A Hybrid CNN–LSTM Network for the Classification of Human Activities Based on Micro-Doppler Radar

JIANPING ZHU[1], HAIQUAN CHEN[2], AND WENBIN YE [ID][1], (Member, IEEE)
[1]College of Electronic and Information Engineering, Shenzhen University, Shenzhen 518060, China
[2]College of Physics and Optoelectronic Engineering, Shenzhen University, Shenzhen 518060, China

Corresponding author: Wenbin Ye (yewenbin@szu.edu.cn)

ABSTRACT Many deep learning (DL) models have shown exceptional promise in radar-based human activity recognition (HAR) area. For radar-based HAR, the raw data is generally converted into a 2-D spectrogram by using short-time Fourier transform (STFT). All the existing DL methods treat the spectrogram as an optical image, and thus the corresponding architectures such as 2-D convolutional neural networks (2D-CNNs) are adopted in those methods. These 2-D methods that ignore temporal characteristics ordinarily lead to a complex network with a huge amount of parameters but limited recognition accuracy. In this paper, for the first time, the radar spectrogram is treated as a time sequence with multiple channels. Hence, we propose a DL model composed of 1-D convolutional neural networks (1D-CNNs) and long short-term memory (LSTM). The experiments results show that the proposed model can extract spatio-temporal characteristics of the radar data and thus achieves the best recognition accuracy and relatively low complexity compared to the existing 2D-CNN methods.

INDEX TERMS Radar signal processing, human activity recognition, convolutional neural network, recurrent neural network, deep learning.

I. INTRODUCTION

Human activity recognition (HAR) provides excellent potential for various applications, including personal health systems (PHS), human-computer interaction (HCI), and anti-terrorism monitoring [1]–[3]. There are generally two types of HAR: video-based HAR and sensor-based HAR [4]. Video-based HAR takes advantages of the videos or images from optical cameras to resolve human motion, whereas sensor-based HAR relies on the data from smart sensors such as a gyroscope, accelerometer, and radars. Given protecting individual privacy, sensor-based HAR is becoming more popular and extensively used. Among various monitoring sensors, radar-based devices offer unique advantages, such as penetrating opaque objects, adapting to any lighting conditions, and working around the clock [5]. Hence, radar-based HAR methods are attracting growing interests.

The activity recognition using a radar depends typically on the micro-Doppler effect caused by the vibration or rotation of an object, which contains the information of its range, velocity, and other properties [6]. Since the echo based on Doppler radar contains the time-varying kinematic information of human motion, they can be used for activity recognition. Most methods currently use time-frequency analysis to obtain the time-Doppler map. Micro-Doppler signatures (also called time-Doppler map) can be regarded as a statistical pattern. Therefore, how to extract useful features from micro-Doppler signatures becomes a crucial factor for recognition or identification. Conventional approaches [7]–[10] adopted machine learning algorithms in classification, such as multilayer perceptron, principal component analysis (PCA), support vector machine (SVM) and linear discriminant analysis. In these methods, manual features extracted from micro-Doppler signatures are typically used as input to the classifier. However, the efficiency of these features is limited by prior knowledge and the complexity of classification problems [11]. Due to those limitations,
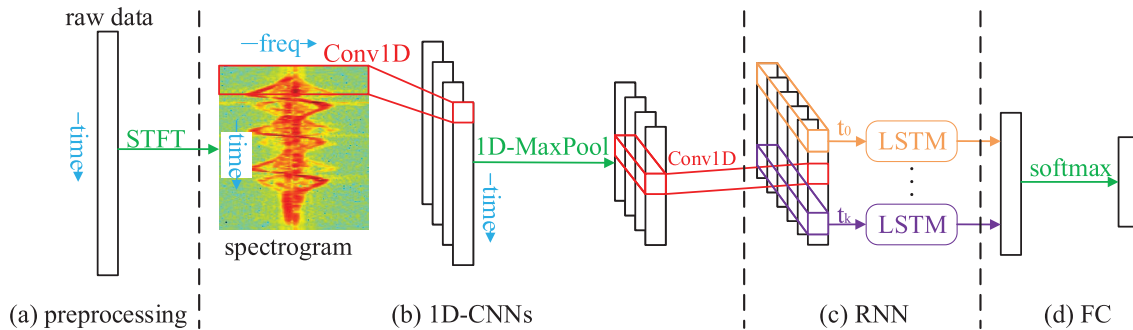
The associate editor coordinating the review of this manuscript and approving it for publication was Chengpeng Hao [ID].

FIGURE 1. This is the overall architecture of the proposed method.

conventional approaches cannot afford most daily HAR tasks.

