IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received December 24, 2021, accepted January 27, 2022, date of publication February 10, 2022, date of current version March 9, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3150838

Multi-View CNN-LSTM Architecture for Radar-Based Human Activity Recognition

HABIB-UR-REHMAN KHALID[1,2], ALI GORJI[2], ANDRÉ BOURDOUX[2], (Senior Member, IEEE), SOFIE POLLIN[2,3], (Senior Member, IEEE), AND HICHEM SAHLI[1,2]

[1]Department of Electronics and Informatics (ETRO), Vrije Universiteit Brussel (VUB), 1050 Brussels, Belgium
[2]Interuniversity Microelectronics Centre (IMEC), 3001 Leuven, Belgium
[3]Department of Electrical Engineering (ESAT), Katholieke Universiteit Leuven (KU Leuven), 3000 Leuven, Belgium

Corresponding author: Habib-Ur-Rehman Khalid (hkhalid@etrovub.be)

This work was supported in part by the Flemish Agency for Innovation and Entrepreneurship via the project WATCHDOG.

ABSTRACT In this paper, we propose a Multi-View Convolutional Neural Network and Long Short-Term Memory (CNN-LSTM) network which fuses multiple ''views'' of the time-range-Doppler radar data-cube for human activity recognition. It adopts the structure of convolutional neural networks to extract optimal frame based features from the time-range, time-Doppler and range-Doppler projections of the radar data-cube. The CNN models are trained using an unsupervised Convolutional Auto-Encoder (CAE) topology. Afterwards, the pre-trained parameters of the encoder are fine-tuned to extract intermediate frame based representations, which are subsequently aggregated via LSTM networks for sequence classification. The temporal correlation among the views is explicitly learned by sharing the LSTM network weights across different views. Moreover, we propose range and Doppler energy dispersion and temporal difference based features as an input to the CNN-LSTM models. Furthermore, we investigate the use of target tracking features as an auxiliary side information. The proposed model is trained on datasets collected in both cluttered and uncluttered environments. For validation, an independent test dataset, with unseen participants, in a cluttered environment was collected. Fusion with auxiliary features improves the generalization by 5%, yielding an overall Macro F1-score of 74.7%.

INDEX TERMS Smart-Homes, feature fusion, data fusion, deep learning, FMCW radar, time-range, Time-Doppler, Range-Doppler, indoor human activity recognition (HAR), convolutional neural networks (CNN), long short-term memory (LSTM), generalization.

I. INTRODUCTION

Human Activity Recognition (HAR) is an active area of research since decades and has been a key enabler of various emerging technologies, such as, smart-homes, smart-health, smart-security and smart-offices. Activity recognition is essential to these application domains as it allows computer systems to monitor and analyze human behavior and assist us in our daily lives. A reliable HAR is still a challenging problem and is faced with many technical issues. On one hand, the *privacy* issue can be solved by employing methods that involve ambient-sensors instead of camera based methods, but on the other, a reliable and a robust *feature extraction* in the presence of *heterogeneous* sensory data, is still quite challenging and requires significant research efforts.

Most of the work in radar-HAR assumes human-centric uncluttered background scenes, where participants are free to perform actions. For an automated indoor HAR system to work reliably and be able to classify actions with low error-rate, it is vital for the system to take into account *heterogeneity* in the sensor data, which may arise due to various external factors, such as, the users, the sensors and the environment. Firstly, each user or participant is different, hence their actions may also differ depending on their

The associate editor coordinating the review of this manuscript and approving it for publication was Brian Ng.

habits and morphology. Secondly, these habits may change over time, making this phenomenon time-variant. Moreover, variations in the sensor placement may result in variations in the aspect angle, relative to the radar sensor, which can greatly affect the feature extraction. Finally, the environment layout might vary from room to room, hence resulting in a different background clutter, and partial occlusion of the human body from the nearby objects, thus giving rise to a shadowing effect leading to a wide range of intra-class variations in the observed features. A reliable radar-HAR system should work regardless of these external factors. For example, when we move from an ideal uncluttered scenario to a more realistic cluttered scenario, a robust radar-HAR system should be able to generalize over the heterogenous sensory data, without much loss of performance.

A complex human-activity can be decomposed into motion sequences of walking, stand-to-sit or sit-to-stand actions [1], [6]. Hence, it is vital for radar-HAR systems, to be able to map an observed human motion to either of these actions. Therefore, in this work, we focus on classifying the following four action classes, 1) *Walking-Towards* (WT) and 2) *Walking-Away* (WA) from the radar-sensor, 3) *Stand-to-Sit* or *Sitting-Down* (SD) and 4) *Sit-to-Stand* or *Standing-Up* (SU). We pose the radar-HAR problem as a 4-class sequence classification problem.

In this paper we investigate a radar-sensor based approach to HAR. Our main goal is to investigate the generalization capability of the proposed learning methods and the models in the presence of different persons, different aspect angles, cluttered environment and multi-path. The major contributions of this paper include:

- A Multi-View Convolutional Neural Network and Long Short-Term Memory (CNN-LSTM) network which fuses multiple ''views'' of the time-range-Doppler radar data-cube for human activity recognition. The views are obtained by projecting the 3D data-cube in the range, Doppler and slow slow-time dimensions;
- A kind of soft-attention in the radar data-cube via range and Doppler energy dispersion based pre-processing;
- The use of target tracking features as auxiliary side information for the CNN-LSTM architecture;
- A multi-view fusion approach, achieved by sharing the weights across the views in each LSTM layer to learn the correlation between the views and auxiliary tracking features;
- A model training, combining an unsupervised feature extraction step, followed by a fine-tuning step to make feature extraction class agnostic and more robust to initializations.

The rest of the paper is organized as follows. In Section II we give a brief overview of the current state-of-art. Section III briefly describes the sensor-setup and the input from the radar-sensor. In Sections IV and V we describe the proposed methods and the models used in this research, which is followed by the discussion about the evaluation setup and the results of the proposed methods. Finally, Section VI concludes the paper with the future work.

II. RELATED WORK

A large and growing body of literature has investigated radar-based HAR classification. An exhaustive revision of the previous methods lies beyond the scope of this paper. We, therefore, briefly give an overview of the related works. For an in-depth review of the research area, the interested reader is referred to [11], [16].

Different representations of the radar signal, that include time-frequency domain based time-Doppler images, and integrated slow-time based range-Doppler images, have been considered. The time-frequency based features contain information about the rate of change of motion, of human body parts over time, while range-Doppler images provide both velocity and range information.

Over the years, several studies have adopted hand-crafted feature based representations extracted from time-range, range-Doppler, cadence velocity maps [3], [25], along with machine learning models such as support vector machines and random forests. In [2], the authors estimated the phase, velocity, rate of change, mean, standard deviation and range of the I and Q signals, and applied a random forest classifier for HAR classification. The authors of [7] proposed a dynamic range-Doppler trajectory (DRDT) method to recognize various human motion. First, range-Doppler frames consisting of a series of range-Doppler images are obtained from the backscattered signals. Next, the DRDT is extracted from these frames to monitor human motion in time, range, and Doppler domains in real time. Then, a peak search method is applied to locate and separate each human motion from the DRDT map. Finally, range, Doppler, radar cross section, and dispersion features are extracted and combined in a multidomain fusion approach as inputs to a machine learning classifier. In [10], the authors trained a random forest classifier with time-Doppler and range-Doppler based region-of-interest features such as velocity centroid, dispersion, and instantaneous energy based features, as proposed in [17], together with tracking based (such as target location, velocity, acceleration, range and azimuth) and point cloud features.

