Early Gesture Recognition with Reliable Accuracy Based on High Resolution IoT Radar Sensors

Rui Min, *Member, IEEE,* Xing Wang, *Student Member, IEEE,* Jie Zou, Jing Gao, Liying Wang, *Student Member, IEEE,* and Zongjie Cao, *Member, IEEE*

***Abstract*—Early recognition of gestures gives a better user experience during human-computer interaction in a near real-time way. However, the inadequate information accumulated in the early stages of the gesture will make the machine ambiguous and respond wrongly. Therefore, the challenge of early recognition lies in choosing the right time to trigger gesture prediction and using limited information to achieve reliable accuracy. In this paper, an early gesture recognition method is proposed to achieve almost the same accuracy as a complete gesture sequence. For this purpose, we design a new architecture to modify the current feature sequence according to the existing information, and introduce a new loss function to maximize the probability of correct gesture label as early as possible. At the same time, gesture recognition is triggered when the maximum probability of model output is greater than a set threshold. The publicly available dataset used for evaluation comes from Google's Project Soli sensor. Experimental results show that the proposed method for early recognition can achieve a gesture recognition rate of 88.85% with an average time saving of 69.47%. Compared with the results of all sequence recognition, the recognition rate drops by only 1.7%. In addition, the proposed method based on the dataset constructed by terahertz radar also shows similar early recognition performance and can be applied to other radar sensors.**

***Index Terms*—Early recognition, gesture prediction, reliable accuracy, radar sensor.**

I. Introduction

THE innovation of communication technology makes the vision of interconnection of all things come true gradually. With the lower latency, higher transmission speed and capacity brought by the next-generation (6G) network, the Internet of Things (IoT) will provide more intelligent and convenient services, such as human-computer interaction, autonomous driving, smart home and other application scenarios [1]. With the increase of access sensors and the complexity of application scenarios, 6G-enabled IoT is also expanding towards high-frequency spectrum resources such as millimeter wave and terahertz [2]. This will lead to better connectivity and intelligent perception [3], [4]. Especially for high-frequency radar sensors, it can capture the tiny movement of the target with its high resolution. Combined with its other characteristics, radar sensors will have a wide range of applications in the 6G-enabled IoT. Among them, gesture recognition based on radar sensors is a new type of human-computer interaction.

Gesture recognition is an exploration of a new generation of human-computer interaction, aiming to use a more natural and convenient way to talk with machines. A typical application scenario is vehicle interface interaction [5], [6]. Through the interaction between the gesture and the control system in the vehicle, it can avoid the change of vision caused by looking for the interface and button position, and ensure the safety in the driving process. Besides, gestures are used as a way for people to express their thoughts and feelings [7]. It strengthens the information conveyed in our daily conversations and is easy to learn and understand. Therefore, many advantages of gesture interaction make the gesture become a new information carrier, which is used in a wide range of human-computer interaction scenarios.

The realization of gesture recognition is closely related to the development of sensor technology [8]. Different sensors perceive the movement of gesture in different ways. At present, there are four kinds of sensors used in gesture recognition, including optical sensor, ultrasonic sensor, inertial sensor and radar sensor [9], [10], [11], [12]. Compared with other sensors, radar sensors have some significant advantages in gesture recognition. First of all, the high-resolution radar can effectively sense the small spatial changes of fingers and muscles during the hand movement [13]. Secondly, radar is robust to environmental changes in application scenarios and can adapt to environments such as darkness, occlusion, and smoke. Moreover, the signal collected by the radar sensor has the function of privacy protection compared with video [14]. Finally, the radar does not need to wear sensors on the hands to sense gesture movements like data gloves. This flexible setting makes our solution has a wide range of application scenarios.

The combination of gestures and radar has further promoted the development of gesture recognition. The radar prototype used to sense gesture motion has also been designed, laying the hardware foundation of this field [13], [15], [16]. One of the projects called Google Soli has attracted the attention of many people. It is a new, robust, high-resolution, low-power, miniature gesture sensing technology for human-computer interaction based on millimeter-wave radar, and it constructs a framework from the abstraction of the bottom hardware to the top software application. And the open source dataset can be used by more researchers to conduct research on gesture

This work was supported by the National Natural Science Foundation of China under Grant 62071100, 61971101 and 61971104. *(Corresponding author: Zongjie Cao.)*

R. Min, X. Wang, L. Wang and Z. Cao are with the School of Information and Communication Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China (e-mail: minrui@uestc.edu.cn; xwang@std.uestc.edu.cn; lywang_uestc@163.com; zjcao@uestc.edu.cn).

J. Z and J. G are with the Second Research Institute of CAAC, China (e-mail: zoujie@caacsri.com; gaojing@caacsri.com).

recognition. Although gesture movement is obvious from the perspective of the human, there are differences in speed and range of motion between different people even if they perform the same gesture. At the same time, the duration of gesture is relatively short, which brings a huge challenge to the accurate recognition of gesture. In addition, some gesture interaction systems have a high demand for low delay, such as game interaction and industrial control [17], [18]. The time cost of gesture recognition and response in interactive system will greatly affect the user experience. This requires identifying the gesture category at the early stage of the gesture, leaving more time for the interactive system to respond to the control instructions indicated by the gesture. Therefore, there are two main challenges in radar-based gesture interaction: high accuracy and low time delay of gesture recognition. However, the relationship between the two is opposed to each other. Although the early recognition of gestures reduces the time delay, the insufficient accumulation of motion information reduces the accuracy, and vice versa. In order to reduce the time cost as much as possible while ensuring the accuracy of early recognition, it is necessary to find a right time in the gesture movement to trigger the early prediction of gesture.

In this paper, we propose a direct but effective early recognition framework to achieve reliable accuracy in the early stages of gesture movement. Considering the gesture data stream obtained by different radar systems, the early gesture recognition model is constructed through the enhancement of early motion features and the loss function based on motion process modeling. Specifically, compared with the existed related research, the contribution of this paper can be summarized as follows.

1) A framework for early recognition of gesture is designed and constructed. The framework introduces the CumSoft block on the basis of recurrent neural network to modify early insufficient motion information and improve the contribution of important features to early recognition.
2) A new loss function based on gesture motion process modeling is proposed. The loss function makes the probability of correct gesture category in the early stage of motion maximum and increases with time, so as to ensure the accuracy of early recognition.
3) An early trigger strategy is proposed to determine the appropriate time for early prediction of gesture, and make a trade-off between the accuracy and delay of gesture recognition. This strategy can complete the early gesture recognition based on the online data stream.

The gesture dataset based on Google Soli proves that our framework can achieve high recognition accuracy in the early stage of gesture movement. Compared with the complete gesture sequence recognition, the recognition accuracy is only slightly decreased while obtaining time performance. What's more, our method can be used in radar data streams with inconsistent gesture durations to build low-latency gesture recognition systems.

The rest of this paper is organized as follows. Section II reviews the related work of early gesture recognition. Section III systematically introduces the proposed method, including the model architecture, newly defined loss function, and early triggering strategy. The relevant details and results of the experiment are presented in Section IV. Section V analyzes and discusses the impact of different modules on early recognition. Finally, Section VI is the conclusions and future work.

II. Related Work

Early gesture recognition is an important research direction of human-computer interaction. This research is of great significance for improving user experience and optimizing information transmission. This section will discuss several issues related to this topic, namely radar-based gesture recognition and early prediction of time series.

