Parametric Convolutional Neural Network for Radar-based Human Activity Classification Using Raw ADC Data

Thomas Stadelmayer, Avik Santra, Robert Weigel, Fabian Lurz

Abstract—Radar sensors offer a promising and effective sensing modality for human activity classification. Human activity classification enables several smart homes applications for energy saving, human-machine interface for gesture controlled appliances and elderly fall-motion recognition. Present radar-based activity recognition system exploit micro-Doppler signature by generating Doppler spectrograms or video of range-Doppler images (RDIs), followed by deep neural network or machine learning for classification. Although, deep convolutional neural networks (DCNN) have been shown to implicitly learn features from raw sensor data in other fields, such as camera and speech, yet for the case of radar DCNN preprocessing followed by feature image generation, such as video of RDI or Doppler spectrogram, is required to develop a scalable and robust classification or regression application. In this paper, we propose a parametric convolutional neural network that mimics the radar preprocessing across fast-time and slow-time radar data through 2D sinc filter or 2D wavelet filter kernels to extract features for classification of various human activities. It is demonstrated that our proposed solution shows improved results compared to equivalent state-of-art DCNN solutions that rely on Doppler spectrogram or video of RDIs as feature images.

Index Terms—2D Sinc Filters, 2D Wavelets, Parametric CNN, Deep Learning, Human Activity Classification.

I. INTRODUCTION

People sensing and activity classification have increasing application potential in various areas, such as physical security, defense and surveillance. In industrial and consumer space, human activity recognition finds applications in smart homes, human-machine interfaces and elderly fall-motion monitoring systems. Knowledge of the performed activity in a room can enable smart control of the energy consumption, such as HVAC and lighting [1]–[3]. Furthermore, knowledge

of the performed human activity facilitates true ubiquitous smart home solution by discerning the user's intent.

Most of the human activity recognition systems are based on cameras and computer vision approaches. These systems have the advantage that they are quite easy to implement and benefit from several years of research, as a result the accuracy of such systems is quite high. However, camera systems suffer from lack of privacy and are sensitive to illumination conditions, thus they are not a favorable choice for smart home solutions. On the other hand, radar sensors have shown to be an effective sensing modality for human activity classification [4]–[13]. The radar sensors offer privacy-preserving, illumination-invariance properties and are capable of being aesthetically concealed in the operating environment. Recent innovation in the semiconductor technologies have facilitated integration and antenna-in-package solutions making the radar sensor into a small form factor [14]. Human activity recognition also enables sensing and recognition of elderly fall-motion, which is a leading cause of death in elderly population and in some cases if medical assistance are not provided immediately can lead to major restriction to overall mobility of the individual

This paragraph of the first footnote will contain the date on which you submitted your paper for review. It will also contain support information, including sponsor and financial support acknowledgment. For example, "This work was supported in part by the U.S. Department of Commerce under Grant BS123456."

T. Stadelmayer is with the Friedrich-Alexander University Erlangen-Nuremberg, 91054 Erlangen Germany and Infineon Technologies AG, 85579 Neubiberg Germany (e-mail: thomas.stadelmayer@fau.de; thomas.stadelmayer@infineon.com).

A. Santra is with Infineon Technologies AG, 85579 Neubiberg Germany (e-mail: avik.santra@infineon.com).

R. Weigel and F. Lurz are with the Friedrich-Alexander University Erlangen-Nuremberg, 91054 Erlangen Germany (e-mail: robert.weigel@fau.de; fabian.lurz@fau.de).

[15]–[19]. Thus, automatic sensing of fall-motion in particular among other activities have major social implications.

Radar sensors can sense and recognize human activities utilizing micro-Doppler signatures [20] that are generated by non-rigid-body motions of moving targets. Unique micro-Doppler signatures have been shown to be key features for object classification such as distinguishing among humans, animals and vehicles. The micro-Doppler signatures have been captured using hand-crafted features, a radar Doppler spectrogram or a video of range-Doppler images (RDI) and shown capable in recognizing human activities such as walking, running, crawling, standing, sitting, etc. as well as recognizing human gaits. In [4], authors propose to use six hand-crafted features such as torso Doppler frequency, total bandwidth (BW), offset of the total Doppler, BW without micro-Dopplers, the normalized standard deviation (STD) of the signal strength, and the period of the limb motion to classify seven different human activities, such as running, walking, walking while holding a stick, crawling, boxing while moving forward, boxing while standing in place, and sitting still. In [13], authors have used various deep convolutional neural network architectures to learn from range spectrogram, Doppler spectrogram and video of range-Doppler images for different activity classification. In [21], authors propose a continuous activity monitoring solution by augmenting the tracker's state by the classifier's activity classes. In [18], authors propose a novel deep auto-encoder based solution to sense elderly fall-motion from Doppler spectrogram. In [19], authors propose a deep deformable convolutional network to learn subject's aspect invariance features from radar Doppler spectrogram.