In recent years, HAR has made remarkable advance by applying deep learning (DL) algorithms [12]–[18] to micro-Doppler signatures. Instead of handcrafted feature selection, DL makes it possible to extract features automatically. Compared to traditional ML methods, feature extraction and classification processes are usually performed simultaneously in DL models. For instance, [12], [13] proposed similar deep convolutional neural networks (DCNNs) to classify different groups of physical activities and achieved satisfying results. Nevertheless, these DCNN methods still suffer for the small amount of data available, which limits the depth of CNNs and the generalization ability, leading to weakened performance. A feasible solution to insufficient sample support is to use sparsely connected layers. Convolutional autoencoders (CAEs) [14] has been proposed that uses unsupervised pretraining to alleviate the demand for training data. Despite this, another study [15] showed that transfer learning [16]–[18] was superior to CAEs when less than 550 samples were acquirable.

However, all of the methods [12]–[16], [18] mentioned above ignored the fact that the raw signals received by radars are complex time-series data, whose amplitude and phase could be connected with the kinematics of the observed target [5]. In general, the raw data is preprocessed by a short-time Fourier transform (STFT) to obtain spectrograms. Then, the state-of-art 2D-CNN networks in computer vision are used to classify the time-Doppler maps. Although spectrograms can be treated like optical images, each pixel is time and frequency samples. Compared to the image, which is spatially related, the spectrogram has a strong temporal correlation. Hence, conventional 2-D methods can mainly learn spatial features, resulting in a complex network with a huge amount of parameters but limited recognition accuracy. In this paper, we take into consideration of the temporal characteristics of radar signals and propose a DL architecture that combines 1D-CNNs and recurrent neural network (RNNs). More detailed, a 1D-CNN is firstly employed to extracting spatial features from the spectrograms and the long short-term memory (LSTM) is then introduced to learn global

time-dependent information. Our work is one of the first efforts that the time-Doppler map is treated as a time-sequential vector, and thus 1D-CNN and LSTM are used in HAR to extract spatio-temporal features of radar data. The proposed network is trained and tested on a seven-class HAR data set. It can achieve the best accuracy and relatively low complexity compared to other networks.

The rest of this paper is organized as follows. The design of the proposed network architecture is described in details in Section II. The experiment results is presented in Section III. Finally, Section IV gives the conclusions of this paper.

II. PROPOSED METHOD

Fig. 1 shows the overall network architecture for HAR tasks, which consists of an STFT for data preprocessing, 1D-CNNs for local feature learning, an LSTM layer for global temporal information extraction, and a fully connected layer for classification. First of all, we perform the N points STFT on the raw data to obtain a spectrogram. Next, the spectrogram is treated as a 1-D time series with multiple channels and is fed to a neural network composed of CNNs and RNNs. The CNN part has two 1-D convolution layer and the first one is followed by a max pooling layer for downsampling. The 1D-CNN is performed in the time dimension, which extracts features of adjacent time frames and can preserve the temporal information of the spectrogram. After processed by STFT and 1D-CNNs, the feature map can also be seen as a 1-D time series with multiple channels, which retains unbroken temporal characteristics. Therefore, we use an LSTM layer to handle the global temporal information. Finally, the LSTM layer is connected to a softmax layer to get prediction results.

A. STFT

The raw data is a series of 1-D time-varying signals, including I/Q channels in our case. Most studies in HAR utilize a time-frequency (TF) transform to obtain suitable input for DNNs. The STFT is an efficient linear TF algorithm that transforms time-dependent signals into the time-frequency domain during each short-time section. Although some DCNN methods can abandon the STFT to implement an end-to-end network, we believe that adopting STFT is beneficial to improve the

expression ability of the model with low sample support of radar data.

As early as 1946, Gabor [19] proposed STFT (also called Gabor transform), which adds a Gauss window to the traditional Fourier transform. In 1992, Mann [20] applied STFT to the radar signal processing, and since then, STFT has been widely used in this field. In practice, the STFT is first to split a long-term signal $x[n]$ into M segments through a time window with length L and overlap K. Then the N points fast Fourier transform (FFT) is performed on each segment. A series of FFT results are obtained by sliding the window function, and the results are arranged to get a two-dimensional (2-D) representation $X_{M \times N}$, which is referred to as the spectrogram (i.e. the energy spectral density).