A. Gesture recognition based on radar sensor

At present, many radar-based gesture recognition researches have been further proposed, including the acquisition of radar signals [19], the extraction of micro-Doppler features [20], and the construction of deep network models [21]. These works have laid a good foundation for certain research directions of gesture recognition and met the requirements of common applications for recognition accuracy. Due to the complexity and diversity of gesture movement, feature extraction and classification models based on radar signals have become important parts of gesture recognition. Zhou *et al*. [22] used terahertz radar to acquire high-resolution distance profile sequences of gesture motion and corresponding Doppler signals to characterize the motion features of gestures, and calculated the similarity between gestures based on the extended Dynamic Time Warping (DTW) algorithm to complete the classification. Berenguer *et al*. [23] developed an unsupervised frame-level representation learning strategy with the NetVLAD approach [24] for gesture recognition. It can encode important information across time and acquire highly differentiated gesture features. Hazra *et al*. [25] proposed to use 3D Deep Convolutional Neural Network (3D-DCNN) architecture to learn the embedding model base on a video of range-Doppler images (RDI), which achieved high accuracy of gesture recognition on 60GHz short-range radar sensor. Zhang *et al*. [26] proposed a network based on 3DCNN Long Short Term Memory (3DCNN-LSTM) and introduced a Connectionist Temporal Classification (CTC) loss function to infer the boundaries between gestures and improve the classification accuracy of variable-length gestures.

All of the above methods show exciting gesture recognition accuracy in the corresponding dataset. And it also gradually develops from manual feature extraction to model training of deep network, which further improves the generalization performance of the model. However, these methods focus on sequence-level gesture recognition, which is not conducive to the construction of real-time response system. In addition, the average accuracy of per-sample and per-gesture in gesture recognition are compared [13]. The higher recognition rate of complete gesture indicates that the acquisition of insufficient gesture information results in the decrease of accuracy. In order to be closer to the actual application scenario, the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2021.3072169, IEEE Internet of Things Journal

research on gesture recognition needs to pay more attention to the early accuracy of gesture recognition, hoping to maximize the accuracy at the earliest possible time.

B. Early prediction of time series

The work related to early gesture recognition is classification of time series [27]. Given a gesture database *TS*, sample $(x, y) \in TS$ is a time series $x = (x_1, x_2, \cdots, x_T)$ corresponding to the gesture label y, and T is the length of time series. $\{x_t\}_{t=1}^{T}$ represents the movement state of gesture at time t. Generally, the judgment of gesture category is realized by extracting the complete motion change corresponding to a gesture. Limited by the time delay of complete gesture sequence recognition, part of the research shifted to early recognition of time series. Similar to the early prediction of time series in [28], [29], it is desirable to output the prediction label as early as possible before the gesture is completed.

To achieve this challenging task, Aliakbarian *et al.* [30] developed a multi-stage recurrent architecture based on LSTM to fully extract context-aware and action-aware information in the activity. At the same time, a loss function is used to penalize the error classification in each time step linearly, encouraging the model to make correct category predictions as soon as possible. However, this method only uses part of the activity sequence and lacks process modeling for activities, which is not conducive to the construction of early recognition systems. A rule is needed to guide the early gesture recognition system online whether to wait for more data input or make decisions immediately.

Rußwurm *et al.* [31] estimates a stopping probability based on the dual loss function in the model, and optimizes the accuracy and earliness jointly. The stopping probability is used to determine the response time of the model. Compared with the corresponding baseline, the accuracy of early recognition is slightly reduced. Therefore, it is necessary to maximize the recognition accuracy as much as possible while gaining time performance in the early stage of the gesture.

Ma *et al.* [32] proposed a new ranking loss function to encourage the model to make more confident decisions during the activity. It simulates the probability of the correct category or the probability difference between the correct category and the wrong category in the activity is non-decreasing over time. Although this method is helpful for the model to make a decision in the early stage, it does not explicitly constrain the probability distribution of the category at the time when the model triggers the prediction, so as to ensure the accuracy of the early recognition.

Motivated by the above methods, the framework we constructed is based on the gesture data stream and introduces a loss function to constrain the probability of correct gesture category. Through modeling the process of gesture movement, it is used to guide the model to trigger the early recognition of gesture at the appropriate time. Given the insufficient accumulation of gesture information in the early stage, the early motion features are corrected based on the degree of contribution of gesture recognition, so as to achieve time advance and reliable accuracy at the same time.

III. Proposed Method

To achieve early gesture recognition with reliable accuracy, this paper accomplishes this challenging task from three aspects. First, the designed model pays more attention to the characteristics of key moments on the known input sequence. Second, the loss function maximizes the probability of correct category in the early stages of the gesture. Finally, a trigger strategy for early gesture recognition is proposed.

A. Model architecture

Gesture motion can be abstracted into a time series $x = (x_1, x_2, \cdots, x_T)$, where x_t corresponds to state information at time t. In our work, the range-Doppler map (RDM) is used to describe the movement state of gestures, reflecting the distance, speed, and energy distribution of multiple scattering centers on the hand [16]. In order to extract the spatiotemporal features of gesture motion and process modeling, Convolutional Neural Network (CNN) and Gated Recurrent Unit (GRU) are introduced respectively [33], [34]. Therefore, the overall architecture of our model shown in Fig. 1.

First of all, the RDM generated by gesture motion is used as the input of CNN, which reflects the distance and velocity information of multiple scatters on hand [16]. By learning features automatically, CNN maps the RDM at time t_k to an intermediate representation vector f_k. Subsequently, the feature sequences $(f_1, f_2, \cdots, f_T)$ model the process of gesture movement based on GRU. The GRU takes the feature f_k and the hidden state h_{k-1} of the previous time step as input, and performs the following conversion:

$$z_k = \sigma\left(W_z\left[h_{k-1}, f_k\right]\right), \tag{1}$$

$$r_k = \sigma\left(W_r\left[h_{k-1}, f_k\right]\right), \tag{2}$$

$$\tilde{h}_k = \tanh\left(W\left[r_k \odot h_{k-1}, f_k\right]\right), \tag{3}$$

$$h_k = \left(1 - z_k\right) \odot h_{k-1} + z_k \odot \tilde{h}_k, \tag{4}$$

where σ and tanh are activation functions in neural network, corresponding to sigmoid and hyperbolic tangent function respectively. $\odot$ is Hadamard product and z_k, r_k correspond to the output of update gate and forget gate in the GRU unit, respectively. Besides, W_z, W_r and W correspond to the internal parameters learned by the update gate z_k, the forget gate r_k and the candidate activation $\tilde{h}_k$ during the model training process. These parameters are shared in time sequence and are continuously learned through the back propagation algorithm. Finally, the output of the GRU unit at each time step is sent to the Dense layer, and the softmax activation function is used to obtain the probability of different gesture categories at the current moment.