However, almost all deep learning solutions for human activity classification proposed in literature are based on learning features from radar Doppler spectrogram or video of radar range-Doppler images (RDI), but learning directly from radar raw ADC data was generally not feasible for a scalable solution so far. On the contrary, deep convolution neural networks (DCNN) applied in other domains such as computer vision and speech processing have been shown to be capable of implicitly learning commendable features from the raw input data without need of explicit features or preprocessing. So far DCNNs that are capable of learning directly from raw radar ADC data have not found much success in terms of scalable solution. Inspired from 1D SincNet in speech processing, in this paper we propose two DCNN architectures based on 2D-sinc filters and 2D wavelet filters that directly learn from radar raw ADC data to classify among different human activities using *Infineon's* 60-GHz frequency modulated continuous wave (FMCW) radar chipset *BGT60TR13C*. We demonstrate that the proposed trained DCNN architectures are able to achieve classification accuracy equal or better than equivalent DCNNs with Doppler spectrogram or video of RDIs as input. Thus, for the first time the true potential and advantages of DCNNs in radar processing and classification tasks have been demonstrated that do not rely on any preprocessing and can exploit the raw radar data as it has been demonstrated for data from other domains much earlier.

The rest of the paper is organized as follows, we present the

Fig. 1: (a) *Infineon's* *BGT60TR13C* 60-GHz radar sensor. (b) Functional block diagram of FMCW radar RF signal chain depicting 1TX, 1RX channel

radar system design and system parameters in Section II, the conventional signal processing involving Doppler spectrogram and video of RDI is presented in Section III, the contribution is also presented in Section III. The proposed solution is presented in section IV, the proposed architecture & learning in section V. The results and discussion along with the associated setup is presented in section VI, and we conclude in Section VII.

II. RADAR SYSTEM DESIGN

The work in this paper is based on *Infineon's* *BGT60TR13C* FMCW radar chipset. Its operating frequency ranges from 57 GHz to 64 GHz with an adjustable chirp duration. Its block diagram is shown in Fig. 1. The transmit path consists of a voltage controlled oscillator (VCO) that is regulated by a phase locked loop (PLL) to a reference frequency of $f_{\text{ref}} = 80$ MHz. Highly linear frequency chirps between 57 GHz and 64 GHz are produced by adjusting the divider value and an additional tuning voltage ranging from 1 V to 4.5 V. In the receive path the echo returning from the target object is down-converted with a replica of the transmitted frequency chirp. Therewith the baseband frequency spectrum can be sampled by the 12 bit analog-digital converter (ADC). Moreover the receive path contains an intermediate frequency buffer amplifier and an analog IF-filter that can be adjusted corresponding to the received frequency range. The radar chip is package in an embedded waver level ball grid array package including 4 integrated patch antennas realized by a metal redistribution layer. Three of them are receive antennas having an antenna gain of 10 dBi and one is the transmit antenna with a gain of 6 dBi. Consequently, the radar sensor contains three identical receive paths and one transmit path. The radio frequency (RF) signal is distributed by an active RF distribution network to the receive paths.

The transmitted up-chirp from the FMCW radar's ramp generator is reflected by a moving object and is received at the receiver after round trip delay caused by the target's range from the radar and the velocity of the target. The received signal is mixed at the receiver with the transmitted signal and the resultant signal is low-pass filtered, thus performing the matched filtering operation. The phase of the resultant intermediate frequency or IF signal due to single point target can be expressed as

$$\Phi(t) = 2\pi \left(f_{\min} \tau + \frac{B}{T_c} t \tau - \frac{B}{2T_c} \tau^2 \right) \quad (1)$$

where $f_{\min}$ is the ramp start frequency, B and T_c denote the chirp bandwidth and the chirp time respectively, and $\tau = \frac{2(x + \sum_{i=1}^1 v_i t)}{c}$ is the round trip propagation delay between the transmitted and received signal after reflection from the point target with range x and radial velocity components v_i . The Doppler frequency relates with radial velocity v_i as $\nu_i = 2v_i/\lambda = 2vf_{\min}/c$, the macro-Doppler is represented by the centroid Doppler of the Doppler components due to the target $\nu^c = \sum_{i=1}^1 \nu_i$, while micro-Doppler components are represented as $\{\nu_i - \nu^c\}_{i=1}^1$.

For our demonstrator, we configured the chip to transmit a single chirp per physical frame, where the frames are spaced $T_{\text{PRT}} = 1$ ms. Figure 2 presents the frame configuration for seamless activity detection and classification. The rationale behind choosing this chirp configuration is mainly scalability. The user can chose variable number of frames to build the slow-time data resulting in configurable Doppler resolution. The pulse repetition time, therefore, is 1 ms resulting in unambiguous maximum velocity of $v_{\max} = 1.25 \text{ m s}^{-1}$ which is sufficient in most indoor activity sensing applications. Furthermore, the chosen configuration enables the option to build continuous Doppler spectrogram, which can be fed into a classifier through sliding window.

Fig. 2: Chirp sequence in the proposed system configuration

The bandwidth is set to $B = 1$ GHz and the up-chirp time is set to $T_c = 64 \mu\text{s}$ accounting for a range resolution of 15 cm. The maximum detectable unambiguous range is 9.6 m. The set system parameters are provided in Tab. I.

III. CONVENTIONAL PIPELINE & CONTRIBUTIONS

The conventional signal preprocessing involves 1D moving target indication (MTI) filtering to remove the response from

TABLE I: Operating Parameters.