Recent radar-HAR efforts have applied deep neural networks. In [27], a CNN with three convolution layers, has been applied on range-Doppler images for multiclass HAR classification. To take into consideration the temporal characteristics of radar signals, the authors of [32] applied a 1D-CNN to extract spatial features from the spectrograms, followed by an LSTM network to learn time-dependent information. The authors of [20], applied a similar CNN architecture containing three convolutional layers, two max pooling layers and two fully connected layers, for classifying kitchen activities. The CNN model, as an input, takes an image with two spectrograms from two radar sensors. A multiscale residual attention network, for joint activity recognition and person identification, has been proposed in [12]. The architecture consists of a CNN, with a residual attention mechanism,

which extracts features from the time-Doppler images. The embeddings are fed to a fully connected layer performing the classification task. In contrast with the above works using time-Doppler or range-Doppler images as an input, the authors of [19] applied a LeNet-5 based CNN on the features extracted from an auto correlation function. The authors of [31] proposed an end-to-end deep learning based framework called the Fourier Convolutional neural Network (F-ConvNet). The input of an F-ConvNet consists of raw frames of radar data. Next, multi-scale features are extracted using three convolutional layers. The results are sent to a so called Fourier layer, learning the real and imaginary parts separately. Compared to the above CNN based approaches, the authors of [28] proposed a stacked Bidirectional LSTM (Bi-LSTM) network on spectrograms to perform radar-HAR. Bi-LSTMs can capture both temporal forward and backward correlated information within the radar data-cube.

Considering that the deep learning based models require a large amount of training data, some researchers propose transfer learning based methods. In [8], a ResNet based pre-trained CNN on ImageNet database, is fine-tuned on time-Doppler spectrograms. The authors of [29] proposed a generative adversarial based image-to-image translation approach to transform time-Doppler signatures into a pseudo-audio representation, and fine-tuned a pre-trained VGGish CNN to classify the obtained representations.

Unsupervised feature learning based methods were also investigated. The authors of [26] used a three-layer CAE that used an unsupervised pretraining to alleviate the demand for training data, followed by a supervised fine-tuning of the CNN to extract spatially localized features for classification. The authors of [4] extended the CAE based unsupervised feature extraction, and proposed an attention-augmented CAE, wherein the convolutional maps are concatenated with a multi-headed attention output [30]. The CAE model is first pre-trained in an unsupervised fashion, and then both the convolution and attention parts of the encoder are fine-tuned separately through supervised training. Next, the convolution and attention parts are trained jointly to learn the final configuration for classification.

Different from the above image-based models, the authors of [15] employed a generalized point cloud model to simultaneously represent the time-range-Doppler signature. First, the radar echoes are transformed via range-Doppler processing along the time axis. Then, the target information is gathered by a constant false alarm rate detection algorithm. The point cloud features are aggregated by motion sculpture construction and iterative farthest point sampling. Finally, the resulting point cloud is fed to a hierarchical Point-Net module [5] to recognize human activities.

Recently, researchers considered information fusion by merging complementary radar information at different abstraction levels: signal, feature and/or decision. In [13] a combination of time-range, spectrogram and integrated range-Doppler information are processed with sparse autoencoder to extract features that are then classified by a softmax layer for each of the three inputs. A voting principle is then used as decision fusion. In [9], the radar data-cube is pre-processed with an extended CLEAN algorithm to eliminate unwanted noise/distortions. Then, a multi-dimensional Principal Component Analysis (PCA) approach is applied for feature extraction, followed by a linear discriminant analysis implemented as a shallow neural networks for classification.

III. FMCW RADAR SENSOR SETUP

In this study we use a single 3×4 multiple-input multiple-output Frequency Modulated Continuous Wave (FMCW) radar sensor setup. The sensor operates, in an indoor environment, at a center-frequency (f_0) of 60 GHz with a bandwidth (B) of 2.3 GHz. The radar sensor setup and the waveform parameters are described in more detail in [10].

In the following, we give an overview of the used 3D radar data-cube (x_{raw}) and describe its related 2D feature ''views'' and the auxiliary features used in this study. The reader is referred to [10] for the details of the detection and tracking steps together with the radar signal processing and 3D radar data-cube creation.

A. 3D RADAR DATA-CUBE STRUCTURE AND CHARACTERISTICS

We follow a conventional FMCW signal processing pipeline (as described in [10]). The targets are detected and tracked in the radar's Field of View (FoV), and an estimate of the target's centroid in the radial-range and the azimuth-angle dimensions is used to create a 3D radar data-cube.

We extract 32 range bins and 2 angle bins around the target's centroid. The STFT based micro-Doppler processing yields a ($T \times D$) 32×32 time-Doppler or *micro-Doppler* spectrum, which is observed in ($R \times A$) 32×2 range and angle bins, thus resulting in a uniform-sized 3D radar data-cube of cardinality ($T \times R \times D \times A$), where ($R \times D \times A$) are the three spatial dimensions evolving in slow slow-time (T). Furthermore, we keep 1 *sec* worth of activity by maintaining a buffer of size $N = 12$ previous frames of radar data-cubes. This results in an ordered sequence of 3D radar data-cubes ($x_{raw}^{(1\ldots N)}$).

B. 3D FEATURE STREAMS AND AUXILIARY FEATURES FOR RADAR HAR

After extracting the 3D radar data-cube (also known as the 3D time-range-Doppler data-cube [16]), the range, Doppler and the slow slow-time features are pre-processed before being used for activity classification. The pre-processing step involves, projecting the 3D data-cube in the range, Doppler and slow slow-time dimensions, scaling the features with the range and Doppler energy dispersion based profiles, and calculating the difference features based on the temporal differences of features. We consider four categories of features: 1) raw features, 2) energy dispersion based features, 3) temporal difference based features, and 4) auxiliary features.

- *Raw features:* The 3D time-range-Doppler data-cube, x_{raw}^{n}, contains the reflected energy by the participant's

