IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received August 17, 2020, accepted September 24, 2020, date of publication October 7, 2020, date of current version October 22, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3029247

Human Motion Serialization Recognition With Through-the-Wall Radar

XIAQING YANG[1], (Student Member, IEEE), PENGYUN CHEN[1], MINGYANG WANG[2], (Member, IEEE), SHISHENG GUO[1], (Member, IEEE), CHAO JIA[1], AND GUOLONG CUI[1], (Senior Member, IEEE)

[1]School of Information and Communication Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China
[2]Southwest Institute of Electronic Technology, Chengdu 610036, China

Corresponding author: Shisheng Guo (ssguo@uestc.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61871080, Grant 61771109, and Grant 61701088; in part by the Chang Jiang Scholar Program, and in part by the 111 Project under Grant B17008.

ABSTRACT Motivated by the intrinsic dynamics of physical motion as well as establishment of target motion model, this article addresses the problem of human motion recognition with ultra wide band (UWB) through-the-wall radar (TWR) in a novel view of range profile serialization. Specifically, we first convert the original radar echoes into range profiles. Then, an auto-encoder network (AEN) with three dense layers is adopted to reduce the dimension and extract the features of each range profile. After that, a gated recurrent unit (GRU) network with two hidden layers is employed to deal with the features of each time-range slice and output the recognition results at each slice in real time. Finally, experimental data with respect to four different behind-wall human motions is collected by self-developed UWB TWR to validate the effectiveness of the proposed model. The results show that the proposed model can validly recognize the human motion serialization and achieve 93% recognition accuracy within the initial 20% duration of the activities (the average durations are 4s, 5.5s, 3s and 4.5s), which is of great significance for real-time human motion recognition.

INDEX TERMS Human motion recognition, ultra wide band through-the-wall radar, auto-encoder network, gated recurrent unit.

I. INTRODUCTION

Through wall sensing techniques have developed for many years and become more attractive in a lot of fields, such as counter terrorism, law enforcement, and security check, etc [1]–[7]. Although these techniques have preferable capability to precisely locate, robustly track, and clearly imaging some hidden human targets, detailed information of the real-time activities of the human targets can still not be dug out clearly.

Recently, the topic of human motion recognition exploiting radar becomes a hot spot, and mass of machine learning approaches have been applied to solve the problem [8]–[10]. Support vector machine (SVM), as one of the most classical machine learning methods, have been applied to distinguish different motion types based on radar data. In [11], Y. Kim *et al.* used SVM to classify 7 kinds of motions, such as running, walking and walking without swinging arms. In [12], J. Bryan *et al.* used principal component analysis (PCA) for feature extraction and SVM for feature classification to realize the recognition of 7 different human motions, such as walking, running, turning and punching. By carefully designing the common features of those motions in a handcrafting manner, different motion data is mapped to several distinct point clusters in a higher-dimensional linear space, and these clusters are available to be split by multiple segmentation hyper planes. Therefore, those motion types could be recognized with a high accuracy [13]. However, the above method significantly relies on the quality of handcrafting feature designs. Unreasonable features would generate inappropriate high-dimensional linear spatial mapping, which leads to the unavailability of achieving the best splitting hyper planes via the optimization methods, and thereby significantly reduces the performance of human motion recognition [14].

To avoid such drawbacks, deep learning methods were adopted to extract the appropriate motion characters

The associate editor coordinating the review of this manuscript and approving it for publication was Hasan S. Mir.

automatically [15]–[21]. Convolutional neural network (CNN) is one of the most utilized deep learning structures to improve the classification accuracy for multiple human motion types [22]. Owing to its capability of learning, CNN is able to automatically extract deep motion features from a given action dataset which has a large quantity of data samples with type labels, and thereby achieves high recognition accuracy [23]. Nevertheless, CNN has a static structure and requires all the input data with exactly the same size along each dimension. It has the following 3 disadvantages: (a) It is not well compatible with the dynamic feature of physical motions, because the lasting time of human motions could be any length; (b) It makes it impossible to give the real-time recognition results in the process of action occurrence while applying CNN in a real-time scenario of detecting motion types. The reason is as a result of the completeness of the input data. (c) It involves repeated computation. Therefore, CNN is not suitable for radar based human real-time motion recognition.

Different from other deep learning architectures like multilayer perceptron (MLP) and CNN, recurrent neural network (RNN) is a significant partition among the deep learning techniques, which is good at dealing with sequential data [24]. It can utilize the sequentiality to learn and extract sequential features from a sequential input data stream. In [25], H. Li *et al.* presented a framework based on multilayer bi-LSTM network (bidirectional Long Short-Term Memory) for multimodal sensor fusion to sense and classify daily activities' patterns and high-risk events such as falls. In [26], J. Zhu *et al.* proposed a deep learning model composed of 1-D convolutional neural networks (1D-CNNs) and long short-term memory (LSTM). The results show that the proposed model can extract spatio-temporal characteristics of the radar data and achieve the best recognition accuracy with relatively low complexity compared to the existing 2D-CNN methods. In our previous work [27], we used the stacked RNN with LSTM units to extract sequential features for automatic motion classification and verified that a stacked RNN with two 36-cell LSTM layers successfully classifies six different types of human motions. Due to the mechanism of recurrent feedback, RNN is capable to remember the relationship between the historical inputs and the current input. Because of the time-dependent and sequential nature of human body and limb motion, RNN is a preferred choice for learning time-varying motion features that improves human body and limb motion classification accuracy.

For traditional line-of-sight scenario, micro-Doppler features are employed for radar based human motion recognition [11], [28]–[30]. In the presence of solid brick wall, low frequency (usually lower than 3 GHz) is usually adopted for ultra wide band (UWB) through-the-wall radar (TWR). As a result, the Doppler frequencies arose from the target motions that lay in a quite narrow band within low frequency range. This requires a very high frequency resolution to achieve enough resolution to distinguish the detailed Doppler frequencies (also called as micro-Doppler frequencies) induced by the motion of limbs. It leads to the invalidation of using time-frequency analysis method to represent the motions of behind-wall human target. Fortunately, the round-trip distance between the target and UWB TWR varies with the motion activating. As long as the UWB TWR is of enough distance resolution, detailed information of target's motion is able to be captured. In our previous work [18], an auto-encoder network (AEN) and a self organized mapping (SOM) network were used to extract the features of human motion from range profiles. We have proved that using AEN network and SOM network to extract the feature of range profile can implement human behavior information representation.

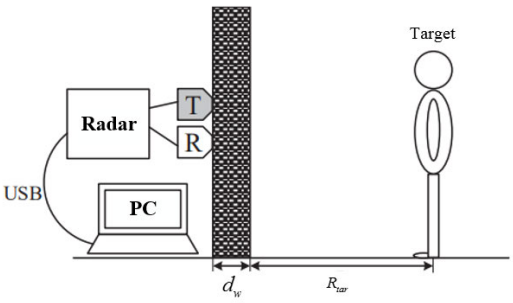

FIGURE 1. The scenario of hidden human motion detection with UWB TWR.