Parameters, Symbol	Value
Ramp start frequency $f_{\min}$	59.5 GHz
Ramp stop frequency $f_{\max}$	60.5 GHz
Bandwidth B	1 GHz
Range resolution δr	15 cm
Number of samples per chirp N_s	128
Maximum range $R_{\max}$	9.6 m
Sampling frequency f_s	2 MHz
Chirp time T_c	64 μs
Chirp repetition time T_{PRT}	1 ms
maximum Doppler $v_{\max}$	1.25 m s^{-1}
Number of chirps per frame	1
Elevation θ_{elev} per radar	70°
Azimuth θ_{azim} per radar	120°

static targets and also Tx-Rx leakage, which effects the first few range bins. The reflections from stationary objects such as chair, tables and wall, etc. can overwhelm the reflections from the other moving targets limiting their visibility on the RDI or Doppler spectrogram. Thus, MTI filter is used to suppress the contribution of these stationary objects and leakage. Among several MTI filters, a simple 1D MTI filter subtracts the mean along the fast-time to remove the Tx-Rx leakage that perturbs the first range bins, followed by mean subtraction along the slow-time to remove the reflections due to static or zero-Doppler targets.

The range information of the target is extracted by performing the first FFT after applying 1D windowing along fast-time, which is the intra-chirp time. The Doppler information of the target is extracted by monitoring the change of target peak along slow-time, which is the inter-chirp time. One common approach is applying FFT along the fast time as well as slow-time dimension. The outcome of this operation is a two-dimensional matrix representing the received power spectrum over range and velocity, also known as range-Doppler image (RDI). The received and deramped IF data is stored in matrices of size $N_c \times N_s$, where N_c being the number of chirps considered in a frame and N_s is the number of transmit samples per chirp.

In a conventional processing pipeline, the above preprocessing is followed by feature image generation, such as Doppler spectrogram and video of RDI, which are then input to deep convolutional neural network (DCNN) or long-short term memory (LSTM) for classification.

A. RDI video

The video of range-Doppler images computed at frame time k can be obtained by applying a 2D STFT on the mean removed ADC data and is expressed as

$$v_{\text{RDI}}(p, l, k) = \left| \sum_{m=1}^{N_{\text{st}}} \sum_{n=1}^{N_{\text{ft}}} w(m, n) s(m, n, k) e^{-j2\pi(\frac{mp}{N_{\text{st}}} + \frac{nl}{N_{\text{ft}}})} \right| \quad (2)$$

where $w(m, n)$ is the 2D weighting function along the fast-time and slow-time, $s(m, n, k)$ is the mean removed ADC data

on k^{th} frame. The index n, m sweep along the fast-time and slow-time axis respectively, while l, p sweep along the range and Doppler axes respectively. N_{st} and N_{ft} are the FFT size along the slow-time and fast-time respectively. Figure 3 presents the video of RDI for walking and working activity at time frame 0, 15 and 30.

Fig. 3: Example video of RDIs for walking and working activities

B. Doppler Spectrogram Features

The Doppler spectrogram is generated by marginalizing the video of RDIs v_{RDI} across the range axis and can be expressed as

$$S(p, k) = \sum_{l=1}^{N_{\text{r}}} v_{\text{RDI}}(p, l, k). \quad (3)$$

where l is sweeps along the range axis, N_{r} is the number of range bins and p, l, k are the Doppler, range and frame indices respectively. The Doppler spectrum at frame k contains both the macro-Doppler components as well as micro-Doppler components due to hand and leg movements while performing an activity. The stacked Doppler spectrum across consecutive frames is referred as Doppler spectrogram that captures information about the instantaneous Doppler spectral content and the variation of the Doppler spectral content over time. Figure 4 presents the Doppler spectrogram of the different activities, namely empty room, walking, standing idle, arm movement, waving and working on laptop.

Figure 5(a) presents the conventional pipeline that involves explicit preprocessing and feature generation followed by a neural network such as DCNN or LSTM for classification. The novel aspect of the proposed architecture implicitly performs the preprocessing and feature generation in the neural network itself. Thus the input to the neural network is the raw radar ADC data directly as depicted in Fig. 5(b). The initial layer of the proposed DCNN learns 2D Sinc filter kernels or 2D wavelet filter kernels, which is representative of the preprocessing and feature extraction. The capability of DCNN to directly operate on the raw ADC data helps in reducing the computation complexity dramatically as well in practical implementation eliminates the need for digital signal processor (DSP) for preprocessing.

Fig. 4: Example Doppler Spectrogram

Fig. 5: (a) Conventional processing pipeline, involving explicit preprocessing, feature generation and neural network (b) Proposed processing pipeline, involving 2D Sinc filter DCNN or 2D Wavelet filter DCNN for implicit preprocessing, feature generation and classification.

IV. PROPOSED APPROACH

Different activities can be distinguished by analyzing their unique range-velocity profiles. Some activities may have a very different profile such as walking and standing idle, but some may have only slight differences. For example working on the laptop and sitting idle on a chair only differ from slight hand movements to control the laptop. Thus, a higher resolution on specific frequency bands is required in order to accurately distinguish these actions. However, when applying a 2D STFT the whole observable range-velocity space is discretized in equal bins. Besides applying a STFT, time domain bandpass filters can be used to analyze the frequency composition of a signal. Time domain bandpass filters yield the ability to adjust the cutoff frequencies according to the needs of the application and can therefore be learned within a neural network.

A. 2D Sinc Filters

The benefit of integrating preprocessing into the neural network itself by learning the hyperparameters of a set of time domain filters was already shown in [22] for 1D audio signal processing. In the paper constraining the first convolutional layer to the use of 1D bandpass filters defined as the difference of two lowpass sinc filters with different cutoff frequencies is proposed. While training the lower and higher cutoff frequencies are optimized for the needs of the application. Replacing

classical preprocessing by this layer shows improved results in speaker recognition.