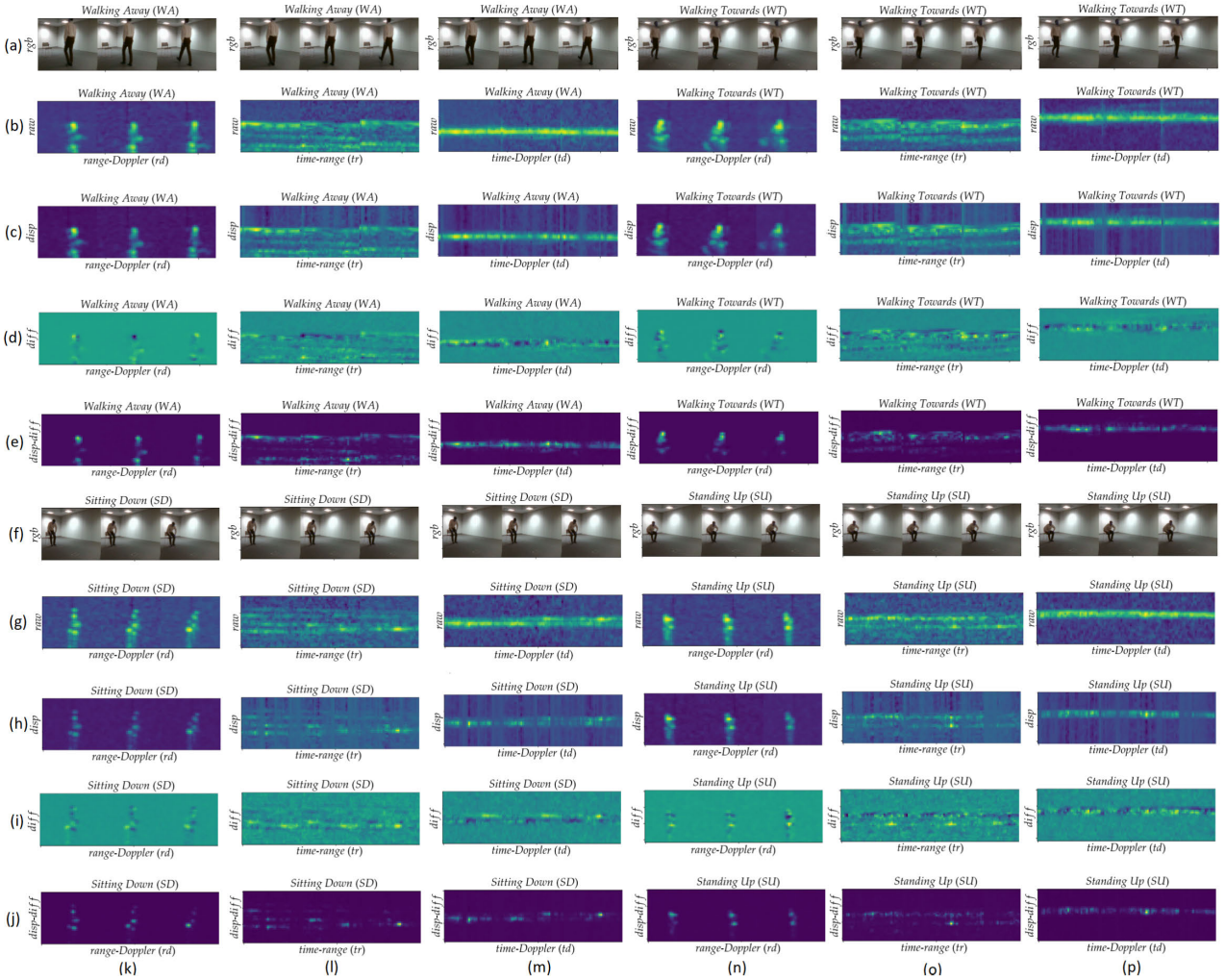

FIGURE 1. The 2D projections of the raw, energy dispersion, difference, and dispersion-difference based 3D data-cubes for all four, WA, WT, SD, and SU actions are shown, where rows (b) and (g) represent projections from the raw data-cube, rows (c) and (h) are projections from the energy dispersion based data-cube, rows (d) and (i) from the difference, and rows (e) and (j) from the dispersion-difference based data-cubes. Columns (k) and (n) represent range-Doppler projections, columns (l) and (o) represent time-range, and columns (m) and (p) time-Doppler projections. We show 0.25 *sec* (i.e. $N = 3$ frames) of 4 actions: action WA is shown in columns (k)-(m) and rows (a)-(e), columns (n)-(p) and rows (a)-(e) show WT action, columns (k)-(m) and rows (f)-(j) show SD action, while SU action is shown in columns (n)-(p) and rows (f)-(j).

limbs and torso as they move in the radar's FoV, and therefore is a function of range (r), Doppler (d), angle (a) and slow slow-time (t) dimensions.

A 2D range-Doppler map (or view) is obtained by projecting the 3D time-range-Doppler data-cube in the range-Doppler space ($x^n_{rd,raw}$), in the same way we obtain 2D time-range view ($x^n_{tr,raw}$) and 2D time-Doppler view ($x^n_{td,raw}$), by projecting the 3D time-range-Doppler data-cube in the time-range and time-Doppler space, respectively as illustrated in Figure 1.

- *Energy Dispersion based features:* are derived from the 3D time-range-Doppler data-cube x^n_{raw}. The idea is inspired by [17], however unlike [17] we use range and Doppler profile energy dispersion for creating a kind of soft-attention in the slow slow-time dimension of the 3D data-cube.

  We first, estimate the instantaneous range (r^n_{cent}) and Doppler (d^n_{cent}) profile energies, by using the 2D time-range and the 2D time-Doppler views, respectively as follows:

$$r^n_{cent}(t,a) = \sum_{d=1}^{D}\sum_{r=1}^{R} x^n_{tr,raw}(t,r,a)x^n_{raw}(t,r,d,a) \tag{1}$$

$$d^n_{cent}(t,a) = \sum_{r=1}^{R}\sum_{d=1}^{D} x^n_{td,raw}(t,d,a)x^n_{raw}(t,r,d,a) \tag{2}$$

The energy profiles are then used for creating a soft-attention in the slow slow-time dimension of the 3D time-range-Doppler data-cube:

$$x^n_{disp}(t,r,d,a) = \frac{||x^n_{td,raw}(t,d,a) - d^n_{cent}(t,a)||_2^2}{||x^n_{tr,raw}(t,r,a) - r^n_{cent}(t,a)||_2^2} x^n_{raw}(t,r,d,a) \tag{3}$$

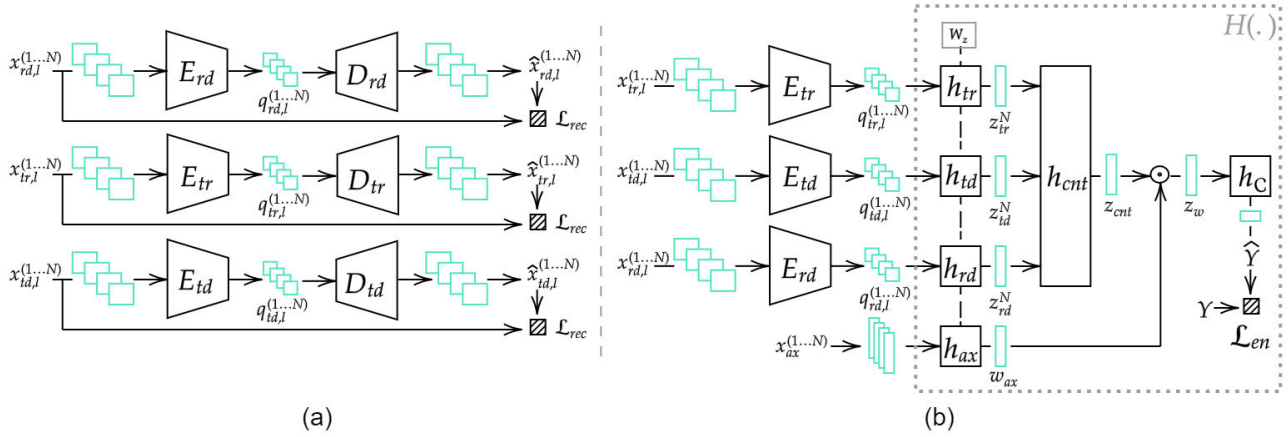

FIGURE 2. The proposed radar-HAR framework is shown, where a) illustrates the CAE based unsupervised feature learning, with E_S and D_S as the respective encoder and decoder CNN models for each view $s = \{\text{rd, td, tr}\}$ and stream $l = \{\text{raw, disp, diff, disp-diff}\}$, while b) illustrates the MV RADAR-Net that performs frame-based space-time feature extraction via a pretrained (or finetuned) encoder CNNs E_S for each view first. These features are further converted to sequence-based features, where the multi-view data-fusion including auxiliary features is performed in LSTM layers (defined in h_S and h_{ax}, in $H(.)$). Finally, the sequential features are concatenated in $h_{cnt} \in H(.)$ and adapted using the auxiliary feature based context w_{ax}. The final classification is performed by $h_C \in H(.)$ on the adapted sequential features.