As a continuation, in this article, we address the problem of human motion recognition using UWB TWR in a serialization manner. Specifically, we first convert the original radar echoes into range profiles. Then, an AEN with three dense layers is adopted to reduce the dimension and extract the features of each range profile. After that, a gated recurrent unit (GRU) network with two hidden layers is employed to deal with the features of each time slice and output the recognition results at each slice in real time. Finally, experimental results with respect to four behind-wall human motions validate the effectiveness of the proposed model. The proposed method has the following two highlights:

1) The human motion recognition results can be provided at each temporal frame, which meets the application requirement of UWB TWR in real time;
2) Our approach is able to deal with the recognition problem for temporal-length-varying human motions.

The rest of the paper is organized as follows. In Section II, the signal model of UWB TWR is established. Section III represents the mechanism of human motion serialization recognition based on GRU network. Section IV elaborates the procedures of dataset collecting, implementation processing, and experimental evaluating. Section V ultimately concludes this article.

II. SIGNAL MODEL

Consider a scenario that a UWB TWR is located against to a wall to detect a single human who is performing some motions on the other side, as shown in FIGURE 1.

The transmitted signal of the UWB TWR in this article is stepped frequency signal [31]. The stepped-frequency signal can be expressed as

$$s(t) = \sum_{k=0}^{K-1} \text{rect}\left(\frac{t - T/2 - kT}{T}\right) e^{j2\pi(f_0 + k\Delta f)t}, \quad (1)$$

where T is the lasting time of each carrier frequency point, K is the total number of sampling points, KT is the total time needed to transmit a complete frequency step signal of K sampling points in time-sharing (It is generally called "the duration of slow time period" or "one slow time period" for short). f_0 represents the starting carrier frequency, Δf denotes frequency step, and the function rect$(\cdot)$ is defined as

$$\text{rect}(t) = \begin{cases} 1, & |t| \leqslant \frac{1}{2}, \\ 0, & \text{otherwise.} \end{cases} \quad (2)$$

The echos received by the receiver can be expressed as

$$s_{\text{r}}(t) = \sum_{p=1}^{P} A^{(p)} s(t - \tau^{(p)}) + s_{\text{w}}(t) + s_{\text{noise}}(t), \quad (3)$$

where the summation item is the echo reflected from the human target, P denotes the total number of the scattering points on the target, $A^{(p)}$ represents the amplitude of the echo corresponding to point p, $\tau^{(p)}$ is the round-trip temporal delay of point p, which is expressed as

$$\tau^{(p)} = \frac{R_{\text{air}}^{(p)} + \sqrt{\varepsilon_{\text{w}}} R_{\text{wall}}^{(p)}}{c}, \quad (4)$$

c is the velocity of electromagnetic wave spreading in the air, ε_{w} denotes the permittivity of the wall, signal $s_{\text{w}}(t)$ signifies the strong clutter reflected by the wall, and $s_{\text{noise}}(t)$ expresses the other clutters as well as the environmental noises.

After orthogonal demodulation processing, we sample the result in the temporal rate of T and achieve a discrete vector, expressed as

$$\boldsymbol{s}_{\text{r}} = \boldsymbol{s}_{\text{tar}} + \boldsymbol{s}_{\text{w}} + \boldsymbol{s}_{\text{noise}}, \quad (5)$$

where

$$\boldsymbol{s}_{\text{tar}} = \sum_{p=1}^{P} a^{(p)} e^{-j2\pi f_0 \tau^{(p)}} \boldsymbol{s}_0, \quad (6)$$

$$\boldsymbol{s}_0 = \left[1, \quad e^{-j2\pi\Delta f \tau^{(p)}}, \quad \cdots, \quad e^{-j2\pi(K-1)\Delta f \tau^{(p)}}\right]^{\dagger}, \quad (7)$$

$\boldsymbol{s}_{\text{w}}$ and $\boldsymbol{s}_{\text{noise}}$ are the vectors corresponding to signal $s_{\text{w}}(t)$ and $s_{\text{noise}}(t)$ respectively, $\dagger$ denotes the operation of transposition.

Afterwards, by performing inverse fast Fourier transform (IFFT) with Q points, we get the range profile of a single K-sampling-points period (also called *slow time*), expressed as

$$\boldsymbol{s}_{\text{rp}} = \left[\boldsymbol{s}_{\text{rp}}(q)\right]^{\text{T}} + IFFT\{\boldsymbol{s}_{\text{w}}\} + IFFT\{\boldsymbol{s}_{\text{noise}}\}, \quad (8)$$

where $q = 0, 1, \cdots, Q-1$ is the index of range bins, and $\boldsymbol{s}_{\text{rp}}(q)$ is the temporal form of the stepped frequency signal after pulse compression, expressed as

$$\left|\boldsymbol{s}_{\text{rp}}(q)\right| = \left| \sum_{p=1}^{P} a^{(p)} e^{j\phi(q,\tau^{(p)})} \frac{\sin \pi \frac{K}{Q}(q - Q\Delta f \tau^{(p)})}{\sin \pi \frac{1}{Q}(q - Q\Delta f \tau^{(p)})} \right|. \quad (9)$$

With Eq. (9), the following remarks can be inferred:

1) The spatial states of the human motions have been implicated in $\boldsymbol{s}_{rp}(q)$. Each slow time period KT records the short-term posture information of the hidden human target at that time.
2) With the continuous accumulation of slow time periods, dynamic information of target movement can be obtained. As the UWB TWR continuously emits the stepped frequency signal represented by (1) in a slow time period KT and receives the behavioral echo signal of the hidden human target in the scene, a two-dimensional data block that contains the human body motion information is finally formed. In the process of human motion, the range profile received by the radar is arranged by a one-dimensional range profile corresponding to a plurality of slow time periods in the direction of time lapse.

Therefore, the range profile has the capability to represent hidden human motions and each time slice represents a one-dimensional range profile.

Additionally, since the strong clutters reflected by the wall significantly interferes with the energy reflected from the hidden target, moving target indicator (MTI) technique is utilized to inhibit the static background clutters [32].

III. HUMAN MOTION SERIALIZATION RECOGNITION

A. MOTIVATION

Human brain is of highly complicated, smart, and collaborated to recognize a large number of motion types in short time. In practice, the recognition procedure can be roughly described in the following.

Suppose a scenario that an observer O watches a target S who performs a type of motion M. Then the following judgment stages would be occurred:

1) The initial stage: When S prepares to perform M, O does not know what motion S is going to perform at such instance. But O still has the conscious that S is going to perform some motion, which is different from the state of no motion performing.
2) The transient stage: When S starts to perform part of M, O gradually recognizes which type of motions that S is doing, and quickly rules out the possibility of other motion types except for M.
3) The steady stage: When S continues to perform M until cease, O provides M, the result of the motion recognition, with a high confidence by combining the historical motion information that O has observed.