The similar combination of CNN and GRU has achieved good results on many complete time series classification tasks [32], [35]. For early gesture recognition, the motion feature displayed in the early stage of the gesture is vague, especially for micro gestures with high similarity. This leads to the limited accuracy of early recognition. On this basis, we introduced a CumSoft block to modify the feature output after CNN. Its main function is to use the accumulated motion information to evaluate the overall contribution of input data at the current

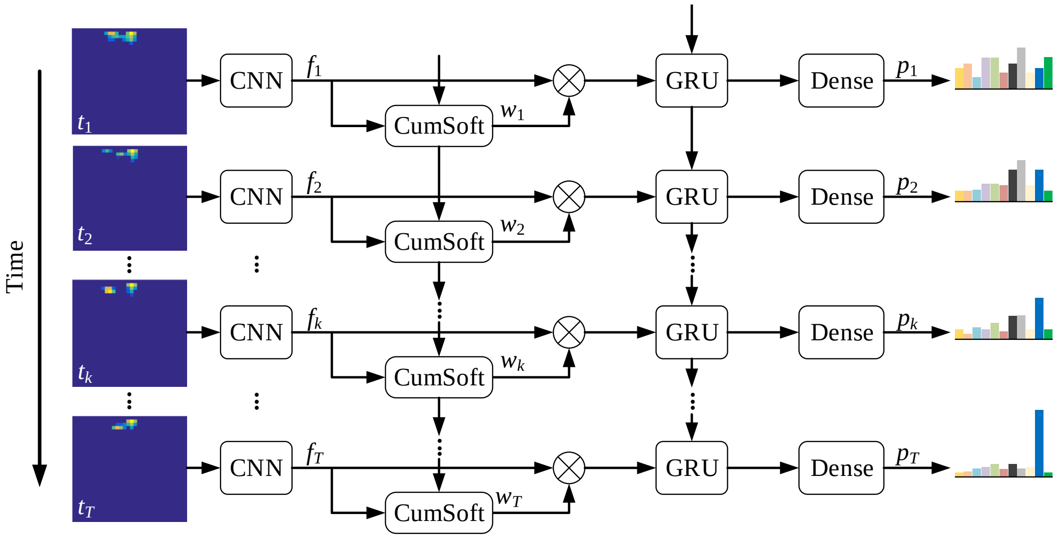


Fig. 1. Model architecture overview. First, CNN performs gesture feature extraction on RDM output by radar sensor, and then these features are modified by a CumSoft block and sent to GRU for early gesture recognition.

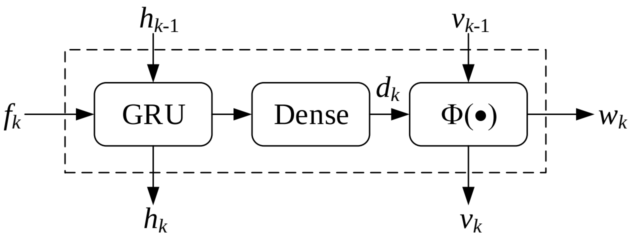


Fig. 2. The structure of CumSoft block. The current feature input f_k is used to evaluate the contribution to gesture recognition.

time step to gesture recognition. Next, the contribution score is used to weight the feature information.

As shown in Fig. 2, the gesture feature f_k extracted by CNN is sent to GRU and Dense layer, and its output d_k is considered to be an important representation of the information provided by the current input data. For early recognition, it is difficult to obtain the proportion of the global contribution based on partial information. Therefore, an approximate estimate is used. Here, $\Phi(\bullet)$ is an approximation of Softmax function, which uses the current known sequence to estimate the contribution score in the complete gesture sequence. It can be defined as

$$w_k = \frac{1 + d_k + d_k^2/2}{T + \sum\limits_{t=1}^{k} \left(d_t + d_t^2/2\right)}, \tag{5}$$

where T is the total time step included in a complete sequence of gesture. In order to adapt to the structure of our model, let

$$v_k = \begin{cases} T, & k = 0 \\ v_{k-1} + \left(d_k + d_k^2/2\right), & 1 \le k \le T \end{cases}, \tag{6}$$

then Eq. 5 can be replaced by

$$w_k = \frac{1 + \left(d_k + d_k^2/2\right)}{v_k}. \tag{7}$$

These parameters can be recursive based on the past, without the need to introduce the later gesture information. This is conducive to the early recognition of gesture data flow. What's more, the output of this block will guide model to pay more attention to the important moment features in the early stage of the gesture, so as to further alleviate the contradiction between early accurate recognition and insufficient gesture information.

B. Loss function

In the traditional sequence recognition framework, the model loss only focuses on the category probability distribution output at the last time step, and does not restrict the early probability distribution strongly. For early recognition, ambiguous gesture motion makes the category probability distribution output by the model very different from the actual one, resulting in limited recognition accuracy. In addition, the human have an intuitive feeling for gesture judgment. As gesture progresses, the shape and movement trajectory are more and more clearly displayed, and the judgment of gesture category is more determined. Analogously, the more accumulated gesture motion features, the greater the probability output of the model for the correct category.

Therefore, we redefine the loss function based on two basic principles. On the one hand, the output of the correct category probability is maximized as early as possible. On the other hand, the probability output of the correct category gradually increases with time. The detailed instructions are as follows:

$$L_1 = -\sum_t \frac{t}{T} \log\left(1 - \left(\max\left(p_t\right) - \hat{p}_t\right)\right), \tag{8}$$

$$L_2 = \sum_t \frac{t}{T} \left|\hat{p}_t - \mu \frac{t}{T}\right|, \tag{9}$$

where $p_t = (p_{t,1}, p_{t,2}, \cdots, p_{t,m})$ is the probability score of m gesture categories output by the model at the t-th time step, and $\hat{p}_t$ is the probability score of the correct category in p_t.

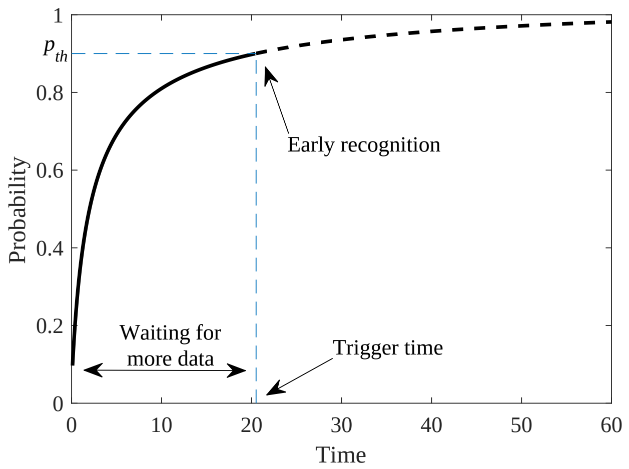


Fig. 3. The probability curve of correct gesture category. The newly defined loss function expects the probability of the correct gesture category to be the largest and gradually increases with time. p_{th} is used to guide the model whether to wait for data or trigger early recognition.

In addition, μ is a positive scalar that control the slope of the probability curve for the correct gesture category.

Therefore, the final loss function applied to our model is defined as:

$$L = L_1 + \lambda L_2, \tag{10}$$

where λ is also a positive scalar for weighting the respective losses.

Among the loss functions defined above, the time factor t/T is introduced in both L_1 and L_2, which makes the cost of misclassification in the later stage of the gesture increasingly higher, prompting the model to make decisions as early as possible. In L_1, as long as the probability of correct gesture category is greater than other categories in the model training process, no loss will be introduced. This loose condition makes the model easier to train. What's more, L_2 ensures that the probability of correct gesture category increases gradually, which makes the probability output of the model more confidently fit the actual distribution.

C. Early trigger strategy

The proposed loss function ensures that the probability output of the correct gesture category reaches a high level in the early stage of the gesture from two aspects, and then gradually approaches 1. In turn, the gesture probability output by the model can guide the selection of the triggering moment of the gesture.

For early recognition, low latency and high precision are always mutually restrictive [36]. Choosing the right trigger point is also the result of compromise between various conditions [37]. Therefore, this paper introduces a trigger threshold $p_{th} \in [0, 1]$ to determine the triggering moment of early recognition shown in Fig. 3. When the following condition is met for the first time:

$$\max(p_t) \geq p_{th}, \tag{11}$$

the corresponding time t will trigger early gesture recognition. If the condition cannot be satisfied when the gesture is completed, gesture recognition is triggered at the last time step. In short, the triggering moment of early gesture recognition is defined as:

$$t^* = \begin{cases} T, & \max(p_t) < p_{th} \\ \min\{t | \max(p_t) \geq p_{th}\}, & otherwise \end{cases} \tag{12}$$

More specifically, the implementation of this strategy is divided into two parts. First, the model parameters are trained using a complete gesture sequence, aiming to use the global timing information to better achieve the expected goal of the loss function. Then, these trained model parameters are used to predict the probability that the input data stream belongs to different gesture categories, so as to select the appropriate time to trigger recognition. It should be noted that the parameters of each block in the model are shared, and the data information after the current time will not be introduced, which makes the early prediction and the complete sequence identification have the same output at the same time.