- *Temporal Difference based features:* are derived from the raw 3D time-range-Doppler data-cubes, with the idea of putting more emphasis on the most recent events (compared to the previous data-cube) occurring in the raw radar data-cube:

$$x^n_{diff}(t, r, d, a) = x^n_{raw}(t, r, d, a) - x^{n-1}_{raw}(t, r, d, a) \quad (4)$$

  Moreover, these recent events can be further highlighted by estimating a dispersion-difference based data-cube:

$$x^n_{disp\text{-}diff}(t, r, d, a) = ||x^n_{diff}(t, r, d, a)||_1 x^n_{disp}(t, r, d, a) \quad (5)$$

  From the above 3D energy based data-cubes we extract 2D time-Doppler views $x^n_{td,l}$, time-range views $x^n_{tr,l}$ and range-Doppler views $x^n_{rd,l}$ for $l \in \{\text{disp, diff, disp-diff}\}$ (Figure 1).
- *Auxiliary features:* are estimated using the target tracking results, encoded as target location (x, y), range (r), azimuth (θ) and Doppler $(d^n = \arg\max_d \sum_{t=1}^{T} \sum_{a=1}^{A} x^n_{td}(t,d,a))$ as follows (Equation 6):

$$x^n_{aux} = [x^n\text{-}x^{n\text{-}1} \;\; y^n\text{-}y^{n\text{-}1} \;\; r^n\text{-}r^{n\text{-}1} \;\; d^n\text{-}d^{n\text{-}1} \;\; \theta^n\text{-}\theta^{n\text{-}1} \;\; \dot{x}^n \;\; \dot{y}^n \;\; \ddot{x}^n \;\; \ddot{y}^n \;\; \dot{d}^n \;\; \dot{\theta}^n]^T \quad (6)$$

IV. PROPOSED MODEL

We propose a multi-view fusion approach, where we fuse multiple 2D views and auxiliary features from a single radar sensor. For data-fusion and sequential feature learning, an LSTM layer with shared parameters is proposed. As a result, the proposed model, denoted as Multi-View RAnge Doppler Activity Recognition Network (MV RADAR-Net), is composed of space-time frame based model, $E_s(.)$ with $s \in \{\text{rd, td, tr}\}$, and a sequential model $H(.)$.

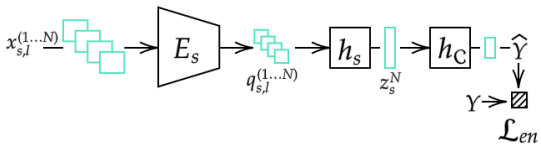

FIGURE 3. Single view CNN-LSTM model for optimal view-stream pair selection using an E2E learning approach is shown, where E_S, h_S and h_C are the respective CNN based encoder, LSTM network based sequential model and a fully connected layer based classification head, defined with same configration as in Section IV-A, and Section IV-B, while $s = \{\text{rd, td, tr}\}$ and $l = \{\text{raw, disp, diff, disp-diff}\}$.

The space-time frame based $E_s(.)$ are pre-trained using CAE models as illustrated in Figure 2a and discussed in Section IV-A. The pre-trained encoders $(E_s(.))$ are then used with the sequential model $(H(.))$, illustrated in Figure 2b and discussed in Section IV-B, where we optionally fuse the auxiliary features.

A. SPACE-TIME CAE MODEL

The CAE models for each view with an encoder and decoder can be formally defined as follows:

- For each 2D view s in {rd, td, tr}, we define a space-time feature stream from l in {raw, disp, diff, disp-diff}. For a given view-stream pair, a space-time frame based encoder model $E_s(.)$ and a decoder model $D_s(.)$ is defined (Figure 2a).
- Each space-time encoder $(E_s(.))$ model consists of two CNN layers, where each layer has two convolution operations. Prior to applying convolutions the features are symmetrically padded. Activations, after both convolution operations, are normalized and rectified using batch-normalization and RELU-activation functions, respectively. The first CNN layer implements a 2D convolution operator with a stride length of 2 and a

spatial-filter of size ($3 \times 3 \times 8$). The second CNN layer uses the same configuration (i.e. the same stride length and spatial-filter size is used), however extracts 16 features instead of 8. This results in a compressed space-time feature sequence ($q_{s,l}^{(1\ldots N)}$) of cardinality ($N \times 8 \times 8 \times 16$). Similarly, the space-time decoder ($D_s(.)$) models also consists of two CNN layers, where in each layer, the deconvolution operation is realized by an upsampling and a convolution operation. Prior to applying convolutions the features are upsampled and symmetrically padded, in both CNN layers. The activations are normalized and rectified, resulting in a reconstructed input sequence of cardinality ($N \times 32 \times 32 \times 2$).

B. SEQUENTIAL MODEL

After training the CAE for each view, the encoders are used for training the sequential model ($H(.)$), and is formally defined as follows:

- For each 2D view $s \in \{\text{rd}, \text{td}, \text{tr}\}$, we define a sequential model (h_s), consisting of a spatial-pyramid pooling layer, a fully connected layer and an LSTM network sharing it's parameters with the LSTM networks of the other views. h_s starts by transforming the compressed space-time codes ($q_{s,l}^{(1\ldots N)}$) to a sequential feature vector. This is accomplished by using a spatial-pyramid pooling operator, where each (8×8) 16-dimensional code map in N, is max-pooled by using three spatial grids of sizes (2×2), (4×4) and (6×6). A sequence of 16-dimensional 56-element feature vector sequence is generated. The pooled feature vector sequence, is transformed via a fully connected layer into a 64-dimensional vector sequence. The activations are normalized and rectified, before being fed to an LSTM network.
- For each view s, we define an LSTM network with 64-dimensional hidden and cell -states, which is unrolled for $N = 12$ frames, to learn from 1 *sec* worth of activity. The LSTM network performs two tasks, 1) vector sequence to sequential feature conversion, and 2) space-time feature fusion. The first task is accomplished by the input weight matrix ($\mathbf{W}_{p,s} = [W_{p,s}^g, W_{p,s}^f, W_{p,s}^i, W_{p,s}^o]$), which is uniquely defined for each space-time sequence by its respective sequential model. The second task requires the definition of a hidden-state weight matrix ($\mathbf{W}_z = [W_z^g, W_z^f, W_z^i, W_z^o]$), which is shared between the different views. For a given normalized and rectified 64-dimensional vector sequence $p_s^{(1\ldots N)}$, the LSTM network is formally defined in Equations 7, 8, 9, 10, 11 and 12 [21]:

$$\tilde{c}_s^n = tanh\left(W_z^g z_s^{n-1} + W_{p,s}^g p_s^n\right) \tag{7}$$

$$f_s^n = \sigma\left(W_z^f z_s^{n-1} + W_{p,s}^f p_s^n\right) \tag{8}$$

$$i_s^n = \sigma\left(W_z^i z_s^{n-1} + W_{p,s}^i p_s^n\right) \tag{9}$$

$$o_s^n = \sigma\left(W_z^o z_s^{n-1} + W_{p,s}^o p_s^n\right) \tag{10}$$

$$c_s^n = f_s^n \odot c_s^{n-1} + i_s^n \odot \tilde{c}_s^n \tag{11}$$

$$z_s^n = o_s^n \odot tanh\left(c_s^n\right) \tag{12}$$

where, z_s^{n-1} is the hidden-state vector in, c_s^n is the cell-state, i_s^n is the input control gate, f_s^n is the forget control gate, $\tilde{c}_s^n$ is the intermediate cell-state vector and $\odot$ represents the element-wise product. Once the N^{th} cell-state (c_s^N) vector is available, the N^{th} hidden-state (z_s^N, Equation 12) vector from each view, is made available for feature concatenation to the subsequent model.