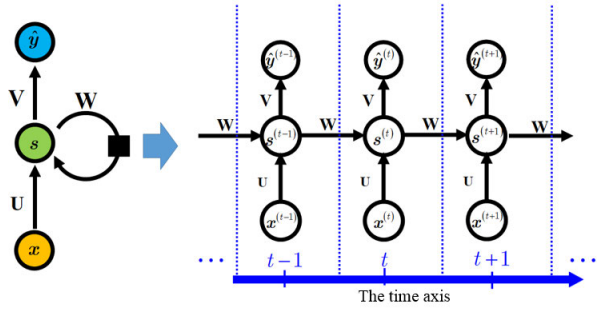

FIGURE 2. The structure of vanilla RNN.

4) The final stage: When S stops performing M, O gives the assertion within a short time that M has been stopped and S is doing nothing.

In the above description, we can find that the entire recognition procedure conducted by O's brain is a temporal serialization process in essence. The process can be regarded as a transfer that multiple latent motion features or states vary with time flows. During transferring, the brain is of strong capability to conduct the visual motion features abstraction and the motion type analysis. The recognition result outputs at each temporal instance.

Therefore, motivated by the aforementioned behavior of brain, we expect that UWB radar could carry out similar capability to extract motion features, judge motion types and provide recognition results at each slow time instance.

B. RECURRENT NEURAL NETWORK

RNN has three common structure forms: vanilla RNN, long short-term memory (LSTM), and GRU [33], [34]. All of them are deep in both spatial layers and temporal flows.

Vanilla RNN is the simplest structure of RNN, which is conducted from the architecture of MLP with three layers, as shown in Figure 2. The only difference between MLP and vanilla RNN is the loop link in the hidden layer. Therefore, the work flow of vanilla RNN can be expressed as

$$\boldsymbol{s}^{(t)} = \boldsymbol{g}\left(\boldsymbol{U}\boldsymbol{x}^{(t)} + \boldsymbol{W}\boldsymbol{s}^{(t-1)} + \boldsymbol{b}_s^{(t)}\right), \tag{10}$$

where $\boldsymbol{x}^{(t)}$ is the input of the network, $\boldsymbol{s}^{(t)}$ represents the output of the hidden layer at time t, $\boldsymbol{U}$ denotes the link weights between the input layer and the hidden layer, $\boldsymbol{W}$ signifies the feedback loop weights in the hidden layer, $\boldsymbol{b}_s^{(t)}$ is the bias item in the hidden layer, and $\boldsymbol{g}$ is the activation function. According to the flow shown in Figure 2, the final output $\boldsymbol{y}$ is able to be expressed as

$$\begin{aligned}\boldsymbol{y}^{(t)} &= \boldsymbol{g}\left(\boldsymbol{V}\boldsymbol{s}^{(t)} + \boldsymbol{b}_y^{(t)}\right) \\ &= \boldsymbol{g}\left(\boldsymbol{V}\left(\boldsymbol{U}\boldsymbol{x}^{(t)} + \boldsymbol{W}\boldsymbol{s}^{(t-1)} + \boldsymbol{b}_s^{(t)}\right) + \boldsymbol{b}_y^{(t)}\right),\end{aligned} \tag{11}$$

where $\boldsymbol{g}$ is also an activate function.

Theoretically, vanilla RNN could deal with sequences at any length and remember all the historical information that has been passed through. However, it has been proved to be impossible for vanilla RNN, since it faces the severe problem of gradient explosion or vanishing during training, which leads to the short-term memory of vanilla RNN.

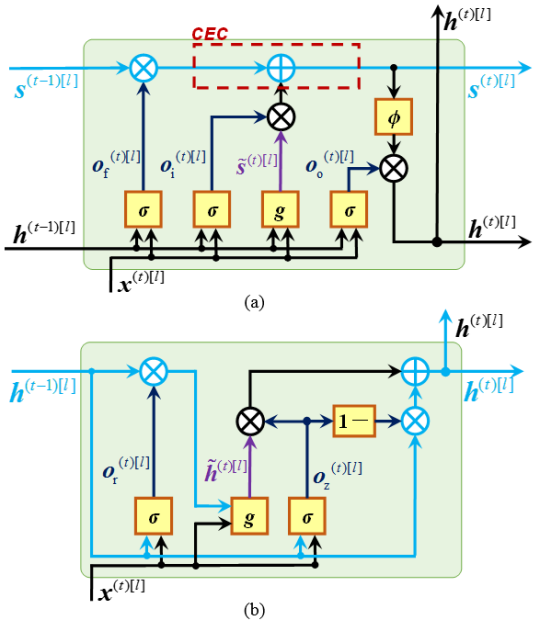

FIGURE 3. The structure of LSTM and GRU: (a) LSTM structure; (b) GRU structure.

To address the above problem, gated mechanism was proposed and applied in LSTM as well as GRU. Figure 3 depicts the structures of LSTM and GRU. Both of these two networks induce multiple extra "gates" to control the quantity of information flowing within the hidden layer. LSTM contains three gates: the forget gate, the input gate, and the output gate, whereas GRU includes only two gates: the reset gate and the update gate. From the aspect of spatial complexity, LSTM embrace more parameters than GRU. Therefore, GRU has less computation costs than LSTM. As for the aspect of performance, LSTM and GRU almost have the same effects on multiple recognition tasks. In consequence, we just describe the work flow of GRU as below.

$$\begin{aligned}\boldsymbol{o}_{\mathrm{r}}^{(t)[l]} &= \boldsymbol{\sigma}\left(\boldsymbol{W}_{\mathrm{r}_h}\boldsymbol{h}^{(t-1)[l]} + \boldsymbol{W}_{\mathrm{r}_x}\boldsymbol{x}^{(t)[l]} + \boldsymbol{b}_{\mathrm{r}}^{[l]}\right), \\ \widetilde{\boldsymbol{h}}^{(t)[l]} &= \boldsymbol{g}\left[\boldsymbol{W}_{\tilde{h}_h}\left(\boldsymbol{o}_{\mathrm{r}}^{(t)[l]} \odot \boldsymbol{h}^{(t-1)[l]}\right) + \boldsymbol{W}_{\tilde{h}_x}\boldsymbol{x}^{(t)[l]} + \boldsymbol{b}_{\tilde{h}}^{[l]}\right], \\ \boldsymbol{o}_{\mathrm{z}}^{(t)[l]} &= \boldsymbol{\sigma}\left(\boldsymbol{W}_{\mathrm{z}_h}\boldsymbol{h}^{(t-1)[l]} + \boldsymbol{W}_{\mathrm{z}_x}\boldsymbol{x}^{(t)[l]} + \boldsymbol{b}_{\mathrm{z}}^{[l]}\right), \\ \boldsymbol{h}^{(t)[l]} &= \boldsymbol{o}_{\mathrm{z}}^{(t)[l]} \odot \boldsymbol{h}^{(t-1)[l]} + (1 - \boldsymbol{o}_{\mathrm{z}}^{(t)[l]}) \odot \widetilde{\boldsymbol{h}}^{(t)[l]},\end{aligned} \tag{12}$$

where l represents the l-th hidden layer, $\boldsymbol{x}^{(t)[l]}$ is the input of the l-th hidden layer, $\boldsymbol{\sigma}$ and $\boldsymbol{g}$ are activate functions, $\boldsymbol{W}$ and $\boldsymbol{b}$ represents weights and bias respectively, $\widetilde{\boldsymbol{h}}^{(t)[l]}$ denotes the input candidates of the GRU, the letter r and z indicate the reset gate as well as the update gate, and $\odot$ signifies the Hadamard product.

