently reducing the input dimension will decrease the chance of overfitting. To support this hypothesis with experimental data, we use the network described in Section V-B and fix the length of the input window to 45 frames.

We compare the models trained on the original 256 Doppler channels and on a downsampled version by a factor of two and four. The MD signatures are downsampled by linear interpolation. Both the effect of removing the zero-Doppler channels (*Remove Static*) and removing the outer Doppler channels, representing high speeds in both directions (*Remove Outer*), are analyzed. More precisely, in the case of removing the static objects, the three center Doppler channels are removed. For removing outer Dopplers, we empirically decided to eliminate 24 Doppler dimensions at both sides. This significantly reduces

[3]http://deeplearning.net/software/theano/
[4]lasagne.readthedocs.io/

TABLE III: Results for the default CNN model when removing and downsampling certain Doppler dimensions from the input (in %). The left half of the table shows the error rate for the five different targets used, for both the original input and for downsampled versions of the input by a factor of two and a factor of four. The right half of the table shows the error rate for the same input dimensions but when making use of thresholded MD signatures.

	Raw MD Signatures						Thresholded MD Signatures					
	original		factor 2		factor 4		original		factor 2		factor 4	
Input	256×45	66.51	128×45	61.97	64×45	**51.16**	256×45	**28.46**	128×45	30.81	64×45	32.64
Remove Static	253×45	67.07	126×45	57.13	63×45	**53.49**	253×45	28.34	126×45	**27.79**	63×45	33.82
Remove Outer	208×45	64.27	104×45	54.18	52×45	**51.34**	208×45	**31.07**	104×45	35.59	52×45	33.42
Remove Both	205×45	47.76	102×45	**46.35**	51×45	50.74	205×45	**26.65**	102×45	31.70	51×45	33.36



(a) 256×45 (b) 253×45 (c) 208×45 (d) 205×45

Fig. 12: Visualization of an example MD signature fed as input to our model: (a) the original signature, (b) *Remove Static*, (c) *Remove Outer*, and (d) *Remove Both*.

the input dimension, while we hypothesize that this will not influence the effectiveness of the trained models. Fig. 12 shows an example MD signature for each scenario.

In Table III, an overview of the results is given. Again, each number is the result of averaging the output of the experiment three times. We can observe there is a significant advantage to using thresholded MD signatures as compared to using the raw counterparts. Moreover, we can observe that the sensitivity of the network to overfitting on the noise obstructs the learning of advanced features. While hard thresholding removes parts of the signal of a target, the loss in information is clearly outweighed by the benefits of a more robust learning process. When raw MD signatures are used, it is clear that removing both static and outer Doppler channels has a positive effect on the performance of the model. This effect is less pronounced when using thresholded MD signatures as input. The lowest error rate is 26.65% and is achieved by removing both static and outer Doppler channels on thresholded MD signatures.

C. Main Results

In this section, we finalize the model configuration based on our previous analyses. As discussed in Section VII-B, reducing the input dimension by removing the outer and static Doppler dimensions has a small but positive influence on the effectiveness of our approach. Thus, a thresholded MD signature with dimension 205×45 is fed as an input to the CNN model described in Section V-B.

TABLE IV: Error rate for the training, validation, and test set (in %) for a number of target combinations. (*) Trained on the combined training and validation set.

trained on	**training**	**validation**	**testing**
targets 1, 2, 3	1.72	6.71	13.20
targets 1, 4, 5	2.71	30.87	20.81
targets 1, 2, 3, 4, 5	**1.95**	**24.70**	**21.54**
targets 1, 2, 3, 4, 5 (*)	0.09	0.07	15.10

The error rate on all five targets for the validation and test set acquired with our best model setup is listed in Table IV. Two combinations consisting of three targets are shown, together with the results obtained by a trained model for all five targets. It is clear that some targets can be more easily classified than others as the best performing triplet achieves a validation and test error of 6.71% and 13.20%, respectively. However, the combination of *Target 1, 4, and 5* results in a significantly lower effectiveness of 30.87% and 20.81% for the validation set and test set, respectively. Fig. 9 shows that *Target 2* and *Target 3* have the highest and lowest average walking speed, respectively. We assume this enables easier separation between the targets. *Target 4* possesses high variability in average walking speed between its training and validation set (*cf.* Fig. 9), while all three targets have similar walking speeds in the validation set.