- For the auxiliary features view, we define a sequential model h_{ax} (Figure 2b), consisting of a fully connected layer and an LSTM network where the parameters are shared across views (s). The auxiliary feature sequence ($x_{ax}^{(1\ldots N)}$) is transformed via the fully-connected layer into a 64-dimensional vector sequence, which is followed by a normalization and RELU-activation operations, before being fed to the LSTM network. The hidden-state weight matrix of the auxiliary LSTM network is shared with the LSTM networks of the other views $s \in \{\text{rd}, \text{td}, \text{tr}\}$, thus facilitating mid-level data-fusion of the auxiliary features with the radar data views. The output of the LSTM is mapped to a 64-dimensional context vector w_{ax} via a fully connected layer.
- The sequential feature vectors (z_s^N) from each view (s) are concatenated in a concatenation layer, and mapped to a 64-dimensional feature vector (z_{cnt}) via a fully connected layer. This feature concatenation and transformation is realized in h_{cnt} of Figure 2b. The concatenated and transformed sequential feature vector z_{cnt} is further adapted as follows:

$$\begin{aligned} z_w &= z_{cnt} \odot w_{ax} \\ &= z_{cnt} \odot h_{ax}(x_{ax}^{(1\ldots N)}) \end{aligned} \tag{13}$$

where, $\odot$ is an element-wise product [24]. Finally, z_w is passed to a fully connected layer ($h_{\mathcal{C}}$ in Figure 2b) and a softmax activation function for a multi-class classification task.

V. EXPERIMENTAL RESULTS

A reliable radar-HAR requires an efficient use of the range, Doppler and slow sow-time features in the radar data-cube. The generalizability and robustness of an indoor radar-HAR system for an unseen room and a cluttered environment is largely based on the training methodology, feature selection and the use of externally available auxiliary information from the tracker. Furthermore, environment clutter can greatly affect the feature extraction step, which directly affects the model's performance. In this study we focus on the generalizability and reliability of the DL based models and methods, proposed for an indoor radar-HAR system. As described in detail in [10], the goal is to be able to generalize for an unseen room and participants, with an unseen environment and cluttered background, i.e. the model's ability to perform in the context of *layout-generalization* and *person-generalization*.

We collect data in three rooms, in both cluttered and uncluttered environments. Section V-A details the data collection campaign carried out during this study, while we explain the feature selection approach in Section V-B. Furthermore, we study the ablation of auxiliary feature fusion in Section V-D, which is followed by a detailed discussion about the results in Section V-E.

A. DATASET

This study uses 4 datasets from [10] containing radar data related to three activities of daily living in three different rooms, with two different layouts (as shown in Figure 4).

We briefly describe the characteristics of the four databases, for more details refer to [10]:

- *Action Primitive Oriented (DB1)*: is layout free and focused entirely on the primitive actions. Participants were asked to perform the SD, SU, WA and WT actions with a single chair placed in the center of the room (as shown in Figure 4a). The actions were performed without varying the aspect angle of the participants with the radar sensor. Fifteen subjects were asked to participate, which resulted in over 2000 3D radar data-cube sequences (of 1 *sec*, $N = 12$) for SD and SU classes and over 5000 samples for WA and WT classes (as shown in Table 1).
- *Aspect Angle Oriented (DB2)*: Unlike DB1, here we asked the subjects to perform the actions with four variations in the aspect angle, i.e. the participants were asked to follow paths *A-D* (as shown in Figure 4a). This resulted in another 5000 samples for WA and WT classes (as shown in Table 1) and approximately 2000 3D radar data-cube sequences for SD and SU classes.
- *Multi-path, Aspect-angle, and Shadowing Aspect Oriented (DB3) for validation*: This database was recorded in the same room as DB1 and DB2, however now included background clutter, from the furniture (as shown in Figure 4b). We asked five new participants, which were not included in the DB1 and DB2, to perform the actions. This resulted in 5000 samples for WA and WT classes (as shown in Table 1) and approximately 2000 3D radar data-cube sequences for all classes.
- *Multi-path, Aspect-angle, and Shadowing Aspect Oriented (DB5)*: Unlike DB3, this database was recorded in a new room with a cluttered background (as shown in Figure 4c) and with twenty one new participants (not included in either DB1, DB2 or DB3). This resulted in approximately over 9000 samples for all the classes (as shown in Table 1).

B. 2D SPACE-TIME FEATURE STREAM SELECTION

Single view/stream CNN-LSTM models are trained in an End-to-End (E2E) learning fashion, to select the best feature stream from $l = \{\text{raw}, \text{disp}, \text{diff}, \text{disp-diff}\}$, for each 2D view $s = \{\text{rd}, \text{td}, \text{tr}\}$. Figure 3 illustrates the architecture in which the CNN has the same configuration as the space-time encoder $E_s(.)$ defined in Section IV-A, and the LSTM part is equivalent to the sequential model h_s of Section IV-B, and the classification head h_C. DB1, DB2 and DB5 are used for

TABLE 1. Number of samples per class in each database

	SD	SU	WA	WT
DB1	3,204	2,326	5,844	6,865
DB2	2,144	1,416	5,323	5,052
DB3	2,400	2,125	2,499	2,645
DB5	10,326	10,174	9,402	11,791

TABLE 2. Recognition results on the validation set DB3 for the single view CNN-LSTM based 2D space-time feature stream selection.

	Macro-F1 Score		
	rd	td	tr
raw	0.695	0.665	**0.600**
disp	**0.704**	**0.670**	0.587
diff	0.550	0.560	0.129
disp-diff	0.673	0.661	0.570

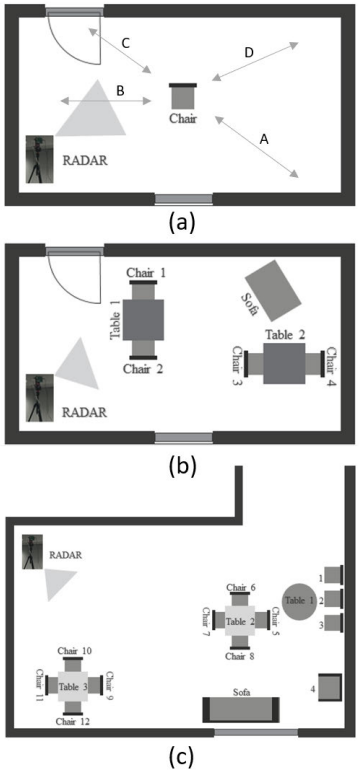

FIGURE 4. The room layouts used for collecting data are shown, room in a) is an uncluttered room where we collected DB1 and DB2, room in b) a cluttered room where we collected DB3, and DB5 was collected in a cluttered room shown in c), which is with a different layout than DB3 (adapted from [10]).

training, while DB3 is used for validation. The results are ranked and the best view-stream pair is chosen for the next steps.

As can be seen from Table 2, the energy dispersion based stream resulted in the best performance, for the 2D range-Doppler ($x_{rd,disp}^{(1...N)}$) and time-Doppler ($x_{td,disp}^{(1...N)}$) views, yielding an average Macro-F1 score of 0.70 and 0.67 respectively. For the 2D time-range view, the original raw feature based stream ($x_{tr,raw}^{(1...N)}$) provided the best Macro-F1 score compared to the other space-time feature based streams ({disp, diff, disp-diff}).

In summary, the energy dispersion based stream was found to be more robust for the 2D range-Doppler and time-Doppler views, while the raw feature stream is preferred for the 2D time-range view. The energy-dispersion data-cubes create focusing in the slow slow-time dimension when the energy dispersion in the Doppler dimension is higher (Equation 3). This, by definition, creates soft-attention in the radar data-cube, when participants are in motion. Whereas, less attention is created when the energy dispersion in the range dimension is higher compared to the energy dispersion in the Doppler dimension (Equation 3). Thus, filtering out events where multipath from the floor or nearby objects is strong, or when the resulting energy in the data-cube is not due to participant's limbs, thorax or head. Soft-attention in the context of a WA action, in time-range, Figure 1 (c,l), and time-Doppler, Figure 1 (c,m), leads to more signal preservation, compared to an SD action, in time-range, Figure 1 (h,l), and time-Doppler, Figure 1 (h,m). The temporal-difference based data-cube highlights the most recent events compared to the previous data-cube, thus creating a data-cube encoding high frequency features (Equation 4), increasing the noise level in the transformed data-cube, and leading to poor discriminative features.