In this paper the extension from 1D audio signals to 2D radar signals is proposed. The 1D sinc filter is defined as

$$h_{K,f_s}(k, f_1, b) = 2(f_1 + b) \text{sinc}(2(f_1 + b) \frac{k - \lfloor \frac{K}{2} \rfloor}{f_s}) - 2f_1 \text{sinc}(2f_1 \frac{k - \lfloor \frac{K}{2} \rfloor}{f_s}) \quad (4)$$

where K is the filter length, f_s the sampling frequency of the signal, f_1 the lower cutoff frequency, b the bandwidth and k the filter parameter index. The hyperparameters of this filter are the lower cutoff frequency f_1 and the bandwidth b that implicitly defines the higher cutoff frequency. By defining a lower cutoff frequency and bandwidth in slow-time as well as in fast-time direction, a 2D bandpass filter that is able to extract joint range and velocity features can be created. The 2D sinc filter is defined as

$$\text{sinc}_{2D}(n, m; f_1^{st}, b_{st}, f_1^{ft}, b_{ft}) = w(n, m) h_{N, f_s^{st}}(n, f_1^{st}, b_{st}) h_{M, f_s^{ft}}(m, f_1^{ft}, b_{ft}) \quad (5)$$

where N and M are the filter-lengths, f_s^{st} and f_s^{ft} the sampling frequencies, f_1^{st} and f_1^{ft} the lower cutoff frequencies, b_{st} and b_{ft} the filter bandwidths respectively in slow-time and fast-time direction. Furthermore, $w(n, m)$ is a 2D cosine weighting function. n is sweeping along slow-time and m along fast-time. An exemplary 2D sinc filter is shown in Fig. 6 in time as well as in frequency domain. In frequency domain the rectangular shape with clear cutoff frequencies can be seen. The first layer of a CNN is initialized according to the definition of 2D sinc filters and only the hyperparameters are allowed to be learned during training.

Fig. 6: Exemplary 2D sinc filter in time and frequency domain

B. 2D Morlet Wavelets

The range-velocity profile is not only defined by its composed frequencies but also by the change of frequencies over time. When transforming a signal to frequency domain the time information is lost. This can be overcome by windowing the time domain signal. However, smaller window sizes mean higher time resolution but to the cost of worse frequency resolution and vice versa. Especially for time varying signals wavelets have several advantages over fourier transformations as they provide a time and frequency resolution [23]. Due to the fact that radar signals are highly time varying, the usage of a 2D wavelet transformation using Morlet wavelets

is proposed. The 2D Morlet wavelet is defined as

$$g_{N,M}(n, m; \sigma_{st}, \sigma_{ft}) = \frac{1}{2\pi\sigma_{st}\sigma_{ft}} \exp\left(-\left(\frac{(n/N - \lfloor N/2 \rfloor)^2}{2\sigma_{st}^2} + \frac{(m/M - \lfloor M/2 \rfloor)^2}{2\sigma_{ft}^2}\right)\right) \quad (6)$$

$$\phi[n, m; f_c^{st}, \sigma_{st}, f_c^{ft}, \sigma_{ft}] = g_{N,M}(n, m; \sigma_{st}, \sigma_{ft}) \cdot \cos\left(2\pi f_c^{st} \frac{n - \lfloor N/2 \rfloor}{f_s^{st}}\right) \cos\left(2\pi f_c^{ft} \frac{m - \lfloor M/2 \rfloor}{f_s^{ft}}\right) \quad (7)$$

where N and M are the filter-lengths, σ_{st} and σ_{ft} the standard deviations, f_c^{st} and f_c^{ft} the center frequencies, f_s^{st} and f_s^{ft} the sampling frequencies respectively in slow-time and fast-time direction. The hyperparameters that can be optimized by the neural network are the center frequency and the standard deviation of the wavelet. Similar to the previous introduced 2D sinc filters the frequency area of interest can be adjusted by the center frequency. But, additionally also the time-frequency resolution can be optimized by changing the standard deviation of the gaussian part of the wavelet. Due to the fact that the defined wavelet is the product of a cosinus and a gaussian window function also the frequency response has the shape of a gaussian. That means that it has no clear cutoff frequencies as it can be seen in Fig. 7, where an exemplary 2D Morlet wavelet in time and frequency domain is depicted. The standard deviations of the gaussian in time domain and in frequency domain are indirect proportional. Consequently, decreasing the width of the gaussian in time domain will lead to an increased width of the frequency response, which in turn shows the time-frequency resolution trade-off.

Fig. 7: Exemplary 2D Morlet wavelet in time and frequency domain

V. ARCHITECTURE AND LEARNING

In this paper two state of the art DCNNs evaluating the preprocessed data, a state of the art DCNN evaluating raw input data and two novel DCNN architectures using raw ADC data as input are proposed. All architectures have several characteristics in common. First, all networks finishing with a softmax classifier layer of size 6 as six different actions should be classified. Second, categorical crossentropy is used as loss function. Third, RMSprop optimizer is used with a learning rate $lr = 0.0001$, $\rho = 0.9$, $\epsilon = 10^{-8}$ and batches of size 128. Fourth, all unconstrained convolutional and dense layers are initialized using the 'Glorot' initialization scheme. Fifth, the common convolutional and dense layers are using a rectifier linear unit as activation. And sixths, after every common convolutional and dense layer a dropout with a rate of 0.2 is implemented in order to prevent over fitting.