C. THE PROPOSED REAL-TIME HUMAN MOTION SERIALIZATION RECOGNITION MODEL

According to the comparisons of different types of RNN mentioned above and inspired by the human brain behavior recognition process, we propose a real-time human motion recognition model based on GRU.

Before using the GRU network for real-time recognition, we need to use a network to extract the hidden information in the range profile. In this article, AEN is utilized to extract the common motion features. The reason why we choose the AEN is that the structure is brief and can implement unsupervised self-learning of dimensionality reduction features of data without additional manual feature selection [35]. The architecture of AEN is shown in Figure 4.

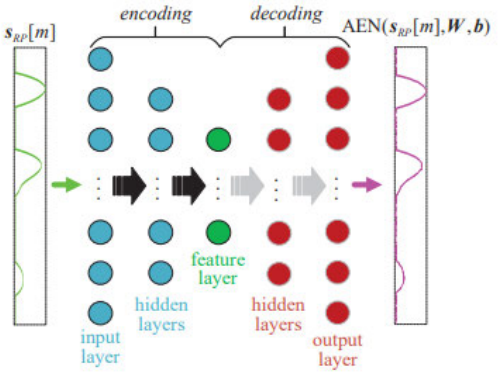

FIGURE 4. The structure of AEN.

Two sub-networks are included in AEN: one of them is the encoder, and the other is the decoder. The encoder aims to compress and map the input data into a smaller feature space and the decoder tries to recover the input data from the mapped feature space. Hence, AEN can be treated as an equivalent representer transforming the data space into a compact feature space, holding the dominated information implied in the motion data. When the output of AEN is quite similar to the input data within an error tolerance during the training phase, the feature layer shared by the encoder and the decoder can represent the features of the input data well. Hence, we only maintain the encoder to transform the range profile into the deep features during usage.

Accordingly, a novel serialized recognition model for hidden human target is proposed, and the structure of the model is shown in Figure 5. By connecting the encoder part of the AEN and the GRU network, the procedure of hidden human motion recognition in serialization manner can be achieved. Since the outputs of GRU network are also some deep features, we add an extra layer with softmax function to translate the features into the possibility of the motion type that the input data belongs to. The softmax function can be expressed as

$$f(x_i) = \frac{e^{x_i}}{\sum_i e^{x_i}}. \tag{13}$$

After the above operations, the proposed model can output recognition results in real time.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

A. DATASET CONSTRUCTION

1) SAMPLES ACQUISITION

To evaluate the recognition performance of the proposed model, we collect a human motion dataset with a self-developed UWB TWR. The block diagram of the radar system is shown in Figure 6. The acquisition scenario is shown in Figure 7. Therein, the radar is placed against to one side of a brick wall with a thickness of 0.70 m. The hight of the radar is 0.90 m away from the ground. The radar is equipped with a single transmitting antenna as well as a single receiving antenna, and the distance between them is 7.5 cm. Stepped frequency starting from 1.6 GHz and ending with 2.2 GHz is employed as the emitting signal. The frequency step is set to 2 MHz, resulting that the total frequency points is 301. Additionally, the lasting time of each carrier frequency is configured as 1×10^{-4} s to achieve a slow time period of 0.03s. The range resolution is 0.25 m and the polarization mode is horizontal transmission and horizontal reception (HH).

During the experimental data acquisition, 4 volunteers were asked to perform the following 4 motion types with the same distance of 1.5 m apart from the other side of the wall along the direction of light-of-sight for the radar: i) boxing; ii) walking on a fixed point without arms swinging; iii) picking; and iv) arm raising and lowering. The reasons why we choose these four motions are that they belong to daily motions and the changes of radial distances are obvious. The 4 volunteers are all males. Their ages ranged from 23 to 28 years and weights ranged from 65 to 90 kg, with height from 1.65 to 1.80 m. For each volunteer, every motion type was asked to be performed 54 times repeatedly. Therefore, the dataset contains 864 samples in range profile format. In this article, the training set contains 544 samples, the validation set contains 160 samples, and the testing set contains 160 samples. Each sample contains 404 slow time periods (time slices). Therefore, the number of time slices used for training is 544×404, and the number of time slices used for validating and testing is 160×404. In addition, since the range ambiguity of radar is 75 m, which is much less than the distance of the volunteers who performs motions, we truncated all the samples and only kept 512 range bins, i.e. 4.7 m, in round trip. So, the size of each time slice is 1×512. For convenience, we use the words ''motion I'', ''motion II'', ''motion III'', and ''motion IV'' to refer to the corresponding four motion types mentioned above.

The examples of motion samples are shown in Figure 8, and we are able to find that abundant interferences as well as the noise energy filled in the range bins other than the zones appeared strong energy resulting from the volunteers motions. According to the statistic learning theory [36], such noises would weaken the recognition ability of the motion

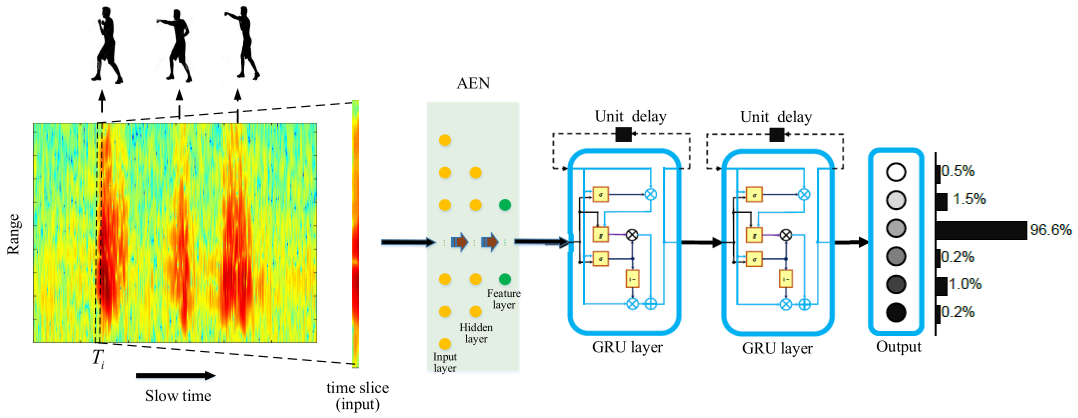

FIGURE 5. The structure of the proposed model.