B. 1D-CNN
Convolutional neural networks, one of the most popular DL algorithms, have been successfully introduced into time series processing like HAR. Instead of learning only shallow features in a heuristic or handcrafted way, DL can extract high-level features automatically and be more capable of some complex tasks [21]. Compared to other models, CNNs offer the advantage of local dependency [4]. It means the adjacent points on the feature map are tend to be correlated, which coincides with the radar signal.

Since the spectrogram obtained by STFT can be seen as a 2-D virtual image, most approaches build models with three convolutional layers connected by two fully connected (FC) layers. In this paper, the spectrogram is treated as a 1-D (time dimension) signal with multiple channels (frequency dimension), where 1D-CNN is applied to. Thus, the temporal characteristics of the spectrogram could be better preserved and exploited later in LSTM. Furthermore, 1D-CNN offers lower computational complexity. As shown in Fig. 1(b), the proposed network contains two 1-D convolutional layers. Both layers use ReLU functions as the nonlinear activation functions. The first convolutional layer is followed by a max pooling layer with a size of 2. According to the design rules in [22], the number of filters in the second convolutional layer is twice that of the first layer due to the downsampling caused by the pooling layer. In the proposed architecture, satisfactory results can be obtained in HAR tasks by applying shallow one-dimensional convolution. Beside the depth, other network parameters are further optimized, such as the number of filters and the kernel size.

C. LSTM
Recurrent neural networks, especially LSTM, play an essential role in natural language processing. Different from feedforward networks, RNNs contain feedback loops and are capable of handling tasks based on the temporal sequence. To solve the issue of gradient vanishing or explosion that may occur in training traditional RNNs, researchers proposed the LSTM as shown in Fig. 2, which commonly includes a cell, a forget gate, an input gate, and an output gate. In some radar-based dynamic recognition problems, LSTM was
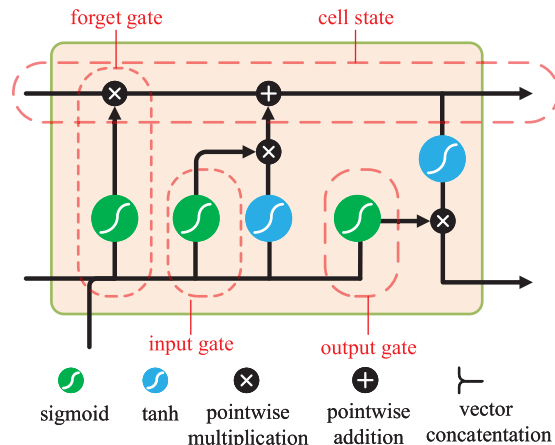

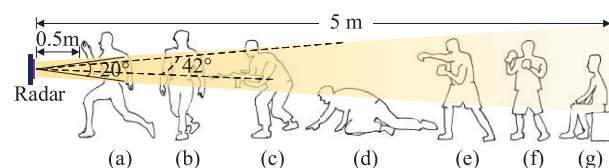**FIGURE 2. The schematic diagram of LSTM.**

FIGURE 3. The setup for data collection of seven human activities. (a) running (b) walking (c) walking while holding a stick (d) crawling (e) boxing while moving forward (f) boxing while standing in place (g) sitting still.

utilized to model dynamic processes with the unsegmented data flow.

Our study reveals that feature extractors and classifiers combined with 1D-CNNs and LSTM outperform DCNN based ones. In addition, adopting LSTM makes it possible to build a much shallower network, which provides outstanding performance while claims fewer parameters. Since the output of 1D-CNNs can be regarded as feature vectors arranged in the time dimension, we feed these time-dependent feature vectors into the LSTM units to learn contextual time information.

III. RESULTS
In this section, the details of data collection and system implementation is introduced first. Then, the performance of the proposed method with different network parameters is discussed. Finally, we give the comparison between our method and the existing state-of-art models.

A. DATA COLLECTION AND SYSTEM IMPLEMENTATION DETAILS
Fig. 3 displays the setup for human activity measurements. The original data for HAR is measured by an Infineon's Sense2GoL Doppler radar operating from 24.05 GHz to 24.25 GHz in free space. The -3dB beamwidth of the radar is 20 degrees in the vertical direction and 42 degrees in the horizontal direction. The measurement range is between 0.5 m and 5 m. There are seven types of human activities designed the same as [8]: (a) running, (b) walking, (c) walking

TABLE 1. The action type and the number of groups.