Fig. 8: Proposed 2D filter learning CNN

Fig. 9: State-of-art DCNN architectures for evaluating (a) Doppler spectrograms and (b) video of RDIs

A. 2D SincNet

The 2D SincNet uses 2D sinc filter convolutions in the first convolutional layer as described in chapter IV-A. As parameter this layer takes the filter lengths, the number of filters, the sampling frequencies, the padding mode and the stride for the slow- as well as fast-time direction respectively. Although there are no separated filters for slow- and fast-time, it is required to explicitly provide the number of filters in slow-

time N_{st} and the number of filters in fast-time N_{ft} . According to this, 2D sinc filters are generated in a way that they form an equal grid of size $N_{st} \times N_{ft}$ covering the complete observable range Doppler domain. The trainable weights in this layer are the lower cutoff frequencies and the bandwidths in slow- as well as fast-time direction. In order to guarantee equal training in both filter dimensions the bandwidths and cutoff frequencies are normalized.

The 2D sinc filter layer is followed by a MaxPool layer with a pooling size of 8×2 . Afterwards a common two dimensional convolutional layers using 50 filters of size 3×3 is implemented. As already mentioned above, the convolutional layer follows a dropout layer with a rate of 0.2. Moreover, after the dropout a max pooling of size 4×2 is applied to decrease dimensionality. Then the tensor is flattened and fed into a dense layer of size 32 followed by the softmax classifier layer. The proposed network is depicted in Fig. 8

B. 2D WaveConvNet

The 2D WaveConvNet (WCN) is designed similar to the 2D SincNet. Only the first convolutional layer is initialized by 2D Morlet wavelets as described in IV-B instead of using 2D sinc filters. Required parameter for this layer are the filter lengths, number of filters, sampling frequencies, padding mode and stride for the slow as well as fast time direction respectively. Similar to the 2D sinc filters the number of filters in slow time direction N_{st} as well as the number of filters in fast time direction N_{ft} have to be explicitly provided in order to distribute the frequency response of the wavelets equally as a grid in the 2D frequency domain. Both time axis were normalized as already discussed in the previous section. As a result, the standard deviations is chosen to be 0.06 in both filter dimensions. In a total N_{st} times N_{ft} 2D wavelets are created. Trainable weights of this layer are the center frequencies and standard deviations in slow- as well as in fast-time dimension. Also in the 2D WCN the learnable weights are normalized.

C. State-of-art Networks

DSNet: The DSNet is a typical state-of-art 2D DCNN architecture followed by a dense and softmax layer. It uses already preprocessed 2D Doppler spectrograms as input. Furthermore, it contains three common 2D convolutions with 4, 8 and 16 filters respectively. Each filter uses a kernel size of 3×3 . After each convolution a dropout layer with dropout rate of 0.2 is used. Moreover, after the first two convolutional layers a MaxPooling of size 2×2 is used. Afterwards the tensor is flattened and fed into a dense layer of size 64 followed by the softmax classifier. The DSNet architecture is shown in Fig. 9 (a).

RDINet: The RDINet has a three dimensional stream of RDI images as input. Therefore three 3D convolutional layers are used to extract information from the video of RDIs. They all have a kernel size of $3 \times 3 \times 3$ and use 4, 8 and 16 channels respectively. Also here a dropout layer with a rate of 0.2 is added after each convolutional layer to prevent overfitting.

After the first two dropout layers a maxpooling of size $2 \times 2 \times 4$ is performed. Afterwards the tensor is flattened and further processed by a dense layer of size 64 before it is classified by the final softmax layer. The RDINet is sketched in Fig. 9 (b).

2D ConvNet: The 2D ConvNet uses the same architecture as the 2D SincNet and 2D WCN. Only the first layer is substituted by a unconstrained 2D convolutional layer with 'Glorot' weight initialization. No predefined time domain filters are used. Therefore each filter parameter can be learned individually.

VI. RESULTS AND DISCUSSION

A. Experimental Setup and Dataset

Fig. 10: Experimental setup for data recording with test person performing activity "working"

To evaluate the approach presented in this paper, a dataset was recorded in a real world environment. The radar was mounted on a tripod at a height of 1.20m and was placed in the corner of the room. The room has about $20m^2$ with a table and chairs inside. The experimental setup is shown in Fig. 10.

The dataset contains five different human activities plus additionally a recording of an empty room. To record the class "walking" a single human was allowed to randomly walk around. The class "idle" is split up in two recordings. First, a person was standing in front of the radar and in the second recording the person was sitting at the table facing towards the radar. As third activity random arm movements while standing were recorded. This class is called "arm movements". In order to record the fourth class called "waving" a person was waving with its hand at different positions in the room facing towards the radar. As last class working at the laptop while sitting at the table was recorded. Each activity was performed by the same person and recorded for about 18 minutes in total.

Samples containing 2048 chirps with an overlap of 512 chirps are cut out from the recordings. Given the fact that the chirp repetition time is 1ms each sample captures 2.048s. For each sample a Doppler spectrogram and a video of RDIs as described in sec. III-B and III-A is created. Therefore a dataset with raw ADC data, Doppler spectrograms and videos of RDIs based on exactly the same chirps per sample is obtained. For each activity about 700 samples are available. Due to slightly different recording times for each activity, the number of samples per class varies. In table II the exact number of samples per class is stated.