The best performing model achieves 24.70% error rate on the validation set and 21.54% error rate on the test set for all five persons. We would like to emphasize that this result is based on a training set that is recorded in a different room and on a different day. To measure the impact of providing information about the room to the training set, we combine both the training and the validation set, and use 80% and 20% of the resulting set of samples for training and validation, respectively. The error rate of the test set decreases to 15.10%, showing the advantage of having a larger training set and having more variety in the data. However, we can observe that the relatively small difference shows that the initial learning of the model is, to some extent, already robust against different environmental conditions

To measure the influence of the proposed feature learning approach, we compare the obtained results with a traditional

dimension reduction technique, namely Principal Component Analysis (PCA) in combination with a Support Vector Machine (SVM) and a Random Forest (RF) classifier. To that end, we apply PCA on the thresholded Doppler dimensions and reduce the 256 channels to 9 components. The input samples are represented by a 405-dimensional vector, containing 9 Doppler components for each of the 45 frames. The hyper parameters of both classifiers are optimized by means of a grid search performed over a range of values, selecting the model with the best outcome for the validation set. Table V lists the results for both combinations. The deep CNN substantially outperforms both PCA-based methods by an absolute margin of 17% on the test set.

TABLE V: Error rate on the validation and test set (in %) for the deep CNN- and PCA-based methods.

method	**validation**	**testing**
PCA plus RF	48.86	38.59
PCA plus SVM	49.20	38.52
deep CNN	**24.70**	**21.54**

D. In-Depth Analysis

In this section, we analyze the results obtained with the above described network, achieving 21.54% on the test set. First, a comparison is given of the accuracy between the different targets. Second, we analyze the focus of the network when classifying MD signatures. Finally, a representative example is shown, demonstrating the effectiveness of the trained model.

In Table VI, the normalized confusion matrix is displayed for the test set, obtained with the original training and validation set (for all five targets). As described in Section VI-B, the test set consists of 1,490 samples, equally distributed over the five targets. We can observe that, on the one hand, distinguishing *Target 1* and *4* is more difficult with 68.12% and 70.47% accurately classified samples, respectively. On the other hand, *Target 2*, *3*, and *5* achieve high scores of 91.28%, 78.19%, and 84.23% correctly classified samples, respectively. According to this table, mainly *Target 1, 4*, and *5* are confused among each other. When inspecting the MD signatures of the training set of both *Target 1* and *Target 4*, we can observe that they show great variability while walking and that their corresponding MD signatures contain relatively more noise.

Fig. 13 shows the activation feature maps of the second to last convolutional layer for two randomly selected samples. The first image of each row represents the original input MD sample. Different feature maps contain different types of information, highlighting specific parts of the signal. Specifically, certain feature maps focus on the entire shape of the MD signature, while others highlight the torso or body part trajectories. This insight shows that the network has learned a wide range of discriminative features that steer the decision of the identification prediction.

Finally, we study the robustness of the network when making predictions for a contiguous MD segment of 47 seconds. More precisely, one fragment of ten seconds per target is randomly selected from the validation set and the resulting

TABLE VI: Confusion matrix for the test set (in %).

True Label \ Predicted Label	Target 1	Target 2	Target 3	Target 4	Target 5
Target 1	**68.12**	2.35	3.36	12.08	14.09
Target 2	3.69	**91.28**	1.34	1.34	2.35
Target 3	10.07	2.01	**78.19**	8.05	1.68
Target 4	4.7	0.00	8.39	**70.47**	16.44
Target 5	7.72	2.35	2.68	3.02	**84.23**

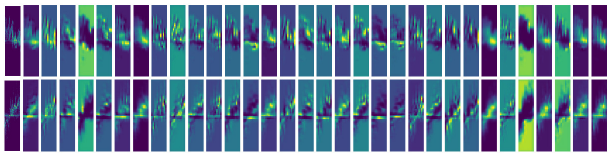


Fig. 13: Resulting 32 feature maps of the second to last convolutional layer for two randomly selected MD fragments.

fragments are then concatenated, again using a random order. Each time step, an MD signature with a length of 3 s is fed to the CNN model and probabilities for each target are returned. The trained model either consistently predicts the correct output or is confused when a transition between two targets happens. In the latter case, the MD signature fed to the model contains gait aspects of two different targets, which explains the confusion of the model. Fig. 14 shows the concatenated MD signal, together with the corresponding probabilities for each target, the predicted target, and the true target.