C. MV RADAR-NET TRAINING

Having selected the space-time feature streams for each view, we proceed with the proposed unsupervised training of the space-time models, and the supervised training of the sequential model. The model training is composed of two phases.

During the first phase the space-time feature encoders ($E_s(.)$) are trained in an unsupervised manner. For this step we utilize the encoder-decoder architecture (as shown in Figure 2a). The data from both cluttered and uncluttered domains (i.e. DB1, DB2 and DB5), is utilized for training. For validation the reconstruction on an independent test dataset from a cluttered domain (i.e. DB3) is used. We use a batch size of 16 and a learning rate of $1E-6$. The models are trained for 70 epochs using an ADAM optimizer [14]. The following L1-L2 reconstruction loss is minimized as an objective:

$$\mathcal{L}_{rec}(.) = \sum_{p=1}^{2} ||x_{s,l}^{(1...N)} - \hat{x}_{s,l}^{(1...N)}||_p \tag{14}$$

In the second phase, the sequential model ($H(.)$) is trained with the space-time feature extractors ($E_s(.)$) (as shown in Figure 2b). The parameters of the feature extractors are initialized using the pre-trained encoders from the previous unsupervised training phase and are frozen during the supervised training of the sequential model. Next, the full architecture is fine-tuned with a reduced learning rate of $1E-7$, while the batch size, epochs and optimizer is kept the same as in phase-one. We train the models using DB1, DB2 and DB5, while DB3 is used for validation and minimize the following cross-entropy based focal-loss [18]:

$$\mathcal{L}_{en}(.) = -\frac{1}{|Y|}\sum_{i=1}^{|Y|} y_i \left(1-\hat{y}_i\right)^{\gamma} log\left(\hat{y}_i\right) \tag{15}$$

where, γ is defined as 2.0 and y_i is an element in the class-label vector (Y) representing the i-th class.

D. ABLATION STUDY

To evaluate the benefit of the proposed approach, an ablation study was carried out. We compare the proposed two-step training (with fine-tuning) of the MV RADAR-Net using the view-stream pairs defined by {{rd, disp}, {td, disp}, {tr, raw}} with auxiliary feature fusion, denoted as U2SF-p3/wAux, to the following models:

- The MV RADAR-Net of Figure 2b without auxiliary features trained in an End-to-end fashion (with parameters trained from scratch), denoted as E2E/woAux.
- The MV RADAR-Net of Figure 2b without auxiliary features with frozen $E_s(.)$ and initialized from the pre-trained CAE models of Figure 2a, denoted as U2SF-p2/woAux. Here $E_s(.)$ is used for features inference.
- The MV RADAR-Net of Figure 2b without auxiliary features trained using the two-step training with fine-tunning approach of Section V-C, denoted as U2SF-p3/woAux.
- The MV RADAR-Net of Figure 2b with auxiliary features trained in an End-to-end fashion (with parameters trained from scratch), denoted as E2E/wAux.
- The MV RADAR-Net of Figure 2b with auxiliary features with frozen $E_s(.)$ and initialized using the pre-trained CAE models of Figure 2a, denoted as U2SF-p2/wAux.

The above models were trained on DB1, DB2 and DB5 and evaluated on DB3. Table 3 lists the recognition accuracy on the validation set DB3.

1) WITHOUT AUXILIARY FEATURE FUSION

In the context of an E2E/woAux, the precision, recall and F1-scores of WA and WT classes remain above 0.80, while the model struggles around 0.50 for the SD and SU classes. This results in a baseline average accuracy score, average macro-F1 and weighted average F1 score, of 0.71, 0.7 and 0.71 respectively. Furthermore, the U2SF-p2/woAux and U2SF-p3/woAux did not improve the performance compared to E2E/woAux, except for adding 1% to the overall average accuracy score, during the fine-tuning stage. As a result, the

TABLE 3. Ablation study results of the MV RADAR-Net model on the validation set DB3.

	Precision				Recall				F1-score						
	WA	WT	SD	SU	WA	WT	SD	SU	WA	WT	SD	SU	Acc	Macro	Weighted
E2E/woAux	0.85	0.8	0.59	0.54	0.86	0.86	0.61	0.48	0.86	0.83	0.6	0.51	0.71	0.7	0.71
U2SF-p2/woAux	0.83	0.79	0.61	0.59	0.84	0.85	0.63	0.5	0.84	0.82	0.62	0.54	0.71	0.7	0.71
U2SF-p3/woAux	0.8	0.77	**0.64**	**0.6**	0.87	0.87	0.58	0.51	0.83	0.82	0.61	0.55	0.72	0.7	0.71
E2E/wAux	0.92	**0.89**	0.57	0.57	**0.94**	0.86	**0.68**	0.46	**0.93**	0.87	0.62	0.51	0.75	0.73	0.74
U2SF-p2/wAux	**0.92**	0.88	0.61	0.58	0.9	0.87	0.63	**0.58**	0.91	0.87	0.62	0.58	0.75	0.75	0.75
U2SF-p3/wAux	0.91	0.88	0.62	0.59	0.92	**0.87**	0.64	0.57	0.91	**0.87**	**0.63**	**0.58**	**0.76**	**0.75**	**0.76**

average accuracy, average macro-F1 and weighted average F1 scores, without auxiliary feature fusion were 0.72, 0.7 and 0.71 respectively (as shown in Table 3).

2) WITH AUXILIARY FEATURE FUSION

Auxiliary feature fusion, in E2E/wAux, significantly improved the performance scores for WA and WT classes. The overall recall, precision and F1-score for WA and WT classes, with auxiliary feature fusion, was more than 0.85, however the performance for the SD and SU classes still struggled around 0.50. In contrast, using the U2SF-p2/wAux approach, the performance score of the SD and SU classes improved, resulting in an average recall, precision and F1-score of more than 0.55. Overall, with auxiliary feature fusion, in the context of U2SF-p3/wAux, the baseline average accuracy, average macro-F1 and weighted average F1 scores improved by 5% as shown in Table 3.

E. DISCUSSION

The proposed 3D raw radar data-cube and it's derived representations encode local state of the targets, while with the auxiliary tracking features we are able to encode the global context of the target. This allows the model to infer with high accuracy, the state of the target, i.e. if the targets are in the state of motion or if they are static. This is evident from the ablation study conducted in Section V-D. As shown in Table 3 using the auxiliary feature fusion, we get an overall improvement of 3% in the average Macro-F1 score, in the context of E2E/wAux compared to E2E/woAux. This is further improved by another 2% when the proposed two-step learning approach is used, in the context of U2SF-p3/wAux. These results are also aligned with the projections of the auxiliary features in the UMAP manifold [23], as depicted in Figure 5.

As shown in the Table 3, most degradation in the performance is resulting from the SD and SU classes. This is precisely because the participants were asked to perform SD and SU actions on two chairs with different orientations as shown in Figure 4b. For a chair facing towards the radar, in case of an SD action, while sitting down, the participant first moves away from the radar, upon completion of the action, they would move towards the radar. For an SU action, on the same chair, the order of these events would reverse.