Action Type	Number of Groups
(a)running	2075
(b)walking	2367
(c)walking while holding a stick	2064
(d)crawling	1972
(e)boxing while moving forward	1967
(f)boxing while standing in place	2429
(g)sitting still	2049
Total	14923

TABLE 2. The initial architecture of the network.

layers	output shape
Input	6000×2
STFT L=51, overlap=12, pad_to=150	153×150
1D-Conv1 $L_1 \times C$, **ReLU**, **1D-MaxPool** 2	$76 \times C$
1D-Conv2 $L_2 \times 2C$, **ReLU**	$76 \times 2C$
LSTM N_3	N_3
FC-softmax 7	7

while holding a stick, (d) crawling, (e) boxing while moving forward, (f) boxing while standing in place, and (g) sitting still. The data are collected by seven subjects, including five males and two females. Note that all data was collected in a restricted environment. The data collection details are listed in Table 1. Each acquisition process lasts for three seconds, with a sampling rate of 2 kHz.

In this experiment, all models are implemented on a server equipped with 64G memory, a 2.5GHz Intel(R) Xeon(R) E5-2678 v3 CPU, and an NVIDIA GeForce GTX1080Ti graphics card. Each model is trained in Python using Keras based on the backend of Tensorflow. We use Adaptive moment estimation as the optimizer for back propagation with a batch size of 32. The learning rate is set to 0.0001 and will be reduced by half if there is no improvement of the test accuracy for 20 epochs. The early stop mechanism is adopted to stop training smartly. We use the 5-fold cross-validation to test the performance of all the architectures mentioned in this section. The whole data set is divided into five subsets without intersections. Each time, one of the five parts is used as the validation set and the remaining parts as the training set. The average accuracy of five folds is accepted as the final result. By the way, in order to reduce the impact of the individual collector, each category of the data collected by every subject is also split into five folds correspondingly. The spectrograms of seven actions are illustrated in Fig. 4.

B. OPTIMIZATION OF THE NETWORK PARAMETERS
The initial architecture of the network is described in Table 2. As a complex time series, the input data has two channels, each with 6000 samples. First of all, a 150 points STFT is performed on the raw data. The time window length is set to 25.5 ms, corresponding to 51 samples calculated by the sampling rate of 2 kHz. To retain more original infor-

(a) runing   (b) walking

(c) walking while holding a stick   (d) crawling

(e) boxing while moving forward   (f) boxing while standing in place

(g) sitting still
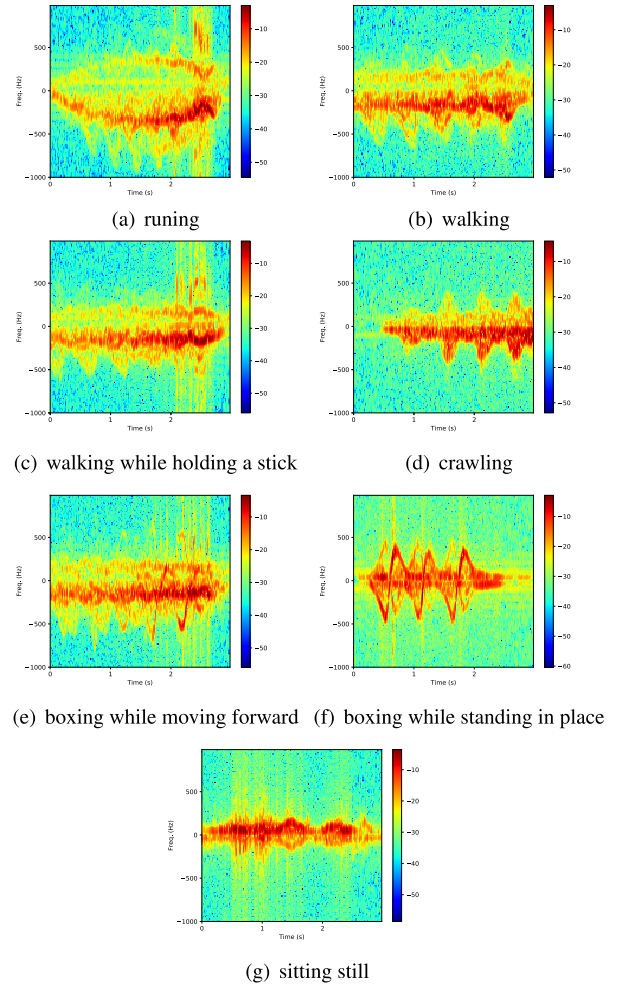
FIGURE 4. The spectrograms of seven actions.

mation, we adopt an overlap of 6 ms and obtain the output spectrogram with 153 time pixels and 150 frequency pixels. Then, the spectrogram is fed to two 1-D convolutional layers. The first layer has C filters with the length of L_1, while the second one has $2C$ filters with the length of L_2. Only the first layer is connected to a max-pooling layer with the length of 2. Both 1-D convolutional layers use ReLU as the activation function and set padding to the same. Finally, we use the LSTM layer with N_3 units to learn global time features and a seven-class softmax layer as a classifier.