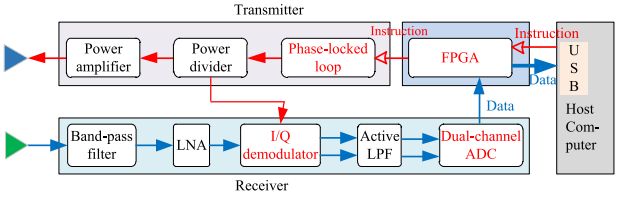

FIGURE 6. The block diagram of the radar system.

FIGURE 7. The experimental scenario for dataset acquisition with a self-developed TWR.

classifier without clutter suppression during the first-time training since the statistical characteristics of noises are the dominants. Therefore, clutter suppression is of importance in the next step of dataset construction.

2) CLUTTER SUPPRESSION

In order to reduce the influence of clutter on the recognition results, we use the maximum entropy threshold method proposed by [37] to determine the threshold E. The reason why the maximum entropy threshold method is used is that the signal-to-noise ratio (SNR) of each sample reaches a high level. Different from the binarization of pixels using the maximum entropy threshold method in image processing,

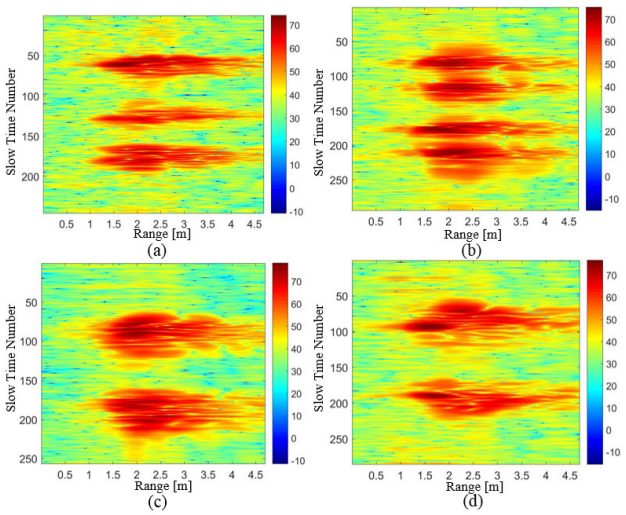

FIGURE 8. The motion samples: (a)motion I; (b)motion II; (c)motion III; (d)motion IV.

after obtaining the threshold E, we use formula (14) to process the sample data. Eq. (14) can be expressed as

$$v_{th} = \frac{E - v_{min}}{v_{max} - v_{min}}, \tag{14}$$

where v_{max} and v_{min} represent the maximum and minimum values (converted to decibels) in the training set, respectively.

Then we compare the value of every behavior distance profile in the whole database with v_{th}. If the value is bigger than v_{th}, it will be retained. If the value is smaller than v_{th}, it will be replaced by the threshold value v_{th}. After that, the sample data after threshold processing is standardized again, the maximum value is set to be 1, the minimum value is set to be 0. Finally, the human motion recognition database of UWB TWR after clutter suppression is obtained.

Figure 9 shows an example of sample data after clutter suppression. Compared with Figure 8, the sample data in Figure 9 can show the characteristics of hidden human motion more clearly, which lays a foundation for the training and testing of

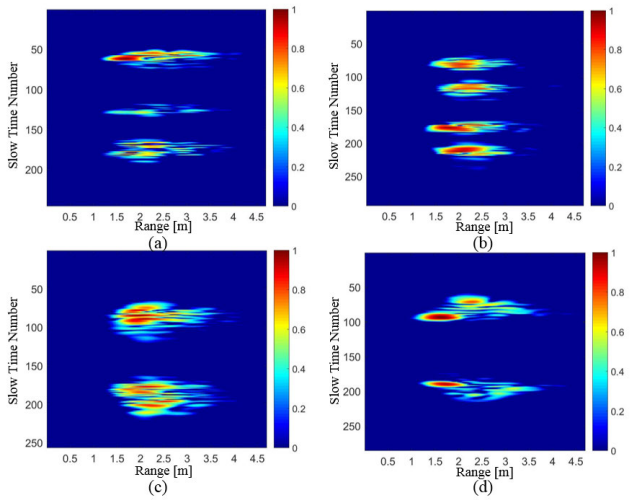

FIGURE 9. The motion samples after clutter suppression: (a)motion I; (b)motion II; (c)motion III; (d)motion IV.

sequential recognition model of hidden human motion based on deep learning.

B. IMPLEMENTATION

Raw radar echo is transformed into range profile at first. This operation consists of *IFFT*, MTI and clutter suppression. After obtaining the preprocessed range profile, the AEN operation is processed to extract local features. Then, a two-layer GRU network encodes the temporal patterns. The output of GRU flows into a fully-connected layer with a softmax activation function to classify the identification of the range-profile at each time step.

Suppose the dataset composed by the range profile signature vectors extracted from the trained AEN contains M examples. The training procedure is summarized as Algorithm 1.

In Algorithm 1, $L(i)$ is called as the loss function. TWR human behavior recognition is actually a classification task. We thus choose cross entropy as the loss function in this article.

In this article, all models are implemented on a server equipped with 64G memory and an NVIDIA GeForce GTX1080 Ti 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 20. Batch normalization technique is also employed in the implementation. The learning rate is set to 0.001.

C. EVALUATION

1) CONVERGENCE CHARACTERISTICS OF THE MODEL

In order to test the convergence characteristics of the model, we carried out experiments on the training set and the verification set. The convergence curve of the network is shown in Figure 10. It can be seen from Figure 10 that the proposed model has good convergence characteristics in training set and verification set. In the first 50 iterations of the model, the loss function decreased rapidly. When the number of iterations reached 300, the network basically converged, and the loss was less than 1×10^{-5}. It can be concluded that the proposed model is suitable for human motion recognition with UWB TWR.

Algorithm 1 Training a GRU Network

Input:
The maximum iterating steps I; The learning rate decay η; Training dataset X_M; Threshold ϕ;
Output:
The GRU net's weights W and bias b;

Initialize $W(i)$, $b(i)$ randomly, where $i = 0$;
Initialize the learning rate $\alpha(i) = 1, i = 0$;
for $i = 1, 2, \cdots, I$ **do**
  Randomly take out an example $X(i)$ from X_M;
  Randomly slice $X(i)$;
  Calculate the loss function $L(i)$;
  Update the weights $W(i) = W(i-1) - \alpha(i) W(i)$;
  Update the bias $b(i) = b(i-1) - \alpha(i) b(i)$;
  Update the learning rate $\alpha(i) = \alpha(0) e^{-\eta(i-1)}$;
  if $L(i) \leq \phi \cup i \geq I$ **then**
    Stop.
  else
    $i = i + 1$.
    Goto step (3).
  end if
end for
Output W, b.

2) REAL TIME RECOGNITION ANALYSIS OF THE MODEL

In order to test the performance of the proposed architecture for processing human motion sequences, we evaluate the model on the verification set. Figure 11 shows the human motion recognition results and the raw range profiles for the four motion types. The upper side of (a), (b), (c), (d) are the range profiles of motion samples. The lower side of (a), (b), (c), (d) are the real time recognition results. Where "1" means "motion I", "2" means "motion II", "3" means "motion III", "4" means "motion IV", and "5" means no action has taken place. T_0 represents the waiting time of an action, T_1 represents the duration of an action, T_2 and T_3 represent the start and end delay of network that correctly identify the action types.