E. Averaging Predictions

To boost the effectiveness of the trained model, we analyze the effect of combining multiple predictions over a longer time period. This also corresponds to the more practical usage of person identification, in wich persons are monitored for longer periods of time. In Fig. 15, the result of averaging predictions for an increasing number of seconds is displayed. The results for the model trained on just the training set are shown, together with the model trained on the combined training and validation set (*cf.* Table IV).

The length of the window over which predictions are averaged ranges from 3 s to 30 s. The number of predictions used depends on the size of this window. A window of 4 s results in an average over 16 predictions, while a window of 30 s results in an average over 406 predictions. We can observe a sharp decrease in error rate on the test set when more than one prediction can be used. For windows of more than 25 s, a minimum error rate of 0% can be achieved.

F. Intruder Detection

An additional use case of person identification is the construction of an automatic intruder detection system that alerts when an unknown person enters a monitored area (for security, theft prevention, and so on). Such a system can be trained on a set of known people living together (e.g., a family), and where

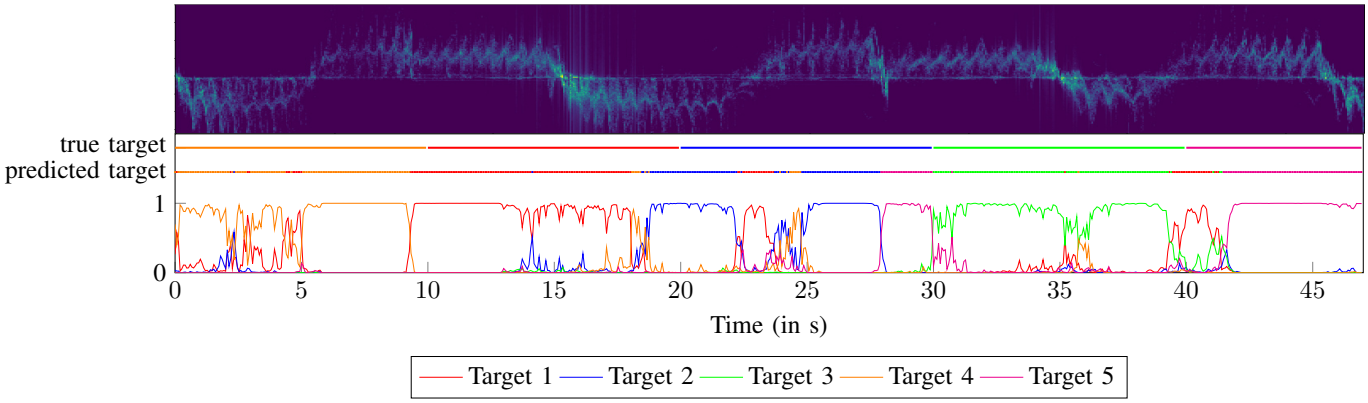


Fig. 14: Experiment showing a 47s MD signature consisting of five targets and its corresponding predictions per frame.

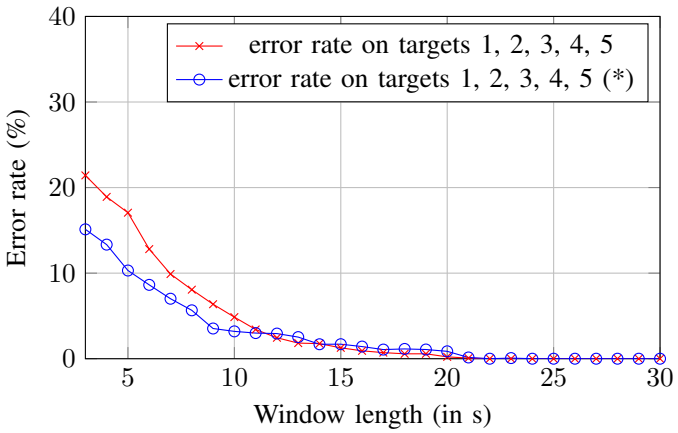


Fig. 15: Classification performance as a function of the length of the window used to average a number of predictions.