IV. Experiments

In this section, different radar gesture datasets used to evaluate the performance of our method are first described, including the publicly available Soli dataset and our own constructed THz dataset. Secondly, we further explain the model structure and parameter configuration in the experiment. Finally, the advantages of this method in early gesture recognition are verified by experiments. Specially, all experiments are conducted on a computer with Intel i7-8750H processor and a Nvidia Geforce GTX 1050Ti GPU card. The operating system is Windows with CUDNN v10.

A. Dataset

The Soli dataset[1] used in this paper comes from Google's Project Soli sensor. Soli is a new sensing technology that uses micro-millimeter-wave radar to recognize gestures. By receiving the electromagnetic waves reflected by hand, soli radar can extract rich motion features about gesture movement, including size, shape, direction, distance, and speed.

This dataset contains 11 types of gestures shown in Fig. 4, such as Swipe, Push, Circle, Palm hold, etc. The more detailed design instructions and considerations are given in [35]. A complete gesture movement is represented by 4 RDM sequences, derived from the 4 receiving antennas of Soli radar. At the same time, the dataset was collected multiple times by different users and contained a total of 5,500 gesture samples. These data come from two parts. In one part, 10 users perform each gesture 25 times, generating 10 × 11 × 25 = 2750 gesture sequences. The other part also contains 2750 samples generated by a single user across six sessions.

Different from the dataset division method in [35], the gesture samples of the training set and the test set correspond to the gesture data completed by different users. In this way, the generalization performance of the trained model to recognize different user gestures can be evaluated, and the

[1] https://github.com/simonwsw/deep-soli

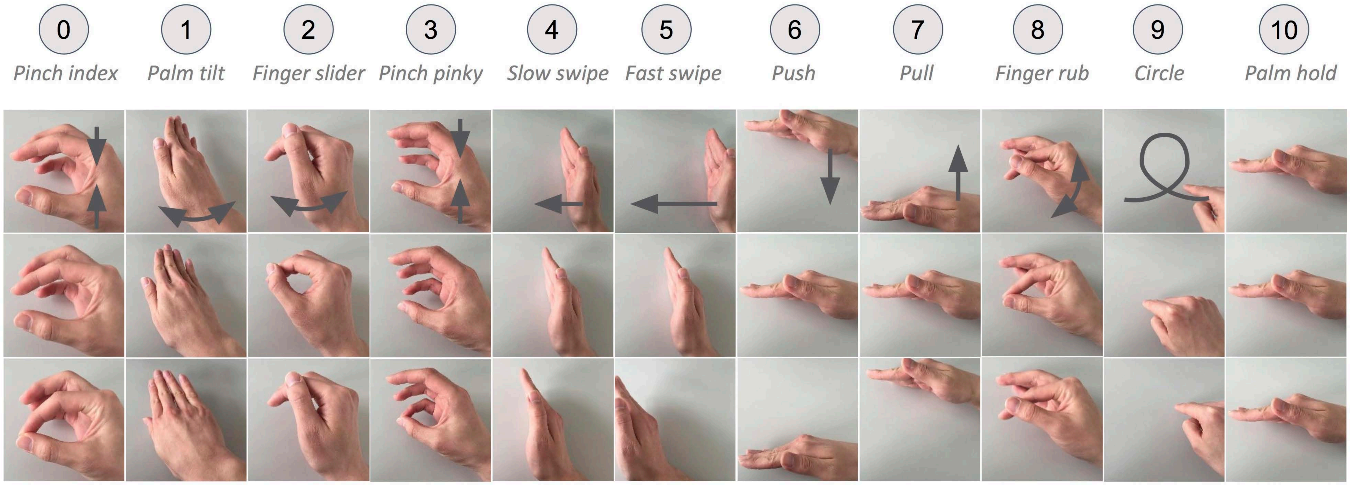


Fig. 4. Each column represents the three important steps of a gesture in Google Soli dataset. The gesture label is indicated by the number in the circle above.

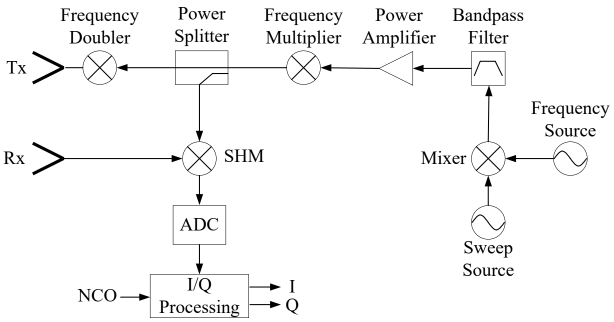


Fig. 5. The structure of the terahertz radar system.

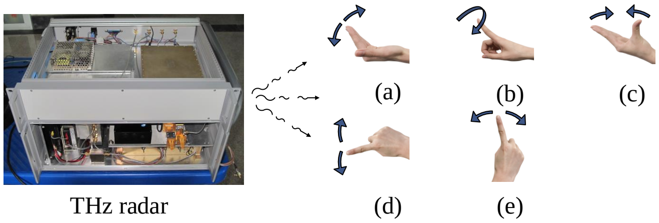


Fig. 6. Five gestures included in THz dataset. (a) beckon, (b) circle, (c) grasp, (d) left and right, (e) forward and backward. The arrows in the figure show the trajectory of hand.

purpose is to hope that the trained model has ideal early recognition performance on unfamiliar users. In the subsequent experiments, 2750 samples generated by different users are used to train and verify the model, and the remaining samples generated by a single user are used to evaluate the early recognition performance of the proposed method. In particular, we perform incoherent accumulation on the RDM obtained by the four antennas, and take its average value as the gesture data input. Inspired by the process modeling in [38], the cumulative input of motion sequences will be explicitly put into the process information to enhance the characterization of gesture sequences.

What's more, the THz dataset is based on terahertz radar sensor. Terahertz (THz) waves refer to electromagnetic waves with a frequency in the range of 0.1 to 10 THz, which are widely used due to their high frequency and large bandwidth. The structure of our terahertz radar is illustrated in Fig. 5. It operates at a central frequency of 220 GHz. The transmitted signal is based on a FMCW modulation scheme that its frequency changing with time. This sweep in frequency is commonly referred to as a chirp. For the terahertz radar system, the chirp bandwidth is 12.8 GHz and chirp time is 1ms. Terahertz radar obtains high resolution in speed and range with high carrier frequency and large bandwidth [39]. Similar to the working principle of the soli sensor, the terahertz radar uses the electromagnetic waves reflected by hand to extract the motion characteristics of gesture. Therefore, the THz dataset is used as a data supplement to verify the adaptability of the proposed method to different radar systems.

In the THz dataset, five common types of gestures are designed for early recognition, as shown in Fig. 6. This dataset was completed by 7 users, each of whom repeated 30 times for each type of gesture, resulting in 7 × 5 × 30 = 1050 samples. To have a unified format with the Soli dataset, the echo signal output by the terahertz radar is transformed into RDM sequence corresponding to gesture through the preprocessing step [40], [41]. For each kind of gesture, the samples of two users selected randomly constitute a test set, and all remaining samples constitute a training set and a verification set. Compared with the number of samples in the Soli dataset, this dataset is expanded by adding Gaussian noise to reduce the chance of over fitting [42].