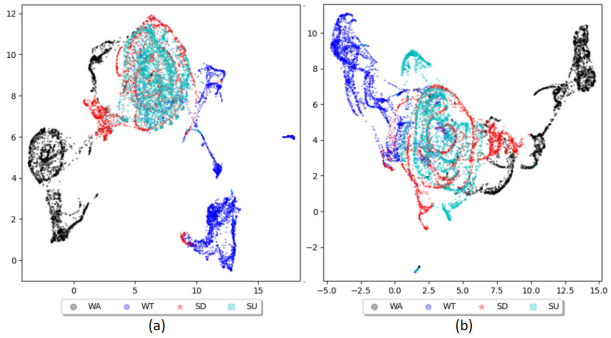

FIGURE 5. UMAP manifold [23] of the auxiliary features (a) and the raw 2D range-Doppler features from DB3 (b), for the four action classes. As shown in (a) the intra-class compactness of the feature space is much better compared to (b), leading to better inter-class separability in (a).

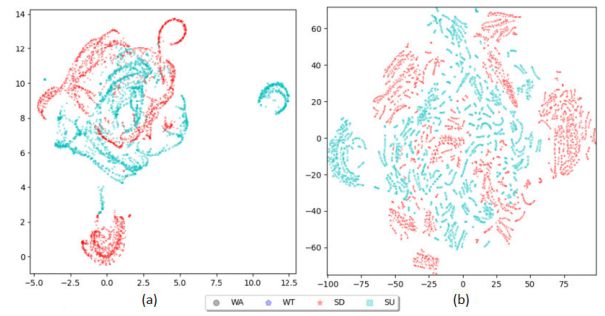

FIGURE 6. UMAP [23] (a) and t-SNE [22] (b) manifolds of the raw range-Doppler maps for SD and SU classes from DB3. As shown in (a) and (b), both SD and SU classes create considerable confusion and thus degrade the overall performance.

However, the problem becomes complicated when the SD and SU are performed on a second chair with different orientation, i.e. facing away from the radar, thus resulting in a confusion between SD and SU classes. This confusion is shown in the UMAP [23] and t-SNE [22] manifolds in Figure 6. Furthermore, apart from the variation in the orientation of the actions, most actions of the participants are occluded behind the tables, especially the SD and SU actions, hence suffering from the shadowing effect due to partial occlusion.

The generalization capability of the learning methods, with a single radar-sensor is still a challenging problem. In this paper we propose to solve this problem by learning from both uncluttered and cluttered domains, and tend to generalize for

TABLE 4. Classification report for WA, WT and SDU classes, with and without auxiliary feature fusion, where MV RADAR-Net was trained using an E2E learning approach. DB1, DB2 and DB5 was used for training while DB3 for validation.

	Precision			Recall			F1-score					
	WA	WT	SDU	WA	WT	SDU	WA	WT	SDU	Acc	Macro	Weighted
E2E/woAux	0.86	0.83	0.83	0.8	0.83	0.86	0.83	0.83	0.85	0.84	0.83	0.84
E2E/wAux	**0.92**	**0.93**	**0.92**	**0.92**	**0.87**	**0.95**	**0.92**	**0.9**	**0.93**	**0.92**	**0.92**	**0.92**

an unseen cluttered environment with unseen participants. However, based on the experimental results, on one hand we could improve the results by fusing the auxiliary features as side information. But on the other hand, it is evident from the results that for some actions, such as SD and SU, with a single radar-sensor, we are fundamentally limited in terms of discriminative features.

One potential solution for this problem could be, to combine the SD and the SU classes, i.e. solving for a multi-class classification problem with Walking Away (WA), Walking Towards (WT), and Sitting Down and standing Up (SDU) classes. The result of this solution with MV RADRA-Net, with and without auxiliary feature fusion, using an E2E learning approach, is shown in Table 4. This approach results in a average accuracy score of 0.84 and 0.92 without and with auxiliary feature fusion, respectively.

VI. CONCLUSION

We study the generalization capability of the learning methods for an unseen cluttered environment, and unseen participants. Our contribution of this research is as follows:

- We propose energy-dispersion based features for a reliable radar-HAR problem. These features were shown to be robust compared to the raw features.
- Apart from the range-Doppler based features, we investigate the utility of the dynamic auxiliary features from the tracker, which allows us to encode the global context of the targets.
- We propose MV RADAR-Net, which utilizes the full radar data-cube, and performs the mid-level multi-view data-fusion using a novel LSTM layer based approach employing a shared hidden-state matrix across multiple views.
- We propose a two-step learning framework, which learns class-label agnostic features, in an unsupervised manner and allows us to reuse pre-trained feature extractors, thus making the overall learning procedure less dependable on the initialization process.

The generalization capability of the learning methods, with a single radar-sensor is still a challenging problem. A potential solution could be, to utilize the absolute context of the targets, however this requires a robust online continual learning framework, to be able to generalize to an unseen environment and will be considered in a future work.

REFERENCES

[1] S. Abdulatif, Q. Wei, F. Aziz, B. Kleiner, and U. Schneider, "Micro-Doppler based human-robot classification using ensemble and deep learning approaches," in *Proc. IEEE Radar Conf. (RadarConf18)*, Apr. 2018, pp. 1043–1048, doi: 10.1109/RADAR.2018.8378705.

[2] M. A. A. H. Khan, R. Kukkapalli, P. Waradpande, S. Kulandaivel, N. Banerjee, N. Roy, and R. Robucci, "RAM: Radar-based activity monitor," in *Proc. IEEE INFOCOM 35th Annu. IEEE Int. Conf. Comput. Commun.*, Apr. 2016, pp. 1–9, doi: 10.1109/INFOCOM.2016.7524361.

[3] S. Björklund, H. Petersson, and G. Hendeby, "Features for micro-Doppler based activity classification," *IET Radar Sonar Navigat.*, vol. 9, no. 9, pp. 1181–1187, 2015.

[4] C. Campbell and F. Ahmad, "Attention-augmented convolutional autoencoder for radar-based human activity recognition," in *Proc. IEEE Int. Radar Conf. (RADAR)*, Apr. 2020, pp. 990–995, doi: 10.1109/RADAR42522.2020.9114787.

[5] R. Q. Charles, H. Su, M. Kaichun, and L. J. Guibas, "PointNet: Deep learning on point sets for 3D classification and segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 77–85, doi: 10.1109/CVPR.2017.16.

[6] V. C. Chen, *The Micro-Doppler Effect in Radar*. Norwood, MA, USA: Artech House, 2011.

[7] C. Ding, H. Hong, Y. Zou, H. Chu, X. Zhu, F. Fioranelli, J. L. Kernec, and C. Li, "Continuous human motion recognition with a dynamic range-Doppler trajectory method based on FMCW radar," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 6821–6831, Sep. 2019, doi: 10.1109/TGRS.2019.2908758.

[8] H. Du, Y. He, and T. Jin, "Transfer learning for human activities classification using micro-Doppler spectrograms," in *Proc. IEEE Int. Conf. Comput. Electromagn. (ICCEM)*, Mar. 2018, pp. 1–3, doi: 10.1109/COMPEM.2018.8496654.

[9] B. Erol and M. G. Amin, "Radar data cube processing for human activity recognition using multisubspace learning," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 55, no. 6, pp. 3617–3628, Dec. 2019, doi: 10.1109/TAES.2019.2910980.

[10] A. Gorji, H.-U.-R. Khalid, A. Bourdoux, and H. Sahli, "On the generalization and reliability of single radar-based human activity recognition," *IEEE Access*, vol. 9, pp. 85334–85349, 2021, doi: 10.1109/ACCESS.2021.3088452.

[11] S. Z. Gurbuz and M. G. Amin, "Radar-based human-motion recognition with deep learning: Promising applications for indoor monitoring," *IEEE Signal Process. Mag.*, vol. 36, no. 4, pp. 16–28, Jul. 2019, doi: 10.1109/MSP.2018.2890128.