these people can be accurately recognized using our method. In this experiment, we test the possibilities of our method tackling such a use case. To that end, we recorded test data for a sixth target using the same procedure as in Section VI-A for a total of five minutes in the validation/test room (*cf.* Fig. 8b). The original model, trained on five different targets, is used to predict the outcome of a given sample. The uncertainty of the predictions is modeled and the prediction for an *intruder target* is based on this uncertainty value. Fig. 16 shows the concatenation of random 10 s fragments of *Targets 2, 3*, and *5*, originating from the test set, together with two 10 s fragments of the *intruder target*. Predictions are made frame per frame, but using a window size of 4 s (*cf.* Section VII-E) to predict the correct outcome. The variance over all probabilities per class is computed, comparing the maximum value to a fixed threshold in order to decide whether it is an unknown target (i.e., an intruder). Fig. 16 shows high uncertainty among the different class probabilities when the unknown target appears. Wrong intruder alerts are given when the MD signature transitions from one known target to another.

VIII. Conclusions and Future Work

In this work, a feature learning approach towards the automatic identification of spontaneously walking persons in different rooms was proposed. To that end, we constructed the IDRad data set of five targets that is split into a training, validation, and test set consisting of 20 minutes, 5 minutes, and 5 minutes of data per target, respectively. A deep convolutional neural network was applied to automatically extract features from the processed micro-Doppler signatures and compute accurate probabilities over five targets. An in-depth investigation was conducted of multiple input configurations, leading to the conclusion that an optimal input signal can be obtained by cutting out the outer 24 Doppler channels and the 3 static Doppler channels of a thresholded micro-Doppler signature of 3 s long. With this input, we achieved an error rate of 24.70% on the validation set and an error rate of 21.54% on the test set for five different targets. We validated the effectiveness of our feature learning approach by comparing it to a combination of principal component analysis with a support vector machine and random forest. The deep convolutional neural network significantly outperformed these approaches by an absolute margin of 17%. When experimenting with larger time windows, we were able to further lower the error rate to 0% for above 25 s windows. Moreover, the approach was extended in order to create an intruder detection system that alerts when an unknown person enters a certain area. To summarize, we successfully built a solution to automatically identify persons in an indoor and realistic setting solely based on gait characteristics recorded with a low-power FMCW radar.

To continue this work, we plan to investigate the necessary improvements to enable a wide set of applications. This primarily involves identification of multiple people walking in the same room, while allowing these rooms to be more cluttered. A key challenge in identifying more than one person is isolating the different targets in the range-Doppler(-azimuth) domain. Based on an effective tracking algorithm, one could separate and deduct the MD signatures of each individual person. Indeed, the separation and clear isolation of each MD signature is a non-trivial task and inquires clear separable targets. In order to improve the robustness of the model in cluttered rooms, we will analyze and implement advanced machine learning techniques to alleviate the impact of multi-path reflections and shadowing effects in the range-Doppler domain. Furthermore, we will increase the size and complexity

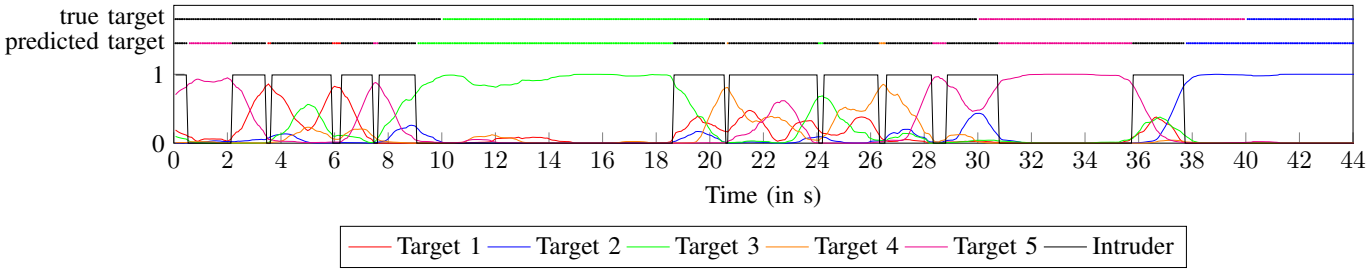


Fig. 16: Experiment on modeling the uncertainty of the network so to enable the prediction of an intruder class.

of our dataset, which will enable true end-to-end learning and mitigate the effects of overfitting. In addition, one can investigate the use of multiple radar setups to incorporate information coming from different recording angles with the aim of significantly increasing the ability to record the gait characteristics of a human in a fine-grained manner. Finally, time-dependent models such as Long Short-Term Memory (LSTM) networks will be investigated so to be able to fully exploit the temporal information.