TABLE II: Samples per class

empty room	walking	idle	arm movements	waving	working
579	695	834	694	687	686

TABLE III: Model sizes of DSNet and RDINet

Layer	DSNet	Layer	RDINet
2D Conv (4, 3x3)	40	3D Conv (4, 3x3x3)	112
2D Conv (8, 3x3)	296	3D Conv (8, 3x3x3)	872
2D Conv (16, 3x3)	1168	3D Conv (16, 3x3x3)	3472
Dense (32)	65569	Dense (64)	102432
Softmax (6)	198	Softmax (6)	198
Total	67,270	Total	529,198

B. Evaluation Details

To evaluate the proposed approach, a filter length of 65 in slow- and 33 in fast-time dimension was chosen for the 2D sinc filters. The same filter size is used for the constrained convolutional layer that substitutes the 2D sinc filter layer in the 2D ConvNet. For the 2D wavelet convolution the filter length in slow time direction was doubled in order to allow the gaussian window function of the Morlet wavelet to expand. Moreover "valid" padding is used in both dimensions to obtain the same output shape after 2D sinc and 2D wavelet layer. This provides the possibility of substituting both layers one by one while keeping the remaining network the same. To reduce computational intensity a stride of 4 and 8 is used in slow- and in fast-time dimension respectively.

As the filter sizes are specified, the final size of the networks can be stated. The composition of parameters per layer is shown in tab. III for the DSNet and the RDINet and in tab. IV for the 2D SincNet, 2D WCN and 2D ConvNet. Given the fact that when evaluating Doppler spectrogram only velocity information has to be evaluated the network can be designed accordingly smaller. The RDINet uses a video of RDIs as input data. Thus, 3D convolutions have to be used which result in a higher number of parameters. Moreover, the first layer of the 2D SincNet as well as of the 2D WCN have significantly less parameters as the corresponding unconstrained convolutional layer of the 2D ConvNet. As a result the network size is reduced by more than 50 %.

TABLE IV: Model sizes of 2D SincNet, 2D WCN and 2D ConvNet

Layer	2D SincNet	2D WCN	2D ConvNet
2D Sinc (64, 65x33)	256	-	-
2D Wave (64, 129x33)	-	256	-
2D Conv (64, 65x33)	-	-	137344
2D Conv (50, 3x3)	28850	28850	28850
Dense (32)	102432	102432	102432
Softmax (6)	198	198	198
Total	131,736	131,736	268,824

TABLE V: Evaluation result

Model	Accuracy ($\pm$ deviation)	F1-score
DSNet	90.9 % ($\pm$ 1.7 %)	91.2 %
RDINet	94.4 % ($\pm$ 0.7 %)	94.7 %
2D ConvNet	94.6 % ($\pm$ 2.9 %)	94.9 %
2D SincNet	98.9 % ($\pm$ 1.0 %)	99.0 %
2D WCN	99.5 % ($\pm$ 0.1 %)	99.5 %

C. Confusion Matrix Classification

For evaluation a 5-fold cross validation is performed. The 2D SincNet, the 2D WCN, the 2D ConvNet and the RDINet are trained for 20 epochs. The DSNet is trained for 100 epochs to allow convergence. Afterwards the accuracy as well as the F1 score are evaluated using the testing part. The obtained accuracies and F1 scores are averaged over all runs. For the accuracy additionally the standard deviation is calculated. The results are shown in tab. V. The state-of-art approaches achieve an accuracy of about 91 %, 94 % and 95 % respectively, whereas the proposed architectures show an improved accuracy of 98.9 % and 99.5 %. In order to analyze the classification results in more detail the confusion matrices of RDINet representing the state-of-art approaches and the confusion matrix of 2D WCN representing the novel architectures are shown in Fig. 11. The limitation of the state-of-art approaches is unveiled by looking at the individual accuracy per class. While most actions are similar well classified by RDINet and 2D WCN, a big uncertainty between the class "idle" and "working" exists. However, the proposed filter learning based approaches do not show this limitation and therefore achieves better accuracy scores.

D. Learned filters

Before training is started the sinc filters as well as the wavelets are initialized as described in sec. V-A and sec. V-B. Moreover, the first unconstrained convolutional layer of the 2D ConvNet is initialized using 'Glorot' scheme. For each approach, cumulating the initial filters leads to an approximately uniform range and velocity gain over the whole space. During training the filter parameters are iteratively optimized. As a result the initial grid structure is dissolved by individual shifts and shape changes of the filters.

The cumulative gain of all filters after training is depicted in Fig. 12. The 2D sinc filters as well as the 2D wavelets have a bandpass characteristic. Therefore the resulting gain of cumulative filters look similar except of the fact that the 2D wavelet gain is smoother caused by its smooth filter shape in frequency domain. However, the resulting weights of unconstrained convolutional layer are quiet different and can not be physically interpreted.

E. Discussion

The limitation of the state-of-art approaches has its origin in preprocessing. The STFT for generating the RDIs equally discretizes the range as well as velocity domain. However, both activities "idle" and "working" contain very slight movements. As a result their features share similar range-Doppler bins and

(a) RDINet

(b) 2D WCN

Fig. 11: Confusion matrix of (a) RDINet and (b) 2D WCN

Fig. 12: Cumulative gain of learned (a) 2D sinc filters, (b) 2D wavelets and (c) unconstrained filters

thus the STFT processed data of those actions is very similar. This complicates a classification. However, due to the ability of learning filters, the 2D WCN as well as the 2D SincNet is able to mitigate this limitation by adapting the hyperparameter of its filters accordingly in order to clearly separate features of similar actions. Furthermore, the lack of range information gets noticeable when predicting classes based on Doppler

spectrograms. The missing range information is expected to have an even higher impact when analyzing activities of multiple humans simultaneously.