Since the shallow initial network structure has been able to give satisfactory performance, we first fix the network depth to reduce the scope of the searching space and seek the optimal kernel length (L_1, L_2), width (C) and the number of LSTM units (N_3). After that, to explore the impact of the number of 1-D convolutional layers, we also test the accuracy of the models with different depths. The influence of network parameters on test accuracy is summarized in Table 3 and Table 4.

1) IMPACT OF THE NETWORK WIDTH AND LSTM NUITS
Base on the initial network structure, we first fix the kernel lengths of the two 1-D convolutional layers at

TABLE 3. Comparison results of the proposed network with different widths and kernel sizes. Bold rows indicate the model with the highest recognition accuracy under certain conditions.

Kernel size $(L_1 - L_2)$	Width $(C - 2C - N_3)$	Acc.(%)	Parameters
5-3	16-32-32	95.23	22k
	16-32-64	96.78	38k
	32-64-32	96.42	42K
	32-64-64	97.42	63k
	64-128-64	97.95	122k
	64-128-128	98.28	205k
	128-256-128	98.41	392k
	128-256-256	98.51	721k
	256-512-256	**98.65**	**1.37M**
5-3	32-64-64	97.42	63k
7-3		97.45	73k
7-5		97.54	77k
9-3		97.30	82k
9-5		**97.69**	**87k**
9-7		97.41	91k
5-3	64-128-128	**98.28**	**205k**
7-3		97.92	122k
7-5		98.26	240k
9-3		98.20	243k
9-5		98.14	260k
9-7		98.10	276k

TABLE 4. Comparison results of the proposed network with different depths.

Depth of CNN	Width	Acc.(%)	Parameters
3	16-32-64-64	97.48	53k
	32-64-128-64	97.80	104k
	32-64-128-128	98.22	187k
	64-128-256-128	98.40	369k
	64-128-256-256	98.48	698k
6	k=32	98.13	369k
	k=64	98.49	1.4M

5 and 3 respectively. Then, we change the widths of the network (including the number of LSTM units) to see their effects. As shown in Table 3, the number of filters (C) is doubled every time from 16 to 256. The number of LSTM units (N_3) is usually half or the same as the kernel width of the previous layer ($2C$). From the table, we can find that an obvious improvement in the performance of the model can be seen by increasing the number of filters and LSTM units. A similar conclusion that wider networks show more powerful performance is also made in wide residual networks (WRN) [23] verified on computer vision data sets. Such phenomenon demonstrates that the proposed method is well-suited for time-dependent series like radar data. When the number of filters, as well as the number of weight parameters, increased significantly, no evidential overfitting can be seen. This allows designers to widen the network for higher
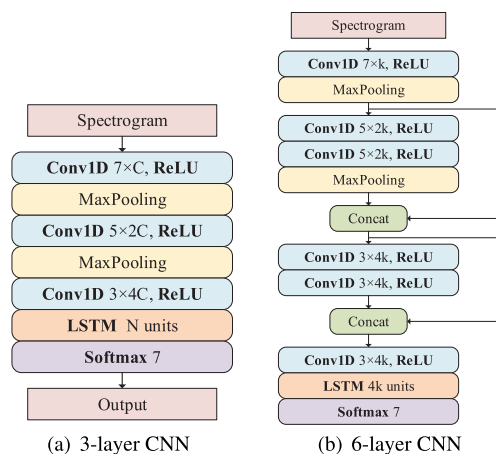

(a) 3-layer CNN

(b) 6-layer CNN

FIGURE 5. Architecture of 1D-CNN-LSTM with 3 convolutional layers and 6 convolutional layers.

demand for accuracy. But it does not mean that the performance could be improved indefinitely. When the width of the first convolutional layer is larger than 128, little improvement can be observed. In addition, the parameter scale and computational complexity will shoot up as the number of filters and LSTM units increase. Hence, the parameter optimization is a trade-off between the pursuit of high performance and the cost of model complexity.