ACKNOWLEDGMENT

The research activities described in this paper were funded by Ghent University, imec, and the Fund for Scientific Research-Flanders (FWO-Flanders).

REFERENCES

[1] V. C. Chen, F. Li, S. S. Ho, and H. Wechsler, "Micro-doppler effect in radar: phenomenon, model, and simulation study," IEEE Transactions on Aerospace and Electronic Systems, vol. 42, no. 1, pp. 2–21, Jan 2006.

[2] K. N. Parashar, M. C. Oveneke, M. Rykunov, H. Sahli, and A. Bourdoux, "Micro-doppler feature extraction using convolutional auto-encoders for low latency target classification," in 2017 IEEE Radar Conference (RadarConf), May 2017, pp. 1739–1744.

[3] Y. Kim and T. Moon, "Human detection and activity classification based on micro-Doppler signatures using deep convolutional neural networks," IEEE Geoscience and Remote Sensing Letters, vol. 13, no. 1, pp. 8–12, Jan 2016.

[4] G. Garreau, C. M. Andreou, A. G. Andreou, J. Georgiou, S. Dura-Bernal, T. Wennekers, and S. Denham, "Gait-based person and gender recognition using micro-doppler signatures," in 2011 IEEE Biomedical Circuits and Systems Conference (BioCAS), Nov 2011, pp. 444–447.

[5] D. Tahmoush and J. Silvious, "Radar micro-doppler for long range front-view gait recognition," in 2009 IEEE 3rd International Conference on Biometrics: Theory, Applications, and Systems, Sept 2009, pp. 1–6.

[6] V. C. Chen, Ed., Radar Micro-Doppler Signatures: Processing and Applications, ser. Radar, Sonar & Navigation. Institution of Engineering and Technology, 2014.

[7] F. Fioranelli, M. Ritchie, and H. Griffiths, "Classification of unarmed/armed personnel using the netrad multistatic radar for micro-doppler and singular value decomposition features," IEEE Geoscience and Remote Sensing Letters, vol. 12, no. 9, pp. 1933–1937, Sept 2015.

[8] M. Ritchie, F. Fioranelli, H. Borrion, and H. Griffiths, "Multistatic micro-doppler radar feature extraction for classification of unloaded/loaded micro-drones," IET Radar, Sonar and Navigation, vol. 11, no. 1, pp. 116–124, January 2017.

[9] L. Liu, M. Popescu, M. Skubic, M. Rantz, T. Yardibi, and P. Cuddihy, "Automatic fall detection based on doppler radar motion signature," in 2011 5th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth) and Workshops, May 2011, pp. 222–225.

[10] M. Wu, X. Dai, Y. D. Zhang, B. Davidson, M. G. Amin, and J. Zhang, "Fall detection based on sequential modeling of radar signal time-frequency features," in Proceedings of the 2013 IEEE International Conference on Healthcare Informatics, ser. ICHI '13. Washington, DC, USA: IEEE Computer Society, 2013, pp. 169–174. [Online]. Available: http://dx.doi.org/10.1109/ICHI.2013.27

[11] S. Z. Gurbuz, C. Clemente, A. Balleri, and J. J. Soraghan, "Micro-doppler-based in-home aided and unaided walking recognition with multiple radar and sonar systems," IET Radar, Sonar Navigation, vol. 11, no. 1, pp. 107–115, 2017.

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

[13] J. Park, R. J. Javier, T. Moon, and Y. Kim, "Micro-Doppler based classification of human aquatic activities via transfer learning of convolutional neural networks," Sensors, vol. 16, no. 12, p. 1990, 2016.

[14] B. Jokanovic, M. Amin, and F. Ahmad, "Radar fall motion detection using deep learning," in 2016 IEEE Radar Conference (RadarConf), May 2016, pp. 1–6.

[15] K. Kalgaonkar and B. Raj, "Acoustic doppler sonar for gait recoginationn," in 2007 IEEE Conference on Advanced Video and Signal Based Surveillance, Sept 2007, pp. 27–32.

[16] Z. Zhang and A. G. Andreou, "Human identification experiments using acoustic micro-doppler signatures," in 2008 Argentine School of Micro-Nanoelectronics, Technology and Applications, Sept 2008, pp. 81–86.