The 2D ConvNet is not limited by using preprocessed data. Therefore, in theory 2D ConvNet can potentially achieve the same accuracy, though not guaranteed, as SincNet and WCN by increasing the training effort. The underlying issue for this is the fact that the learned sinc filters as well as wavelets are within the search space of the unconstrained convolutional layer. However, it ends up with an accuracy of 96.6% with a quiet high standard deviation of 2.9% after 20 epochs. As the number of parameter is significant higher than the number of parameters of the proposed networks, the learning speed is decreased.

Using a predefined set of filters such as the proposed sinc filters or wavelets makes the outcome of the convolutional layer physically interpretability, which helps understanding the system. Moreover, due to filter learning application dependent meaningful range Doppler areas are focused. As a result not only static targets, but also disturbances caused, e.g., by the rotation of a ventilator or possible interferences with wireless signals are mitigated.

VII. CONCLUSION

Human activity classification has several applications in surveillance, human-computer interfaces and smart home applications. We present an activity classifier based on parametric DCNN using 2D sinc filter kernels or 2D wavelet filter kernels that can seamlessly detect and classify human activities directly using raw ADC radar data. We demonstrated the performance of the proposed DCNN in comparison to conventional DCNNs that use Doppler spectrogram or video of RDIs as feature images and demonstrated the proposed solution offers better classification accuracy. While the preprocessing steps in the latter processing are fixed, the parametric DCNN is able to learn hyper-parameters specific to the activity classification task. Additionally, we demonstrate that the conventional un-parametric DCNN is unable to mimic the preprocessing operations due to the large search space and thus is not suitable for a robust solution. As future work, we aim to extend the proposed DCNN to simultaneously classify multiple activities from multiple targets in the field of view.

REFERENCES

- [1] EPRI, "Occupancy sensors: Positive on/off lighting control," Electr. Power Res. Inst., Palo Alto, CA, USA, Rep. EPRI-BR-100323, 1994.
- [2] V. Garg and N. K. Bansal, "Smart occupancy sensors to reduce energy consumption," *Energy Buildings*, vol. 32, no. 1, pp. 81–87, Jun. 2000.
- [3] A. Santra, R.V. Ulaganathan and T. Finke, "Short-Range Millimetric-Wave Radar System for Occupancy Sensing Application," *IEEE Sensors Letters*, vol. 2, no. 3, Sep 2018.
- [4] Y. Kim and H. Ling, "Human Activity Classification Based on Micro-Doppler Signatures Using a Support Vector Machine," in *IEEE Transactions on Geoscience and Remote Sensing*, vol. 47, no. 5, pp. 1328–1337, May 2009.
- [5] D.P. Fairchild and R.M. Narayanan, "Human activity classification using Hilbert-Huang transform analysis of radar Doppler data", In: *Radar Sensor Technology XV*. vol. 8021, International Society for Optics and Photonics, 2011. p. 80210

- [6] D.P. Fairchild and R.M. Narayanan, "Classification and modeling of human activities using empirical mode decomposition with S-band and millimeter-wave micro-Doppler radars," In: *Radar Sensor Technology XVI*. vol. 8361. (International Society for Optics and Photonics, 2012. p. 83610X
- [7] Y. Kim and T. Moon, "Human Detection and Activity Classification Based on Micro-Doppler Signatures Using Deep Convolutional Neural Networks", *IEEE Geoscience and Remote Sensing Letters*, 2016, 13, (1), pp. 8–12
- [8] R.J. Javier and Y. Kim, "Application of Linear Predictive Coding for Human Activity Classification Based on Micro-Doppler Signatures", *IEEE Geoscience and Remote Sensing Letters*, 2014, 11, (10), pp. 1831–1834.
- [9] R.M. Narayanan and M. Zenaldin, "Radar micro-Doppler signatures of various human activities," *IET Radar, Sonar & Navigation*, 2015, 9, (9), pp. 1205–1215.
- [10] M. Zenaldin and R.M. Narayanan, "Radar micro-Doppler based human activity classification for indoor and outdoor environments," In: *Radar Sensor Technology XX*. vol. 9829. (International Society for Optics and Photonics, 2016. p. 98291.
- [11] Y. Lin, J. Le.Kernec, S. Yang, F. Fioranelli, O. Romain and Z. Zhao, "Human Activity Classification With Radar: Optimization and Noise Robustness With Iterative Convolutional Neural Networks Followed With Random Forests," *IEEE Sensors Journal*, 2018, 18, (23), pp. 9669–9681.
- [12] X. Li, Y. He, Y. Yang, Y. Hong and X. Jing, "LSTM based Human Activity Classification on Radar Range Profile," In: *2019 IEEE International Conference on Computational Electromagnetics (ICCEM)*, 2019. pp. 1–2.
- [13] R. Hernangómez, A. Santra and S. Stanczak, "Human Activity Classification with Frequency Modulated Continuous Wave Radar Using Deep Convolutional Neural Networks," *RADAR 2019 International Conference*, 23-27 Sept. 2019, Toulon, France, 2019.
- [14] A. Santra, I. Nasr and J. Kim, "Reinventing radar: The power of 4D sensing", *Microw. J.*, vol. 61, no. 12, pp.26-38, 2018.
- [15] S. Tomii and T. Ohtsuki, "Falling detection using multiple Doppler sensors," *Proc. IEEE Int. Conf. e-Health Networking, Applications and Services*, Beijing, China, Oct. 2012, pp. 196-201.
- [16] B. Y. Su, K. C. Ho, M. J. Rantz, and M. Skubic, "Doppler radar fall activity detection using the wavelet transform," *IEEE Transaction of Biomedical Engineering*, vol. 62, no. 3, pp. 865-875, Mar. 2015.
- [17] Q. Wu, Y. D. Zhang, W. Tao, and M. G. Amin, "Radar-based Fall Detection Based on Doppler Time-Frequency Signatures for Assisted Living," *IET Radar, Sonar & Navigation*, vol. 9, no. 2, 2015, pp. 164-172.
- [18] B. Jekanovic, M. Amin and F. Ahmad, "Radar Fall Motion Detection Using Deep Learning," *IEEE radar conference (RadarConf)* 2016, pp. 1-6.
- [19] Y. Shankar, S. Hazra, A. Santra, "Radar-Based Non-intrusive Fall Motion Recognition using Deformable Convolutional Neural Network", *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, 2019, pp. 1717-1724.
- [20] V.C. Chen: "The Micro-Doppler Effect in Radar", Artech House, 2011.
- [21] P. Vaishnav and A. Santra, "Continuous Human Activity Classification with Unscented Kalman Filter Tracking using FMCW radar", *IEEE Sensors Letters*, 2020 (under review).
- [22] M. Ravanelli and Y. Bengio, "Speaker Recognition from Raw Waveform with SincNet", *2018 IEEE Spoken Language Technology Workshop (SLT)*, 2018, pp. 1021-1028.
- [23] V. Pukhova, E. Gorelova, G. Ferrini and S. Burnasheva, "Time-frequency representation of signals by wavelet transform," *2017 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus)*, St. Petersburg, 2017, pp. 715-718.