2) IMPACT OF THE KERNEL SIZE

Currently, the optimization of kernel size is not an issue for 2D-CNN because of the technique of replacing a large 2-D kernel with several small kernels such as 3×3 mentioned in Inception-v3 [24]. Unfortunately, this solution is not suitable for 1-D convolution as it soars the number of parameters. In view of this, we also change the kernel lengths with other network parameters fixed. More detailed, the kernel length is set from 3 to 9, and L_2 always tends to be smaller than L_1. As shown in Table 3, we can see the accuracy varies slightly as the kernel size of 1D-CNN changes. It seems hard for us to find a rule for this kind of change due to kernel size. However, instead of the rules which are absolute, we can still find some basic ideas. Firstly, the effect on the accuracy caused by the kernel size is not a key factor. Secondly, the difference between the kernel size of two adjacent layers should not be too large. Finally, the optimal kernel size seems to be related to other parameters of the network. For example, a slimmer network tends to require a larger kernel, while a wider network requires a smaller kernel.

3) IMPACT OF THE NETWORK DEPTH

To explore the effect of the number of 1-D convolutional layers, we investigate in two cases. First, we directly convert the original two-layer 1D-CNN into a three-layer one and add a max-pooling layer before the last convolutional layer, as shown in Fig. 5(a). The number of filters in the last layer is twice that of the previous layer. We compare the performance of the three-layer 1D-CNN model with different

widths. After that, we build a much deeper model with six convolutional layers, which can be seen in Fig. 5(b). The recognition accuracy of models with different depths is displayed in Table 4. For the three-layer model, we find some improvements in accuracy when the network width is narrow. However, as the width increases, the effect of depth weakens until it disappears. Even when the convolution part of the network is deepened to 6 layers, the accuracy improvement of the network is not obvious. In short, although the depth of convolutional layers have some impact on accuracy, it does not play a decisive role. In other words, for the proposed method, increasing the width of the convolutional layers is more effective than make the network deeper. As mentioned earlier, the spectrogram of the radar signal can be seen as a multi-channel time series. More specifically, the frequency dimension of the spectrogram used in this experiment is 150 pixels (considered as 150 channels). Due to the characteristic of 1-D convolution, if the number of filters in the first layer is too small (much less than 150), it may cause the information loss. We think this is why the performance is more sensitive to the network width. Meanwhile, too many convolutional layers may destroy the time characteristics of the input signal, which is not conducive to the learning of LSTM. In addition, with the the network getting deeper, the network is more likely to be overfitting, and it takes longer to train the network.

4) ANALYSIS OF THE OPTIMIZED NETWORK

Based on the discussion about the impact of network parameters, we find that increasing the number of filters as well as LSTM units (width) is the most effective way to improve the performance of the 1D-CNN-LSTM model. But it will also cause a surge in the amount of parameters. Balancing the two factors, we choose a network with a kernel size of 5-3 and a width of 32-64-64 (the last bolded row in Table 3) for subsequent discussion. The classification performance of the selected model is evaluated in fold 1 by precision, recall and F1-score. The results are summarized in Table 5 and the confusion matrix is illustrated in Fig. 6. From the table and the confusion matrix, we find that the three activities (b)walking, (c)walking while holding a stick, and (e)boxing while moving forward are relatively easy to be confused. Since the spectrograms of the three activities are similar to some extent, we think it can be considered as a deficiency of time-Doppler map. To further improve the accuracy, we believe that it may be possible to change the form of input signals, such as adding the time-range map, but this is not in the extent of this paper.

C. RESULTS OF DIFFERENT MODELS

To demonstrate the superiority of this method for processing radar data, we compared the performance of our method with recently published ones [7], [8], [13], [15], [17]. Among them, the first two used traditional machine learning methods, while others adopted deep learning algorithms. In [7], a feed-forward artificial neural network, with only one hidden layer, was applied to classify human activities. In [8], Kim et al.

TABLE 5. The classification report of the network with a kernel size of 5-3 and a width of 32-64-64 (the last bolded row in Table 3).

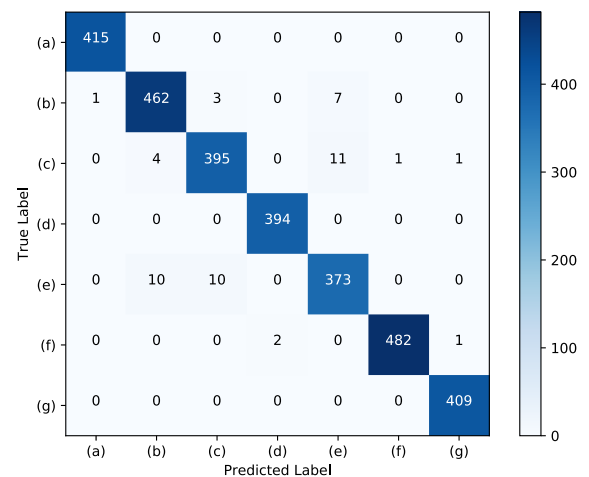|	Precision	Recall	F1-score	Support
(a)	1.00	1.00	1.00	415
(b)	0.97	0.98	0.97	473
(c)	0.97	0.96	0.96	412
(d)	0.99	1.00	1.00	394
(e)	0.95	0.95	0.95	393
(f)	1.00	0.99	1.00	485
(g)	1.00	1.00	1.00	409
Avg/Total	0.98	0.98	0.98	2981

FIGURE 6. The confusion matrix of the network with a kernel size of 5-3 and a width of 32-64-64 (the last bolded row in Table 3).