B. Implementation Details

The model framework of this paper is implemented in Keras [43], [44]. It is an open source neural network library, which is designed to be an effective tool for rapid experiments. The parameters and configuration of the model in keras are described below.

In the CNN block, a Conv2D layer is defined as the input layer with 2 filters and a 2x2 kernel to pass across the input

TABLE I
THE PERFORMANCE COMPARISON OF GESTURE RECOGNITION BASED ON EARLY TRIGGER STRATEGY WITH DIFFERENT MODULE CONFIGURATION. THIS EXPERIMENTAL RESULTS ARE DERIVED FROM THE SOLI DATASET.

Method	C_a(%)	C_e(%)	t_e(timesteps)	Reduced accuracy(%)	Time saving(%)
Baseline	88.84 ± 2.38	73.52 ± 4.81	19.47 ± 2.28	17.24	67.55
Baseline + Our Loss	88.10 ± 1.13	84.71 ± 1.45	**17.21** ± 0.23	3.85	**71.32**
Ours	**90.42** ± 1.39	**88.85** ± 1.10	18.72 ± 1.05	**1.74**	69.47

TABLE II
THE PERFORMANCE COMPARISON OF GESTURE RECOGNITION BASED ON EARLY TRIGGER STRATEGY WITH DIFFERENT MODULE CONFIGURATION. THIS EXPERIMENTAL RESULTS ARE DERIVED FROM THE THZ DATASET.

Method	C_a(%)	C_e(%)	t_e(timesteps)	Reduced accuracy(%)	Time saving(%)
Baseline	70.18 ± 5.89	50.40 ± 10.62	10.12 ± 3.75	28.18	83.13
Baseline + Our Loss	79.34 ± 5.64	75.81 ± 1.62	**8.6** ± 2.57	**4.45**	**85.67**
Ours	**82.61** ± 3.36	**77.54** ± 2.81	12.63 ± 2.71	6.14	78.95

sequences. The Conv2D layer is followed by a MaxPooling2D layer with a pool size of 2x2, which will in effect halve the size of each filter output from the previous layer. Next, a Flatten layer compresses the extracted features and outputs a one-dimensional vector. Finally, each layer in the CNN block is wrapped in TimeDistributed layer, and then added to the main model.

In the CumSoft block, the GRU layer contains 16 memory cells. After the output passes through a Dense layer, the scalar d_k is output at each time step. Finally, the modified feature is fed to another GRU layer containing 64 memory cells, and its output is sent to the Dense layer activated by softmax function, and the corresponding probability of gestures is output.

During the training phase of our model, the output of CNN is added with an additive Gaussian noise whose mean value is 0 and standard deviation is 0.05, so as to reduce the risk of over fitting. At the same time, the Adam optimizer with a learning rate of 0.001 is used to update all parameters of the model, and batchsize is set to 32.

In addition, two parameters μ and λ are introduced into the loss function, which are set to 10 and 2 respectively. The values of these two parameters are determined by the method similar to grid search in machine learning. By setting a series of discrete values in a reasonable range, the optimal experimental results are found. The initialization parameter T of v_0 in the CumSoft block is set to 60. This is obtained by calculating the average length of each RDM sequence corresponding to each gesture. In order to consider the delay and accuracy of gesture recognition, the trigger threshold p_{th} is set to 90%.

C. Results

To evaluate the performance of the proposed method for early accurate recognition, a comparison baseline is set up. The difference is that our model introduces the CumSoft block and defines a new loss function. In the baseline setting, the loss function selects the traditional Categorical Cross Entropy (CCE), and only calculates the loss in the last time step. All performance evaluations are based on our early trigger strategy. In addition, some evaluation indexes are further explained:

1) C_a represents the recognition rate of using complete gesture sequence.
2) C_e represents the recognition rate obtained by the early trigger strategy.
3) t_e represents the average time step spent by early trigger strategy.

Therefore, C_a, C_e and t_e are defined as

$$C_a = \frac{1}{|TS_{test}|} \sum_{(x,y)\in TS_{test}} I\left(\underset{j=1,2,\cdots,m}{argmax}\ (p_{T,j}) = y\right), \quad (13)$$

$$C_e = \frac{1}{|TS_{test}|} \sum_{(x,y)\in \mathrm{TS}_{test}} I\left(\underset{j=1,2,\cdots,m}{argmax}\ (p_{t_x^*,j}) = y\right), \quad (14)$$

$$t_e = \frac{1}{|TS_{test}|} \sum_{(x,y)\in TS_{test}} t_x^*, \quad (15)$$

where $|TS_{test}|$ is the number of samples included in the test data set, and t_x^* represents the timestep taken by sample (x,y) to trigger early gesture recognition. $I(\cdot)$ is the indicator function, that is, when $a=b$, $I(a=b)$ is 1, otherwise it is 0. Compared with the complete gesture sequence recognition, the reduced accuracy and time saving of early recognition can be expressed as

$$\text{Reduced accuracy} = \left(1 - \frac{C_e}{C_a}\right) \times 100\%, \quad (16)$$

$$\text{Time saving} = \left(1 - \frac{t_e}{T}\right) \times 100\%. \quad (17)$$

Table I and II evaluate the performance of different module configurations in the proposed method for early gesture recognition based on the Soli dataset and the THz dataset, respectively. In Table I, the implementation of early trigger strategy under the baseline setting obtained a time advance of 67.55% relative to the complete gesture sequence, but resulted in a 17.24% decrease in accuracy. Such loss of accuracy is unacceptable for high reliability occasions. By analyzing the output probability of some test samples in the experiment, the setting of baseline cannot guarantee the correctness of the corresponding label when the gesture reaches the trigger threshold in the early stage, which leads to the false trigger and reduces the accuracy of early recognition. Subsequently, the newly defined loss function proposed in this paper improved the accuracy of early recognition by about 11%, and the time

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2021.3072169, IEEE Internet of Things Journal

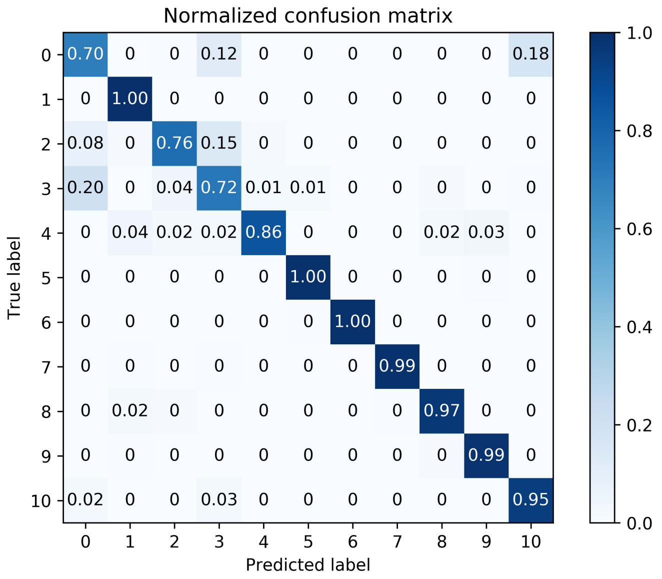


Fig. 7. Confusion matrix of gesture recognition under early trigger strategy. The experimental results are based on the Soli dataset.

performance is slightly improved. This is because our loss function constraint model has the highest output probability of correct category in the early stage of gesture. Finally, the entire method proposed in this paper is more advantageous than the results of baseline in the comparison of various evaluation indexes. At the same time, it is also obvious that the CumSoft block has improved the accuracy of early gesture recognition, but slightly reduced the performance in time.