Thomas Stadelmayer was born in Regensburg, Bavaria, Germany in 1994. He received the B.S. and M.S. degrees in Computational Engineering with focus on digital signal processing from the Friedrich-Alexander University Erlangen-Nuremberg in 2016 and 2019.

Since 2019 he is a Research Assistant with the Institute of Electronics Engineering in the circuits, systems and hardware test group and working in cooperation with Infineon Technologies AG.

His research focus is on digital signal processing and deep learning for mm wave radar systems.

Avik Santra (S'09-M'10-SM'18) received his M.S. degree in Signal Processing with cum laude from Indian Institute of Science, Bangalore in 2010. He is currently working as senior staff algorithm engineer developing signal processing and machine learning algorithms for industrial and consumer radars and depth sensors at Infineon, Neubiberg. Earlier in his career, he has worked as system engineer for LTE/4G modem at Broadcom Communications, and also has worked as research engineer developing cognitive

radars at Airbus. He is a reviewer at various IEEE and Elsevier journals, and is recipient of several outstanding reviewer awards. He is author of the book titled 'Deep Learning Applications of Short-Range radars', published by ArTech House. He has filed over 40 patents and published 30 research papers related to various topics of radar waveform design, radar signal processing and radar machine/deep learning topics.

Robert Weigel (S'88-M'89-SM'95-F'02) was born in Ebermannstadt, Germany, in 1956. He received the Dr.-Ing. and Dr.-Ing.habil. degrees in electrical engineering and computer science from the Technical University of Munich, Munich, Germany, in 1989 and 1992, respectively. From 1994 to 1995, he was a Guest Professor of SAW Technology with the Vienna University of Technology, Vienna, Austria. He was a Research Engineer, a Senior Research Engineer, and a Professor of RF circuits and systems with the

Technical University of Munich until 1996. From 1996 to 2002, he was the Director of the Institute for Communications and Information Engineering, University of Linz, Linz, Austria, where he co-founded the company DICE in 1999, which split into an Infineon Technologies (DICE) and an Intel (DMCE) company, which are devoted to the design of RFICs and MMICs. In 2000, he was appointed as a Professor of RF engineering with Tongji University, Shanghai, China. Since 2002, he has been the Head of the Institute for Electronics Engineering, University of Erlangen-Nuremberg, Erlangen, Germany, where he co-founded the companies eesy-id and eesy-ic, in 2009 and 2012, respectively. He has authored or coauthored more than 900 articles. He has been engaged in research and development of microwave theory and techniques, electronic circuits and systems, and communication and sensing systems. Dr. Weigel served in many roles for the IEEE MTT-S and UFFC-S. He has been the Founding Chair of the Austrian COM/MTT Joint Chapter, a Region 8 MTT-S Coordinator, a Distinguished Microwave Lecturer, an IEEE MTT-S AdCom member, and the 2014 MTT-S President.

Fabian Lurz (S'13-M'20) received the B.Sc. and M.Sc. degree in information and communication technology and the Dr.-Ing. degree from the Friedrich-Alexander University Erlangen-Nuremberg (FAU), Erlangen, Germany, in 2010, 2013 and 2019, respectively. In 2013, he joined the Institute for Electronics Engineering, FAU, as a Research Assistant, where he became the Group Leader of the circuits, systems and hardware test group in 2017. His current research interests include microwave circuits and

system, especially for low-cost and low-power metrology applications. Dr. Lurz was a recipient of the First Prize of the High Sensitivity Radar Student Design Competition of the IEEE International Microwave Symposium in 2014, 2017, and 2018, respectively, and the IEEE Microwave Theory and Techniques Society Graduate Fellowship Award in 2016.