Take Figure 11(a) as an example for detailed analysis. During T_0, there is no information available, so the model determines action 5 stably, which means no action occurs. During T_2, although the model knows that some action is happening, there is too little information available, so there will be a start delay in recognition. As the action continues, the action information accumulates continuously. After T_2, the model has been able to judge the current action according to historical information and current time information.

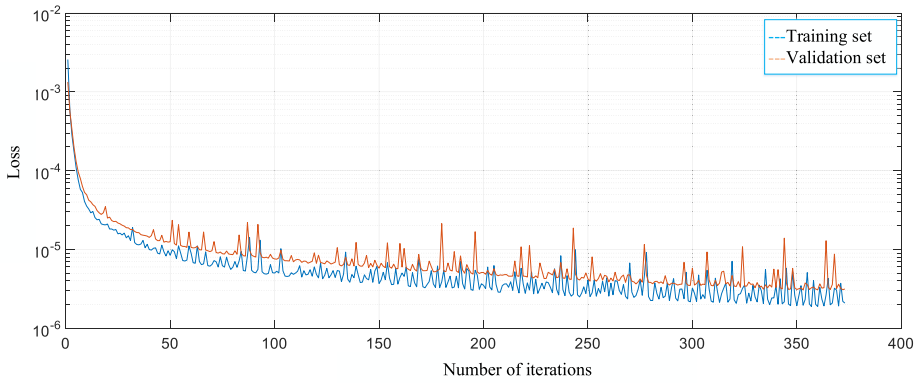

FIGURE 10. Convergence curve of the proposed model.

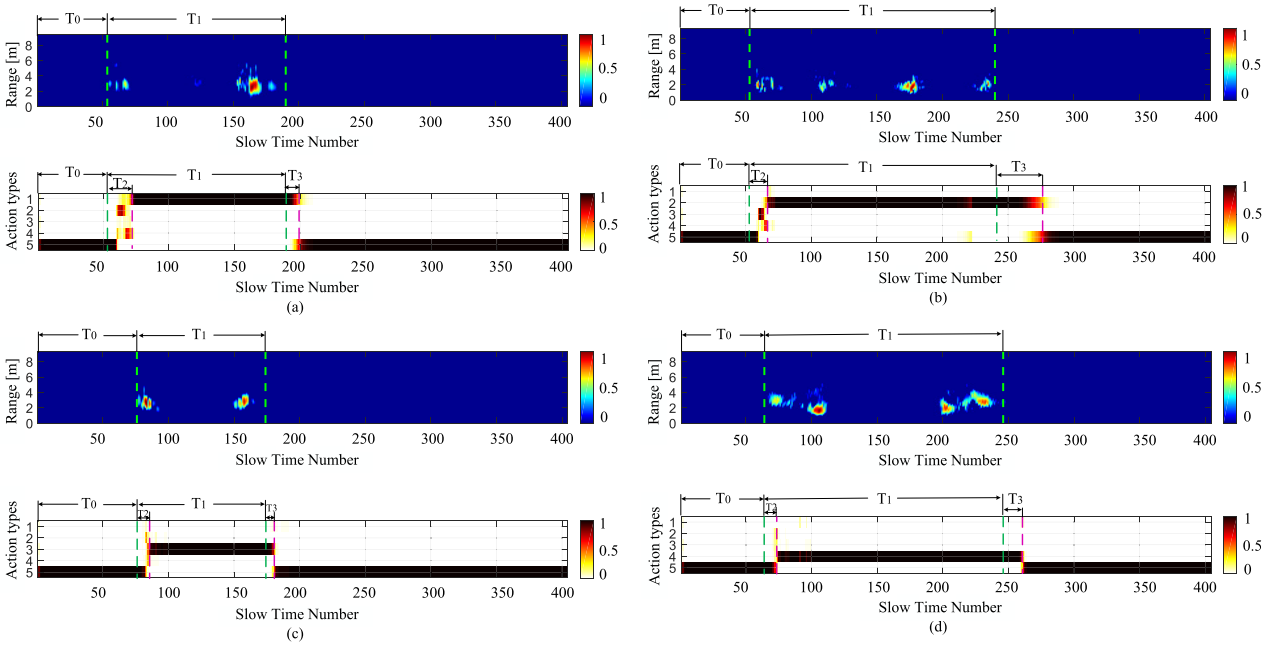

FIGURE 11. Real time recognition results of the motion samples: (a)motion I; (b)motion II; (c)motion III; (d)motion IV.

After T_1, the action ends, but the model still judges that the action is taking place by using the historical information, so there will be an end delay in recognition. During T_1, although there are blank areas caused by the action interval, the model can still judge the action type according to the association information. Note that other graphs in Figure 11 have the same characteristics above.

In the process of motions, they are not maintaining continuous movements, during which there will be some short pause and it's the reason for the blank area. In order to correctly identify the blank area in the process of these motions, we use the GRU network with memory ability. Although the current time is blank, the model can judge the output of the current time according to the historical information.

From the analysis aforementioned, we know that the recognition process of the proposed model is similar to the human brain judgment action type described in part A of Section III. In the initial stage of action, it is impossible to determine which action it belongs to. Only when the action information is accumulated to a certain amount, can the action type be accurately determined. The start and end delay (T_2 and T_3) are unavoidable. The goal is to minimize the delay.

In order to observe the change rule of the recognition ability of the model more intuitively, we regard the effective human motion time T_1 as 1, and take 1% as the step length to test the relationship between the accuracy of model recognition, time delay and the proportion of action occurrence. We take the initial 15%, 18% and 20% of duration of the activities as examples to explain the relationship between them in detail. The test results are shown in TABLE 1. Within the initial 15% of the activities, the recognition rate is more than 80%. With the accumulation of movement information,

TABLE 1. The relationship between accuracy, time delay and action proportion.

Activities	Accuracy	Time Delay	Action Proportion	Accuracy	Time Delay	Action Proportion	Accuracy	Time Delay	Action Proportion
motion I	81.46%	514.39 ms	15%	91.34%	651.00 ms	18%	92.86%	712.68 ms	20%
motion II	82.99%	724.50 ms	15%	91.57%	801.04 ms	18%	93.15%	1330.02 ms	20%
motion III	80.41%	371.29 ms	15%	90.26%	429.05 ms	18%	93.43%	572.18 ms	20%
motion IV	81.28%	507.18 ms	15%	91.07%	608.93 ms	18%	92.56%	681.13 ms	20%
Average	81.54%	529.34 ms	15%	91.06%	622.51 ms	18%	93.00%	824.01 ms	20%

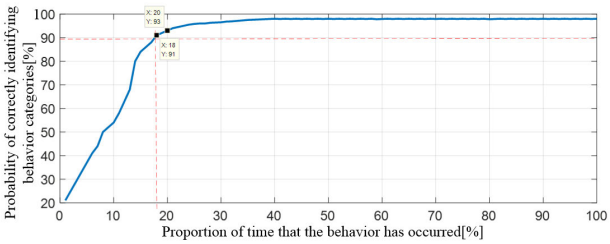

FIGURE 12. The relationship between the correct rate of behavior recognition and the proportion of behavior occurrence.