In summary, the method proposed in this paper can achieve reliable accuracy in the early stage of gesture movement. Compared with the complete gesture sequence recognition, its accuracy drops by only about 1.74%. The confusion matrix for early gesture recognition is shown in Fig. 7. Except for Pinch index, Finger slider, and Pinch pinky, other gesture categories have higher recognition rates. In addition, the main misclassifications are focused on the three types of tiny gestures with high similarity. Similar experimental results are also shown in [35], which indicates that the tiny gesture categories with high similarity selected by the Soli dataset pose challenges for accurate gesture recognition.

The results in Table II are used as the comparison between data sets to evaluate the adaptability of the proposed method to different radar systems. On the whole, the proposed method is basically consistent in the THz dataset, and can improve the accuracy of gesture recognition in the early stage. But the Cumsoft block will delay the gesture trigger time in exchange for further improvement in accuracy. This rule is consistent with conventional cognition, that is, more time is accumulated for higher accuracy. In particular, the results in Table II show that the accuracy of early recognition is lower than that of the soli dataset. This is due to the random selection of data samples from different users when the test set is divided, which puts forward higher requirements for the generalization ability of the model.

Table III shows the experimental results of different early recognition methods on the Soli dataset. It can be seen that our method achieves the best results in accuracy and delay. Although the early accuracy of [31] is similar to ours, it needs more time to wait. The results show that our method has obvious advantages in reducing the time delay.

TABLE III
COMPARISON OF DIFFERENT EARLY RECOGNITION METHODS BASED ON THE SOLI DATASET.

Method	$C_a(\%)$	$C_e(\%)$	t_e(timesteps)
LSTM-m of [32]	86.42	86.36	34.04
LSTM-s of [32]	86.96	86.72	32.37
End-to-end of [31]	90.29	88.89	29.45
Ours	90.42	88.85	18.72

V. ANALYSIS

In this section, we will conduct a detailed analysis of the proposed method to reveal the impact of different modules on early recognition.

A. *Internal visualization of model*

Direct visualization is used to describe the impact of CumSoft block and loss function on early recognition, respectively. For the CumSoft block, the output sequence $(w_1, w_2, \cdots, w_T)$ represents the contribution of the extracted features to the gesture recognition at different moments, so as to modify the motion features. Fig. 8 shows the average w_k changes of 11 gestures after model training. It can be seen that the most important feature segments of gestures appear at different times and have significant differences. Single gesture sequence has a great contribution to early recognition only at certain important moments, while other moments have a little contribution to gesture judgment. For example, gestures 1, 4, 6, 7, and 8 contain important feature information in the early stage, and gestures 0, 2, 5, and 9 contain important feature information slightly behind, but all gestures contribute relatively little in the later stage. It shows that the core stage of gesture movement appears in different periods, mostly concentrated in the early and middle periods. This provides a theoretical basis for the early recognition of gesture.

Therefore, gesture recognition can be completed according to some meaningful sequence segments [45], [46], which will reduce the data redundancy and make the application of mobile terminal become reality. In addition, by further analyzing the core stage of gesture movement, we can optimize the human-computer interaction process and improve the user experience.

Besides, we also visualize the probability changes of different gesture categories to show the impact of the loss function in the modeling of the gesture process. In Fig. 9, the probability of correct gesture category increases with time and approaches 1, while other categories approach 0. In the early stages of gesture movement, the probability of certain wrong categories in the baseline far exceeded the correct category, causing confusion between gestures. This requires the model to accumulate more gesture information to correct the probability distribution between gestures. However, the

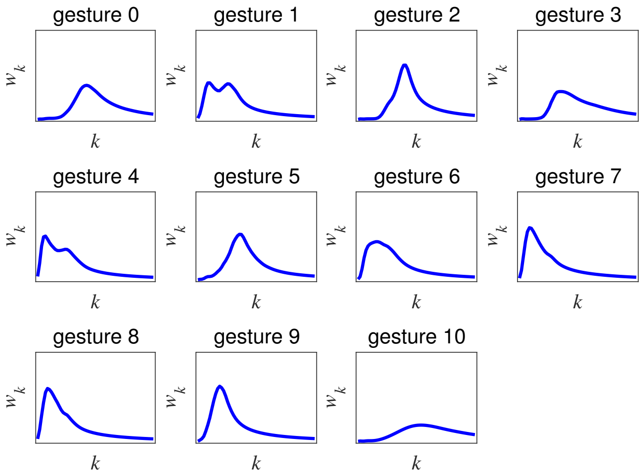


Fig. 8. Different gestures correspond to the average output w_k in the CumSoft block. This experimental results are derived from the test set samples in the Soli dataset. The vertical axis w_k of each subgraph represents the output value of the current gesture at the k-th time step.

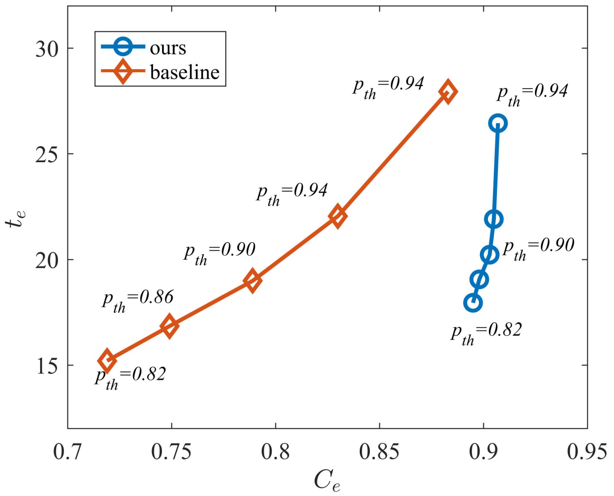
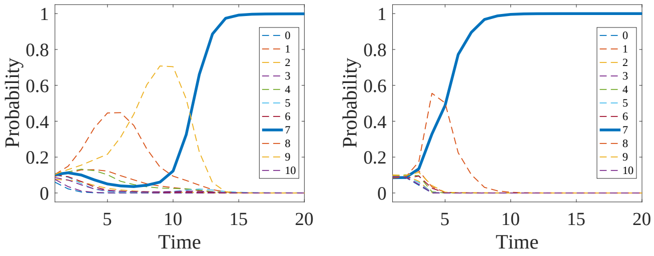


Fig. 10. Performance comparison of early gesture recognition under different trigger thresholds.



(a) baseline (b) ours

Fig. 9. Probability changes corresponding to different categories in gesture movement. The thicker curve in the figure indicates the probability change of the correct category.

probability of error categories in our method is obviously limited in the early stage of gesture, keeping a low level. This makes the probability of the correct category reach the threshold p_{th} earlier, thus triggering early gesture recognition. In other words, the proposed loss function further reduces the probability of error categories in the early stage of gesture by modeling the gesture process. So, the probability of correct category is maximized as early as possible, and the accuracy of early recognition is improved.

B. *Effect of trigger threshold on early recognition*

The setting of trigger threshold p_{th} is a compromise between accuracy and delay in early recognition. The final choice depends on the actual application scenarios. For example, a higher threshold p_{th} results in higher accuracy of early recognition, but more time cost. On the contrary, there are similar results. The impact of the trigger threshold on the proposed method will be further explored below.

When $p_{th} \in \{0.82, 0.86, 0.9, 0.94, 0.98\}$, the accuracy and delay corresponding to early recognition are shown in Fig. 10. On the whole, the increase of p_{th} can obtain higher accuracy, and also delay the triggering of gesture recognition.