TABLE 6. Comparison of the results of different models.

Model	Acc.(%)	Parameters
MLP [7]	66.15	483
SVM [8]	67.67	-
CNN [13]	95.34	738k
CAE [15]	94.88	1.98M
ResNet-18 [17]	94.79	11M
1D-CNN-LSTM	**98.28**	**205k**

used a support vector machine with the Gaussian kernel. As early studies in the field of radar-based HAR, the main contribution of these methods was to confirm the feasibility of using micro-Doppler signatures. Thus, most of the following researches were based on micro-Doppler signatures. In [13], a DNN with five convolutional layers was trained on the resulting spectrograms. For a more challenging 12-class problem, [15] employed an unsupervised pre-training method based on a CAE with three convolutional layers and three deconvolutional layers. In [17], a transfer-learned residual network, composed of 18 residual blocks, was introduced to classify a 6-class data set.

The comparison results of different models are shown in Table 6 and the curves of test recognition accuracy varying
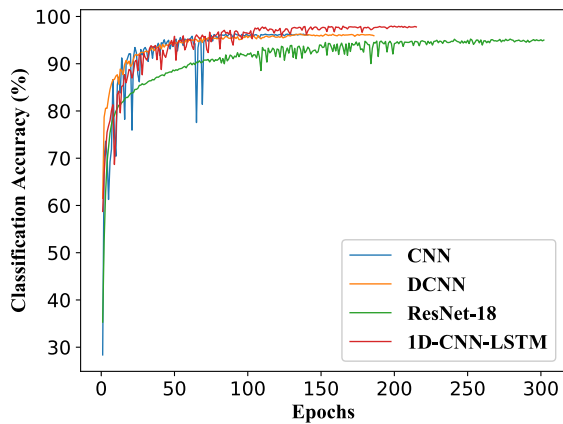
FIGURE 7. Curves of test recognition accuracy varying with the number of epochs.

with the number of epochs are illustrated in Fig. 7. Since the spectrograms obtained by STFT are not as meaningful and well designed as the handcrafted features mentioned in [7], [8], the traditional methods (MLP and SVM) give lower classification accuracy. That is to say, the traditional methods heavily rely on manual extracted features, which leads to poor generalization ability. When it comes to the state-of-art DL methods, their performance is much better with the accuracy varying from 94.79% to 95.34%. The selected architecture of our method for comparison (with the kernel size of 5-3 and the width of 32-64-64) offers the accuracy of 98.28% over the 2-D methods without LSTM by 2.94% to 3.49%. Furthermore, the proposed network has only 205k parameters, which is reduced by 3.6 to 53.6 times.

IV. CONCLUSION

In this paper, we introduced a deep learning model composed of one-dimensional CNNs and RNNs for human activity classification. The model achieved an accuracy of 98.28% verified on a seven-class data set pre-processed with STFT. We optimized the network by searching for parameters such as kernel sizes, widths, and depths. The test accuracy can be improved by building a wider network or adjusting the kernel size. Simultaneously, the performance of our method is better than state-of-art networks. The results indicate that the proposed network architecture is well-suited to extract global temporal features in radar signals. Besides, our network is a kind of efficient models with a small number of parameters.

REFERENCES

[1] O. D. Lara and M. A. Labrador, "A survey on human activity recognition using wearable sensors," *IEEE Commun. Surveys Tuts.*, vol. 15, no. 3, pp. 1192–1209, 3rd Quart., 2013.

[2] O. D. Incel, M. Kose, and C. Ersoy, "A review and taxonomy of activity recognition on mobile phones," *BioNanoScience*, vol. 3, no. 2, pp. 145–171, Jun. 2013.

[3] A. Bulling, U. Blanke, and B. Schiele, "A tutorial on human activity recognition using body-worn inertial sensors," *CSURACM Comput. Surv.*, vol. 46, no. 3, pp. 1–33, Jan. 2014.