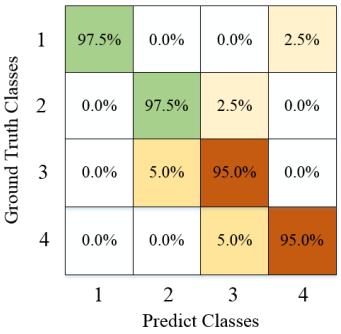

FIGURE 13. Confusion matrix.

the model can achieve 91.06% recognition accuracy within the initial 18% of duration of the activities. Within the initial 20%, the recognition rate of the proposed model reaches 93.00%. After calculation, the maximum delay time for the experimental four behaviors is 1.33s. Such delays can be tolerated in real scenarios.

In addition, to express the relationship between behavior proportion and model recognition accuracy, we take the average of recognition accuracy of all actions in the same proportion, and get the following result curve. The experimental results are shown in Figure 12.

As can be seen from Figure 12, in the initial stage, the recognition accuracy of the model is relatively low. With the continuous accumulation of action information, the recognition rate of the model increases rapidly. After the behavior occurs 20%, the recognition rate of the model changes slowly. In this article, only when the recognition accuracy is higher than 93%, will we determine the type of action that is taking place.

Figure 13 depicts the confusion matrix for the proposed model. From the confusion matrix, we can observe that the micro-F1 score of the model is 0.9543. This shows the effectiveness of our model.

The following discussions can be drawn from the above analysis.

1) As we described in Part A of Section III, the human brain has a delay in judging the beginning and end of an action. Similarly, the model proposed in this article also has a certain delay in the beginning and end. It is consistent with the real scene. From the experimental results, we find that the start and end delay time of an action is related to the complexity and duration of the action. We also find that the proposed model has a good follow-up to the behavior process. Therefore, the ability of the proposed model to recognize human behavior serialization is verified.
2) As we can see, the start time, duration and end time of the four actions are different, which validates the ability of the proposed model to handle non-equal-length human motions.
3) Due to the gated memory mechanism in the GRU unit, the proposed model can memorize historical information an output behadvior recognition results based on the input of the current moment immediately. Therefore, the proposed model has a good real-time performance and can meet the needs of real-time human motion recognition of UWB TWR.

V. CONCLUSION

In this article, we have addressed the problem of human motion recognition with UWB TWR in a novel view of serialization. Firstly, the original radar echoes are converted into range profiles. Then, an AEN with three dense layers is utilized to reduce the dimension and extract the features of each range profile. After that, a GRU network with two hidden layers is employed to deal with the features of each time slice and output the recognition results at each slice in real time. Finally, experimental data with respect to four different behind-wall human motions is collected by self-developed UWB TWR to validate the effectiveness of the proposed model. The experimental results have shown that the proposed model can validly recognize the human motion serializations and achieve 93% recognition accuracy within the initial 20% duration of the activities (the average durations are 4s, 5.5s, 3s, 4.5s). The proposed model is quite different from other neural network models that have to wait until the end of the action to give the action type. Our proposed

model is real-time, which is of great significance for real-time human motion recognition.

ACKNOWLEDGMENT

The authors would like to thank the colleagues in the University of Electronic Science and Technology of China for their assistance in TWR based human motion data collection.