Compared with the baseline setting, our method improves the accuracy of early recognition more based on different trigger thresholds. The curve corresponding to our method rises faster, but the improvement of accuracy slows down. This shows that the proposed method guarantees the maximum probability of correct category in the early stage of gesture movement. At this time, the change of p_{th} has a small impact on the recognition accuracy, but the impact on the time performance is relatively large. Therefore, the proposed method can achieve a higher early accuracy with a lower trigger threshold at the beginning. However, higher performance improvements have become increasingly difficult.

Although the increase of trigger threshold can improve the recognition rate, this effect is only meaningful in a certain range. It is not worth the loss to continue to use more time in exchange for a slight increase in recognition rate. Therefore, in our method, $p_{th} = 0.9$ is selected as the trigger threshold by grid search.

C. *Effect of loss function on early recognition*

The loss function is crucial for deep learning networks. We compare with the loss function used by some other time series tasks. For example, the exponentially weighted and linearly weighted cross-entropy losses are introduced in [47] and [30] for activity anticipation, respectively. This makes the punishment intensity in the later stage of the activity become greater and greater, which helps the model output the correct active label as soon as possible.

Similarly, our loss function also takes such measures to make the model make decisions earlier. In order to improve the accuracy of gesture recognition in the early stage, we use a direct way to restrict only the output probability of the correct category, so as to ensure that the probability of correct gesture is the highest when the gesture is triggered. In Table IV, the experimental results also show that our method has advantages in the accuracy and time delay of early recognition, that is,

TABLE IV
INFLUENCE OF DIFFERENT LOSS FUNCTIONS ON EARLY RECOGNITION PERFORMANCE BASED ON OUR MODEL FRAMEWORK.

Method	C_a(%)	C_e(%)	t_e(timesteps)
Ours+Loss of [47]	86.55	85.38	20.79
Ours+Loss of [30]	88.03	86.31	19.99
Ours	**90.42**	**88.85**	**18.72**

a higher recognition rate is obtained when there are fewer gesture sequences.

VI. CONCLUSION

In this paper, an early gesture recognition method is proposed, which can reduce the delay and obtain almost the same reliable accuracy as the complete gesture sequence. By modeling the process of gesture motion and enhancing the contribution of important moment features to gesture recognition, high recognition accuracy in the early stage of gestures is obtained. In addition, an early trigger strategy is proposed to select the right time for gesture recognition. The experimental results show that this method has the advantages of accuracy and time delay in early gesture recognition. And it has similar performance in other radar sensors. Therefore, the proposed method can achieve early gesture recognition tasks based on continuous data streams on different radar sensors. In future work, we plan to further explore the core feature segments of gesture motion to improve the recognition accuracy of high-similarity micro-gestures in early recognition.

REFERENCES

[1] L. Chettri and R. Bera, “A comprehensive survey on internet of things (iot) toward 5g wireless systems,” *IEEE Internet of Things Journal*, vol. 7, no. 1, pp. 16–32, 2020.

[2] F. Tang, Y. Kawamoto, N. Kato, and J. Liu, “Future intelligent and secure vehicular network toward 6g: Machine-learning approaches,” *Proceedings of the IEEE*, vol. 108, no. 2, pp. 292–307, 2020.

[3] Q. Ren and Q. Liang, “Throughput and energy-efficiency-aware protocol for ultrawideband communication in wireless sensor networks: a cross-layer approach,” *IEEE Transactions on Mobile Computing*, vol. 7, no. 6, pp. 805–816, 2008.

[4] Q. Liang, L. Wang, and Q. Ren, “Fault-tolerant and energy efficient cross-layer design for wireless sensor networks,” *International Journal of Sensor Networks*, vol. 2, no. 3-4, pp. 248–257, 2007.

[5] Q. Ye, L. Yang, and G. Xue, “Hand-free gesture recognition for vehicle infotainment system control,” *2018 IEEE Vehicular Networking Conference (VNC)*, pp. 1–2, 2018.

[6] A. Joshi, C. D. Joshi, A. M. Karambelkar, and S. B. Somani, “Distraction-free car dashboard control through gesture recognition,” 2019.

[7] M. J. Cheok, Z. Omar, and M. H. Jaward, “A review of hand gesture and sign language recognition techniques,” *International Journal of Machine Learning and Cybernetics*, vol. 10, pp. 131–153, 2019.

[8] Q. Ren and Q. Liang, “Energy and quality aware query processing in wireless sensor database systems,” *Information Sciences*, vol. 177, no. 10, pp. 2188–2205, 2007.

[9] K. Czuszyński, J. Rumiński, and A. Kwaśniewska, “Gesture recognition with the linear optical sensor and recurrent neural networks,” *IEEE Sensors Journal*, vol. 18, pp. 5429–5438, 2018.

[10] Z. Wang, Y. Hou, K. Jiang, W. Dou, C. Zhang, Z. Huang, and Y. Guo, “Hand gesture recognition based on active ultrasonic sensing of smartphone: A survey,” *IEEE Access*, vol. 7, pp. 111 897–111 922, 2019.

[11] B. Fang, Q. Lv, J. Shan, F. Sun, H. Liu, D. Guo, and Y. Zhao, “Dynamic gesture recognition using inertial sensors-based data gloves,” *2019 IEEE 4th International Conference on Advanced Robotics and Mechatronics (ICARM)*, pp. 390–395, 2019.

[12] Z. Liu, X. Liu, J. Zhang, and K. Li, “Opportunities and challenges of wireless human sensing for the smart iot world: A survey,” *IEEE Network*, vol. 33, pp. 104–110, 2019.

[13] J. Lien, N. Gillian, M. Karagozler, P. Amihood, C. Schwesig, E. Olson, H. Raja, and I. Poupyrev, “Soli: Ubiquitous gesture sensing with millimeter wave radar,” *ACM Transactions on Graphics*, vol. 35, pp. 1–19, 07 2016.

[14] Q. Ren and Q. Liang, “Fuzzy logic-optimized secure media access control (fsmac) protocol wireless sensor networks,” in *CIHSPS 2005. Proceedings of the 2005 IEEE International Conference on Computational Intelligence for Homeland Security and Personal Safety, 2005.* IEEE, 2005, pp. 37–43.

[15] P. Goswami, S. Rao, S. Bharadwaj, and A. Nguyen, “Real-time multi-gesture recognition using 77 ghz fmcw mimo single chip radar,” 01 2019, pp. 1–4.

[16] P. Molchanov, S. Gupta, K. Kim, and K. Pulli, “Short-range fmcw monopulse radar for hand-gesture sensing,” *IEEE National Radar Conference - Proceedings*, vol. 2015, pp. 1491–1496, 06 2015.

[17] Y. Li, T. Wang, A. Khan, L. Li, C. Li, Y. Yang, and L. Liu, “Hand gesture recognition and real-time game control based on a wearable band with 6-axis sensors,” *2018 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–6, 2018.

[18] M. Simão, P. Neto, and O. Gibaru, “Natural control of an industrial robot using hand gesture recognition with neural networks,” *IECON 2016 - 42nd Annual Conference of the IEEE Industrial Electronics Society*, pp. 5322–5327, 2016.

[19] S. Skaria, A. Al-Hourani, M. Lech, and R. J. Evans, “Hand-gesture recognition using two-antenna doppler radar with deep convolutional neural networks,” *IEEE Sensors Journal*, vol. 19, pp. 3041–3048, 2019.

[20] G. Li, R. Zhang, M. Ritchie, and H. D. Griffiths, “Sparsity-based dynamic hand gesture recognition using micro-doppler signatures,” *2017 IEEE Radar Conference (RadarConf)*, pp. 0928–0931, 2017.