[4] J. Wang, Y. Chen, S. Hao, X. Peng, and L. Hu, "Deep learning for sensor-based activity recognition: A survey," *Pattern Recognit. Lett.*, vol. 119. pp. 3–11, Mar. 2019.

[5] 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.

[6] V. C. Chen, D. Tahmoush, and W. J. Miceli, *Radar Micro-Doppler Signatures: Processing and Applications*. Edison, NJ, USA: IET, 2014.

[7] Y. Kim and H. Ling, "Human activity classification based on micro-Doppler signatures using an artificial neural network," in *Proc. IEEE Antennas Propag. Soc. Int. Symp.*, Jul. 2008, pp. 1–4.

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

[9] J. Li, S. L. Phung, F. H. C. Tivive, and A. Bouzerdoum, "Automatic classification of human motions using Doppler radar," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jun. 2012, pp. 1–6.

[10] W. Li, B. Xiong, and G. Kuang, "Target classification and recognition based on micro-Doppler radar signatures," in *Proc. Progr. Electromagn. Res. Symp.-FALL (PIERS-FALL)*, Nov. 2017, pp. 1679–1684.

[11] S. Z. Gürbüz, B. Erol, B. Çağlıyan, and B. Tekeli, "Operational assessment and adaptive selection of micro-Doppler features," *IET Radar, Sonar Navigat.*, vol. 9, no. 9, pp. 1196–1204, Dec. 2015.

[12] Y. Kim and T. Moon, "Human detection and activity classification based on micro-doppler signatures using deep convolutional neural networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 1, pp. 8–12, Jan. 2016.

[13] T. S. Jordan, "Using convolutional neural networks for human activity classification on micro-Doppler radar spectrograms," *Proc. SPIE, Sensors, Command, Control, Commun., Intell. Technol. Homeland Secur., Defense, Law Enforcement Appl.*, vol. 9825, May 2016, Art. no. 982509.

[14] M. S. Seyfioglu, A. M. Özbayoğlu, and S. Z. Gurbuz, "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.

[15] M. S. Seyfioğlu and S. Z. Gürbüz, "Deep neural network initialization methods for micro-Doppler classification with low training sample support," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 12, pp. 2462–2466, Dec. 2017.

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

[17] 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.

[18] M. S. Seyfioglu, B. Erol, S. Z. Gurbuz, and M. G. Amin, "DNN transfer learning from diversified micro-Doppler for motion classification," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 55, no. 5, pp. 2164–2180, Oct. 2019.

[19] D. Gabor, "Theory of communication. Part 1: The analysis of information," *J. Inst. Electr. Eng.-III, Radio Commun. Eng.*, vol. 93, no. 26, pp. 429–441, Nov. 1946.

[20] S. Mann and S. Haykin, "'chirplets' and 'warblets': Novel time–frequency methods," *Electron. Lett.*, vol. 28, no. 2, p. 114, 1992.

[21] J. Yang, M. N. Nguyen, P. P. San, X. L. Li, and S. Krishnaswamy, "Deep convolutional neural networks on multichannel time series for human activity recognition," in *Proc. 24th Int. Joint Conf. Artif. Intell.*, 2015.

[22] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: https://arxiv.org/abs/1409.1556

[23] S. Zagoruyko and N. Komodakis, "Wide residual networks," 2016, *arXiv:1605.07146*. [Online]. Available: https://arxiv.org/abs/1605.07146

[24] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2818–2826.

JIANPING ZHU received the B.E. degree from the School of Electronic Science and Technology, Shenzhen University, Shenzhen, China, where he is currently pursuing the M.S. degree with the College of Electronic and Information Engineering. His research interests include radar signal and image processing, deep learning, and artificial intelligence.
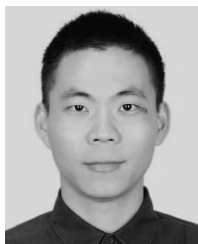
HAIQUAN CHEN received the B.E. degree from the School of Electronic Science and Technology, Shenzhen University, Shenzhen, China, where he is currently pursuing the M.S. degree with the School of Optoelectronic Engineering. His research interests include radar signal and image processing, machine learning, and artificial intelligence.

WENBIN YE (Member, IEEE) received the B.S. degree in microelectronics from Sichuan University, Chengdu, China, in 2009, and the Ph.D. degree from Nanyang Technological University, Singapore, in 2014. From 2014 to 2015, he was a Project Officer with Nanyang Technological University. Since 2015, he has been with the College of Electrical Science and Technology, Shenzhen University, where he is currently an Associate Professor. His research interests include digital filter design, nonuniformly sampled data processing, machine learning, and biomedical signal processing.

. . .