REFERENCES

[1] J. L. Kernerc and O. Romain, "Performances of multitones for ultra-wideband software-defined radar," *IEEE Access*, vol. 5, pp. 6570–6588, 2017.
[2] M. G. Amin, Y. D. Zhang, F. Ahmad, and K. C. D. Ho, "Radar signal processing for elderly fall detection: The future for in-home monitoring," *IEEE Signal Process. Mag.*, vol. 33, no. 2, pp. 71–80, Mar. 2016.
[3] A. Kumar, Z. Li, Q. Liang, B. Zhang, and X. Wu, "Experimental study of through-wall human detection using ultra wideband radar sensors," *Measurement*, vol. 47, pp. 869–879, Jan. 2014.
[4] O. P. Popoola and K. Wang, "Video-based abnormal human behavior recognition—A review," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 42, no. 6, pp. 865–878, Nov. 2012.
[5] F. Qi, H. Lv, F. Liang, Z. Li, X. Yu, and J. Wang, "MHHT-based method for analysis of micro-Doppler signatures for human finer-grained activity using through-wall SFCW radar," *Remote Sens.*, vol. 9, no. 3, p. 260, Mar. 2017.
[6] S. Guo, G. Cui, L. Kong, and X. Yang, "An imaging dictionary based multipath suppression algorithm for through-wall radar imaging," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 54, no. 1, pp. 269–283, Feb. 2018.
[7] H. Li, G. Cui, L. Kong, G. Chen, M. Wang, and S. Guo, "Robust human targets tracking for MIMO through-wall radar via multi-algorithm fusion," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 4, pp. 1154–1164, Apr. 2019.
[8] Z. Sun, J. Wang, J. Sun, and P. Lei, "Parameter estimation method of walking human based on radar micro-Doppler," in *Proc. IEEE Radar Conf. (RadarConf)*, Seattle, WA, USA, May 2017, pp. 0567–0570.
[9] B. Jokanovic, M. Amin, and F. Ahmad, "Radar fall motion detection using deep learning," in *Proc. IEEE Radar Conf. (RadarConf)*, May 2016, pp. 1–6.
[10] B. Cagliyan and S. Z. Gurbuz, "Micro-Doppler-based human activity classification using the mote-scale BumbleBee radar," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 10, pp. 2135–2139, Oct. 2015.
[11] 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.
[12] J. Bryan and Y. Kim, "Classification of human activities on UWB radar using a support vector machine," in *Proc. IEEE Antennas Propag. Soc. Int. Symp.*, Toronto, ON, Canada, Jul. 2010, pp. 1–4.
[13] F. Fioranelli, M. Ritchie, S. Z. Gurbuz, and H. Griffiths, "Feature diversity for optimized human micro-Doppler classification using multistatic radar," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 53, no. 2, pp. 640–654, Apr. 2017.
[14] Y. Kim and Y. Li, "Human activity classification with transmission and reflection coefficients of on-body antennas through deep convolutional neural networks," *IEEE Trans. Antennas Propag.*, vol. 65, no. 5, pp. 2764–2768, May 2017.
[15] 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, May 2019.
[16] A. Kılıç, İ. Babaoğlu, A. Babalık, and A. Arslan, "Through-wall radar classification of human posture using convolutional neural networks," *Int. J. Antennas Propag.*, vol. 2019, pp. 1–10, Mar. 2019.
[17] R. Zhao, H. Ali, and P. van der Smagt, "Two-stream RNN/CNN for action recognition in 3D videos," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Vancouver, BC, Canada, Sep. 2017, pp. 4260–4267.
[18] M. Wang, G. Cui, H. Huang, X. Gao, P. Chen, H. Li, H. Yang, and L. Kong, "Through-wall human motion representation via autoencoder-self organized mapping network," in *Proc. IEEE Radar Conf. (RadarConf)*, Boston, MA, USA, Apr. 2019, pp. 1–6.
[19] M. A. Kiasari, S. Y. Na, and J. Y. Kim, "Classification of human postures using ultra-wide band radar based on neural networks," in *Proc. Int. Conf. IT Converg. Secur. (ICITCS)*, Beijing, China, Oct. 2014, pp. 1–4.
[20] H. T. Le, S. L. Phung, A. Bouzerdoum, and F. H. C. Tivive, "Human motion classification with micro-Doppler radar and Bayesian-optimized convolutional neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Calgary, AB, Canada, Apr. 2018, pp. 2961–2965.
[21] P. Wang, H. Liu, L. Wang, and R. X. Gao, "Deep learning-based human motion recognition for predictive context-aware human-robot collaboration," *CIRP Ann.*, vol. 67, no. 1, pp. 17–20, 2018.
[22] R. Zhang and S. Cao, "Real-time human motion behavior detection via CNN using mmWave radar," *IEEE Sensors Lett.*, vol. 3, no. 2, pp. 1–4, Feb. 2019.
[23] W. Yin, X. Yang, L. Zhang, and E. Oki, "ECG monitoring system integrated with IR-UWB radar based on CNN," *IEEE Access*, vol. 4, pp. 6344–6351, 2016.
[24] H. K. Kwan and J. Yan, "Recurrent neural network design for temporal sequence learning," in *Proc. 43rd IEEE Midwest Symp. Circuits Syst.*, Lansing, MI, USA, vol. 2, Aug. 2000, pp. 832–835.
[25] H. Li, A. Shrestha, H. Heidari, J. Le Kernec, and F. Fioranelli, "Bi-LSTM network for multimodal continuous human activity recognition and fall detection," *IEEE Sensors J.*, vol. 20, no. 3, pp. 1191–1201, Feb. 2020.
[26] 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.
[27] M. Wang, Y. D. Zhang, and G. Cui, "Human motion recognition exploiting radar with stacked recurrent neural network," *Digit. Signal Process.*, vol. 87, pp. 125–131, Apr. 2019.
[28] 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.
[29] C. Ding, L. Zhang, C. Gu, L. Bai, Z. Liao, H. Hong, Y. Li, and X. Zhu, "Non-contact human motion recognition based on UWB radar," *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 8, no. 2, pp. 306–315, Jun. 2018.
[30] H. Du, T. Jin, Y. He, Y. Song, and Y. Dai, "Segmented convolutional gated recurrent neural networks for human activity recognition in ultra-wideband radar," *Neurocomputing*, vol. 396, pp. 451–464, Jul. 2020.
[31] G. Cui, L. Kong, and X. Yang, "Reconstruction filter design for stepped-frequency continuous wave," *IEEE Trans. Signal Process.*, vol. 60, no. 8, pp. 4421–4426, Aug. 2012.
[32] P. Goy, F. Vincent, and J.-Y. Tourneret, "Clutter rejection for MTI radar using a single antenna and a long integration time," in *Proc. 4th IEEE Int. Workshop Comput. Adv. Multi-Sensor Adapt. Process. (CAMSAP)*, San Juan, Puerto Rico, Dec. 2011, pp. 389–392.
[33] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.
[34] D. Zhang, G. Lindholm, and H. Ratnaweera, "Use long short-term memory to enhance Internet of Things for combined sewer overflow monitoring," *J. Hydrol.*, vol. 556, pp. 409–418, Jan. 2018.
[35] S. Lange and M. Riedmiller, "Deep auto-encoder neural networks in reinforcement learning," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2010, pp. 1–8.
[36] S. Groot, R. Harmanny, H. Driessen, and A. Yarovoy, "Human motion classification using a particle filter approach: Multiple model particle filtering applied to the micro-Doppler spectrum," *Int. J. Microw. Wireless Technol.*, vol. 5, no. 3, pp. 391–399, Jun. 2013.
[37] J. N. Kapur, P. K. Sahoo, and A. K. C. Wong, "A new method for gray-level picture thresholding using the entropy of the histogram," *Comput. Vis., Graph., Image Process.*, vol. 29, no. 3, pp. 273–285, Mar. 1985.

XIAQING YANG (Student Member, IEEE) received the B.S. degree in electronic engineering from the University of Electronic Science and Technology of China (UESTC), Chengdu, China, in 2013, where she is currently pursuing the Ph.D. degree in signal and information processing.

PENGYUN CHEN received the B.S. degree in electronic information engineering and the M.S. degree in communication engineering from Xinjiang University, Xinjiang, China, in 2015 and 2018, respectively. He is currently pursuing the Ph.D. degree with the School of Information and Communication Engineering, University of Electronic Science and Technology of China (UESTC), Chengdu, China.

His research interest includes radar-based human target motion recognition.

MINGYANG WANG (Member, IEEE) received the Ph.D. degree from the University of Electronic Science and Technology of China (UESTC), in 2019.

He was a Visiting Student with Temple University, Philadelphia, PA, USA, from November 2016 to February 2018. He is currently an Algorithm Engineer with the Southwest Institute of Electronic Technology. His research interests include information fusion, target recognition, machine learning, statistical signal processing, and pattern discovery.

SHISHENG GUO (Member, IEEE) received the B.S. degree in communication engineering from Nanchang Hangkong University, Nanchang, China, in 2013, and the Ph.D. degree in signal and information processing from the University of Electronic Science and Technology of China (UESTC), Chengdu, China, in 2019.

He is currently an Associate Researcher with the School of Information and Communication Engineering, UESTC. His research interests include through-the-wall radar imaging, signal analysis, and NLOS target detection.

CHAO JIA received the M.S. degree from the University of Electronic Science and Technology of China (UESTC), Chengdu, China, in 2013, where he is currently pursuing the Ph.D. degree with the School of Information and Communication Engineering.

His research interests include NLOS target detection in urban environment and radar signal processing.

GUOLONG CUI (Senior Member, IEEE) received the B.S. degree in electronic information engineering and the M.S. and Ph.D. degrees in signal and information processing from the University of Electronic Science and Technology of China (UESTC), Chengdu, China, in 2005, 2008, and 2012, respectively.

From January 2011 to April 2011, he was a Visiting Researcher with the University of Naples Federico II, Naples, Italy. From June 2012 to August 2013, he was a Postdoctoral Researcher with the Department of Electrical and Computer Engineering, Stevens Institute of Technology, Hoboken, NJ, USA. From September 2013 to July 2018, he was an Associate Professor with UESTC, where he has been a Professor, since August 2018. His current research interests include cognitive radar, array signal processing, MIMO radar, and through-the-wall radar.