[12] Y. He, X. Li, and X. Jing, "A mutiscale residual attention network for multitask learning of human activity using radar micro-Doppler signatures," *Remote Sens.*, vol. 11, no. 21, p. 2584, Nov. 2019.

[13] B. Jokanovic, M. Amin, and B. Erol, "Multiple joint-variable domains recognition of human motion," in *Proc. IEEE Radar Conf. (RadarConf)*, May 2017, pp. 0948–0952.

[14] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[15] M. Li, T. Chen, and H. Du, "Human behavior recognition using range-velocity-time points," *IEEE Access*, vol. 8, pp. 37914–37925, 2020, doi: 10.1109/ACCESS.2020.2975676.

[16] X. Li, Y. He, and X. Jing, "A survey of deep learning-based human activity recognition in radar," *Remote Sens.*, vol. 11, no. 9, p. 1068, Jan. 2019.

[17] J. Lien, N. Gillian, M. E. Karagozler, P. Amihood, C. Schwesig, E. Olson, H. Raja, and I. Poupyrev, "Soli: Ubiquitous gesture sensing with millimeter wave radar," *ACM Trans. Graph.*, vol. 35, no. 4, pp. 1–19, 2016, doi: 10.1145/2897824.2925953.

[18] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," 2017, *arXiv:1708.02002*.

[19] 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 J.*, vol. 18, no. 23, pp. 9669–9681, Dec. 2018, doi: 10.1109/JSEN.2018.2872849.

[20] F. Luo, S. Poslad, and E. Bodanese, "Kitchen activity detection for healthcare using a low-power radar-enabled sensor network," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2019, pp. 1–7, doi: 10.1109/ICC.2019.8761484.

[21] J. Ma, H. Tang, W.-L. Zheng, and B.-L. Lu, "Emotion recognition using multimodal residual LSTM network," in *Proc. 27th ACM Int. Conf. Multimedia*, Oct. 2019, pp. 176–183, doi: 10.1145/3343031.3350871.

[22] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.

[23] L. McInnes, J. Healy, and J. Melville, "UMAP: Uniform manifold approximation and projection for dimension reduction," 2018, *arXiv:1802.03426*.

[24] E. Pei, D. Jiang, and H. Sahli, "An efficient model-level fusion approach for continuous affect recognition from audiovisual signals," *Neurocomputing*, vol. 376, pp. 42–53, Feb. 2020, doi: 10.1016/j.neucom.2019.09.037.
[25] R. Ricci and A. Balleri, "Recognition of humans based on radar micro-Doppler shape spectrum features," *IET Radar, Sonar Navigat.*, vol. 9, no. 9, pp. 1216–1223, Dec. 2015, doi: 10.1049/iet-rsn.2014.0551.
[26] M. S. Seyfioğlu, A. M. Özbayoglu, and S. Z. Gürbüz, "Deep convolutional autoencoder for radar-based classification of similar aided and unaided human activities," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 54, no. 4, pp. 1709–1723, Aug. 2018, doi: 10.1109/TAES.2018.2799758.
[27] Y. Shao, S. Guo, L. Sun, and W. Chen, "Human motion classification based on range information with deep convolutional neural network," in *Proc. 4th Int. Conf. Inf. Sci. Control Eng. (ICISCE)*, Jul. 2017, pp. 1519–1523, doi: 10.1109/ICISCE.2017.317.
[28] A. Shrestha, H. Li, J. Le Kernec, and F. Fioranelli, "Continuous human activity classification from FMCW radar with Bi-LSTM networks," *IEEE Sensors J.*, vol. 20, no. 22, pp. 13607–13619, Nov. 2020, doi: 10.1109/JSEN.2020.3006386.
[29] K. T. Tran, L. D. Griffin, K. Chetty, and S. Vishwakarma, "Transfer learning from audio deep learning models for micro-Doppler activity recognition," in *Proc. IEEE Int. Radar Conf. (RADAR)*, Apr. 2020, pp. 584–589, doi: 10.1109/RADAR42522.2020.9114643.
[30] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. NIPS*, 2017, pp. 1–15.
[31] W. Ye and H. Chen, "Human activity classification based on micro-Doppler signatures by multiscale and multitask Fourier convolutional neural network," *IEEE Sensors J.*, vol. 20, no. 10, pp. 5473–5479, May 15, 2020, doi: 10.1109/JSEN.2020.2971626.
[32] J. Zhu, H. Chen, and W. Ye, "A hybrid CNN–LSTM network for the classification of human activities based on micro-Doppler radar," *IEEE Access*, vol. 8, pp. 24713–24720, 2020, doi: 10.1109/ACCESS.2020.2971064.

HABIB-UR-REHMAN KHALID received the B.Sc. degree in electrical engineering from The University of Lahore (UoL), Pakistan, in 2013, and the master's degree in electrical and electronics engineering from Katholieke Universiteit Leuven (KUL), Belgium, in 2018. He is currently pursuing the Ph.D. degree in engineering sciences with Vrije Universiteit Brussel (VUB) in collaboration with the Interuniversitair Micro-Electronica Centrum (IMEC) under the supervision of Prof. H. Sahli. His main research interests include digital signal processing, real-time embedded and control systems, multi-sensor data-fusion, energy-efficient machine learning, and heterogeneous sensor-based human activity recognition.

ALI GORJI received the B.Sc. and M.Sc. degrees from the Amirkabir University of Technology, Iran, in 2005 and 2008, respectively, and the Ph.D. degree from McMaster University, Canada, in 2012. Previously, he was a Postdoctoral Fellow at the Department of Electrical and Computer Engineering, University of Toronto, Toronto, Canada. His research interests include statistical signal processing, detection and estimation theory, target tracking, and their applications in radar, sensor systems, and robotics.

ANDRÉ BOURDOUX (Senior Member, IEEE) received the M.Sc. degree in electrical engineering from the Université Catholique de Louvain, Belgium, in 1982. He joined the Interuniversity Microelectronics Centre (IMEC), in 1998, where he is the Principal Member of Technical Staff with the Advanced RF Research Group. He is a system level and signal processing expert for both the mm-wave wireless communications and radar teams. He has more than 15 years of research experience in radar systems and 15 years of research experience in broadband wireless communications. He holds several patents in these fields. He is the author and coauthor of over 180 publications in books and peer-reviewed journals and conferences. His research interests include advanced architectures, signal processing and machine learning for wireless physical layer, and high-resolution 3D/4D radars.

SOFIE POLLIN (Senior Member, IEEE) received the Ph.D. degree (Hons.) from Katholieke Universiteit Leuven (KU Leuven), in 2006. From 2006 to 2008, she continued her research on wireless communications, energy-efficient networks, cross-layer design, coexistence, and cognitive radio at UC Berkeley. In 2008, she returned to the Interuniversity Microelectronics Centre (IMEC) to become the Principal Scientist at the Green Radio Team. She is currently an Associate Professor with the Electrical Engineering Department, KU Leuven. Her research interests include networked systems that require networks that are ever more dense, heterogeneous, battery powered, and spectrum constrained. She is a fellow of BAEF and a Marie Curie Fellow.

HICHEM SAHLI received the degree in mathematics and computer science, the DEA degree in computer vision, and the Ph.D. degree in computer sciences from the Télécom Physique Strasbourg, France. Since 2000, he has been a Professor with the Department of Electronics and Informatics (ETRO) and a Scientist at the Interuniversity Microelectronics Centre (IMEC). He coordinates the Audio-Visual Signal Processing Laboratory (AVSP) within ETRO. AVSP conducts research on applied and theoretical problems related to machine learning, signal and image processing, and computer vision. The group explores and capitalizes on the correlation between speech and video data for computational intelligence where efficient numerical methods of computational engineering are combined with the problems of information processing.