[21] Y. Wang, S. Wang, M. Zhou, Q. Jiang, and Z. Tian, “Ts-i3d based hand gesture recognition method with radar sensor,” *IEEE Access*, vol. 7, pp. 22 902–22 913, 2019.

[22] Z. Zhou, Z. Cao, and Y. Pi, “Dynamic gesture recognition with a terahertz radar based on range profile sequences and doppler signatures,” *Sensors (Basel, Switzerland)*, vol. 18, 12 2017.

[23] A. Berenguer, M. Oveneke, H.-u.-R. Khalid, M. Alioscha-Perez, A. Bourdoux, and H. Sahli, “Gesturevlad: Combining unsupervised features representation and spatio-temporal aggregation for doppler-radar gesture recognition,” *IEEE Access*, vol. PP, pp. 1–1, 09 2019.

[24] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, “Netvlad: Cnn architecture for weakly supervised place recognition,” 06 2016, pp. 5297–5307.

[25] S. Hazra and A. Santra, “Short-range radar-based gesture recognition system using 3d cnn with triplet loss,” *IEEE Access*, vol. PP, pp. 1–1, 08 2019.

[26] Z. Zhang, Z. Tian, and M. Zhou, “Latern: Dynamic continuous hand gesture recognition using fmcw radar sensor,” *IEEE Sensors Journal*, vol. 18, pp. 3278–3289, 2018.

[27] Z. z. Xing, J. Pei, and P. Yu, “Early classification on time series,” *Knowledge and Information Systems - KAIS*, vol. 31, 04 2011.

[28] Z. Xing, J. Pei, P. S. Yu, and K. Wang, “Extracting interpretable features for early classification on time series,” in *SDM*, 2011.

[29] N. Parrish, H. S. Anderson, M. R. Gupta, and D.-Y. Hsiao, “Classifying with confidence from incomplete information,” *J. Mach. Learn. Res.*, vol. 14, pp. 3561–3589, 2013.

[30] M. S. Aliakbarian, F. S. Saleh, M. Salzmann, B. Fernando, L. Petersson, and L. Andersson, “Encouraging lstms to anticipate actions very early,” in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 280–289.

[31] M. Rußwurm, S. Lefèvre, N. Courty, R. Emonet, M. Körner, and R. Tavenard, “End-to-end learning for early classification of time series,” *ArXiv*, vol. abs/1901.10681, 2019.

[32] S. Ma, L. Sigal, and S. Sclaroff, “Learning activity progression in lstms for activity detection and early detection,” *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1942–1950, 2016.

[33] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *NIPS*, 2012.

[34] J. Chung, Çaglar Gülçehre, K. Cho, and Y. Bengio, “Empirical evaluation of gated recurrent neural networks on sequence modeling,” *ArXiv*, vol. abs/1412.3555, 2014.

[35] S. Wang, J. Song, J. Lien, I. Poupyrev, and O. Hilliges, “Interacting with soli: Exploring fine-grained dynamic gesture recognition in the radio-

frequency spectrum," in *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*. ACM, 2016, pp. 851–860.

[36] L. Wang, Z. Cao, Z. Cui, C. Cao, and Y. Pi, "Negative latency recognition method for fine-grained gestures based on terahertz radar," *IEEE Transactions on Geoscience and Remote Sensing*, pp. 1–14, 2020.

[37] U. Mori, A. Mendiburu, S. Dasgupta, and J. A. Lozano, "Early classification of time series by simultaneously optimizing the accuracy and earliness," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, pp. 4569–4578, 2018.

[38] V. Gupta, S. K. Dwivedi, R. Dabral, and A. Jain, "Progression modelling for online and early gesture detection," in *2019 International Conference on 3D Vision (3DV)*, 2019, pp. 289–297.

[39] B. Zhang, Y. Pi, and J. Li, "Terahertz imaging radar with inverse aperture synthesis techniques: System structure, signal processing, and experiment results," *IEEE Sensors Journal*, vol. 15, no. 1, pp. 290–299, 2015.

[40] V. Chen and S. Qian, "Cfar detection and extraction of unknown signal in noise with time-frequency gabor transform," *Proc SPIE*, pp. 285–294, 03 1996.

[41] J. Choi, S. Ryu, and J. Kim, "Short-range radar based real-time hand gesture recognition using lstm encoder," *IEEE Access*, vol. 7, pp. 33 610–33 618, 2019.

[42] C. Belloni, N. Aouf, J. L. Caillec, and T. Merlet, "Sar specific noise based data augmentation for deep learning," in *2019 International Radar Conference (RADAR)*, 2019, pp. 1–5.

[43] D. J. Moolayil, "Learn keras for deep neural networks," in *Apress*, 2019.

[44] A. Vani, R. N. Raajan, D. Haretha Winmalar., and R. Sudharsan, "Using the keras model for accurate and rapid gender identification through detection of facial features," in *2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC)*, 2020, pp. 572–574.

[45] D. Gavrila, "The visual analysis of human movement," *Computer Vision and Image Understanding - CVIU*, 01 1998.

[46] P. Molchanov, X. Yang, S. Gupta, K. Kim, S. Tyree, and J. Kautz, "Online detection and classification of dynamic hand gestures with recurrent 3d convolutional neural network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4207–4215.

[47] A. Jain, A. Singh, H. S. Koppula, S. Soh, and A. Saxena, "Recurrent neural networks for driver activity anticipation via sensory-fusion architecture," in *2016 IEEE International Conference on Robotics and Automation (ICRA)*, 2016, pp. 3118–3125.

Rui Min received the B.E., M.S., and Ph.D. degrees in circuits and systems from the University of Electronic Science and Technology of China (UESTC), Chengdu, China, in 2000, 2003, and 2012, respectively. He is currently a Professor with the School of Information and Communication Engineering, UESTC.

His research interests include radar signal processing and image processing.

Xing Wang received the B.S. degree in communication engineering from the Chongqing University of Posts and Telecommunications, Chongqing, China, in 2018. He is pursuing the M.S. degree with the School of Information and Communication Engineering, University of Electronic Science and Technology of China, Chengdu, China.

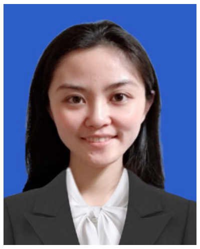
His research interests include high-resolution radar and gesture recognition.

Jie Zou received the bachelor's degree in communication engineering from Southwest Jiaotong University, Chengdu, China, in June 2007. He is currently an Associate Researcher with the Second Research Institute of Civil Aviation Administration of China.

His main research interests include civil aviation air traffic control technology and civil aviation radio technology.

Jing Gao received the master's degree in electronic and communication engineering from the Chengdu University of Technology, in June 2013. He is currently an Assistant Researcher with the Second Research Institute of Civil Aviation Administration of China.

His main research interest includes civil aviation radio technology.



Liying Wang received her B.S. degree from Northeast Petroleum University, Daqing, China. She is currently pursuing her Ph.D. in the School of Information and Communication Engineering, University of Electronic Science and Technology of China, Chengdu, China.

Her research interests include gesture recogntion and deep learning.

Zongjie Cao received the B.E. and Ph.D. degrees from the Xi'an Jiaotong University, Xi'an, China, in 1999, and 2005, respectively.

From 2006 to 2008, he was a postdoctoral researcher with the Communication and Information System Postdoctoral Center, University of Electronic Science and Technology of China (UESTC). In 2008, he joined the School of Electronic Engineering, UESTC. He is currently a professor with the School of Information and Communication Engineering, UESTC. He has authored or co-authored over 50 papers. His current research interests include radar signal processing, synthetic aperture radar imaging and target recognition.