[17] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in Proceedings of the 27th International Conference on Machine Learning (ICML-10), J. Frnkranz and T. Joachims, Eds. Omnipress, 2010, pp. 807–814.

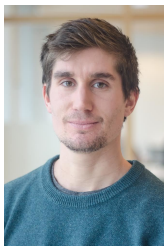
[18] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in Advances in Neural Information Processing Systems 25, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2012, pp. 1097–1105.

[19] "INRAS GmbH," 2017. [Online]. Available: http://www.inras.at

[20] C. Szegedy, S. Ioffe, and V. Vanhoucke, "Inception-v4, inception-resnet and the impact of residual connections on learning," CoRR, vol. abs/1602.07261, 2016.

[21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," CoRR, vol. abs/1512.03385, 2015.

[22] D. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (elus)," CoRR, vol. abs/1511.07289, 2015. [Online]. Available: http://arxiv.org/abs/1511.07289



Baptist Vandersmissen obtained the M.Sc. degree in Engineering from Ghent University, Belgium, in 2012. His Master's studies were followed by an internship at Carnegie Mellon University (CMU) in Pittsburgh, from September 2012 till January 2013. Since March 2013, he has been working as a doctoral researcher at IDLab (Ghent University – imec). His research interests include the automatic detection of semantic concepts in video fragments and the understanding and classification of radar data.

Nicolas Knudde finished his Master of Science in Engineering Physics at Ghent University in 2016. During his master thesis, done during his Erasmus exchange at KTH, he collaborated on several Photonics papers. As from September 2016 he is doing a PhD in Machine Learning and Scientific Computing. His interests are Bayesian modeling and classification of radar data.

Azarakhsh Jalalvand obtained his M.Eng. degree in Artificial Intelligence and Robotics from Iran University of Science and Technology (2009) and his Ph.D. degree in Computer Engineering from Ghent University in Belgium (2015). The focus of his Ph.D. research, which was supported by a European FP7 project, was on noise-corrupted speech and image signal processing using advanced artificial neural networks. He is currently a post-doctoral researcher in the IDLab research group at Ghent University – imec. His main research interests are noise-robust signal processing, machine learning, and time-series data analysis.

Ivo Couckuyt received his M.Sc. degree in Computer Science from the University of Antwerp (UA) in 2007. In October 2007 he joined the research group Computer Modeling and Simulation (COMS) (now merged with CoMP), supported by a research project of the Fund for Scientific Research Flanders (FWO-Vlaanderen). Since January 2009 he is active as a PhD student in the research group IDLab at Ghent University, where he obtained the PhD degree in 2013. He is currently working as an FWO post-doctoral research fellow in the IBCN research group of the Department of Information Technology (INTEC) in the Faculty of Engineering at Ghent University, Belgium.

André Bourdoux received the M.Sc. degree in electrical engineering (specialization in microelectronics) in 1982 from the Universit Catholique de Louvain-la-Neuve, Belgium. He joined IMEC in 1998 and is now Principal Member of the Technical Staff in the " Perceptive Systems for the IoT " Department of IMEC. He is a system level and signal processing expert for the mm-wave and sub-10GHz baseband teams and for the mm-wave radar team. His current research interests are multi-disciplinary, spanning the areas of wireless communications and signal processing, with a special emphasis on broadband systems and emerging physical layer techniques and high resolution radars. He is the author and co-author of over 140 publications in books and peer reviewed journals and conferences.

Wesley De Neve received the M.Sc. degree in Computer Science and the Ph.D. degree in Computer Science Engineering from Ghent University, Belgium, in 2002 and 2007, respectively. He is currently working in the position of Associate Professor for both the IDLab of Ghent University – imec in Belgium and the Center for Biotech Data Science of the Ghent University Global Campus (GUGC) in Korea. His main research interests are natural language understanding, visual content analysis, biotech data processing, and (deep) machine learning.

Tom Dhaene is Professor in Distributed Scientific Computing within the department of information technology (INTEC-IDLab) of the faculty of Engineering at Ghent University. In 1993, after his PhD on electromagnetic modeling at INTEC, Ghent University, he joined the EDA startup company Aphabit, later acquired by Hewlett-Packard, and now part of Agilent Technologies. In 2000, he went back to academia. His current research interests include machine learning, surrogate modeling, bioinformatics and distributed computing. He has published over 200 papers in international journals and conference proceedings, and he is the holder of 5 U.S. patents.