

# Human Gesture Recognition Based on Millimeter-Wave Radar Using Improved C3D Convolutional Neural Network

Wei Li\*, Yang Gao, Jun Chen, Si-Yi Niu, Jia-Hao Jiang, Qi Li

North China University of Technology, Beijing, China

lwsar@ncut.edu.cn, gary\_young228@163.com, 648141353@qq.com, niusiyi\_xxsy@163.com,  
jjhwork@163.com, 1310135983@qq.com

*Received 28 May 2022; Revised 13 September 2022; Accepted 28 October 2022*

**Abstract.** In this paper, we propose a time sequential IC3D convolutional neural network approach for hand gesture recognition based on frequency modulated continuous wave (FMCW) radar. Firstly, the FMCW radar is used to collect the echoes of human hand gestures. A two-dimensional fast Fourier transform calculates the range and velocity information of hand gestures in each frame signal to construct the Range-Doppler heat map dataset of hand gestures. Then, we design an IC3D network for feature extraction and classification of the dynamic gesture heat map. Finally, the experiment results show that the gesture recognition system designed in this paper effectively solves the problems of the difficulty of human gesture feature extraction and low utilization of time series information, and the average recognition accuracy rate can reach more than 99.8%.

**Keywords:** millimeter wave radar, gesture recognition, deep learning

## 1 Introduction

With the advent of the era of artificial intelligence, human gesture recognition is attracting increasing attention due to its various applications, such as intelligent human-computer interactions [1], smart security [2], assisted driving [3], entertainment [4], smart home [5-6], competent medical care [7-8], and patient health monitoring [9]. Various human gesture recognition methods have been proposed, mainly classified as wearable and non-contact sensors.

In wearable sensors like those presented in [10], a three-axis accelerometer, a strong magnetometer, and a gyroscope are connected and worn on the wrist for human gesture recognition. This gesture recognition method can obtain real-time information about human gestures accurately and quickly; however, these sensors are expensive, cumbersome, and quickly forgotten, leading to information loss.

Human gesture detection using non-contact sensors divide into vision-based gesture recognition, infrared sensor-based gesture recognition, ultrasonic wave-based gesture recognition, and electromagnetic signal-based gesture recognition. Human gesture recognition based on visual sensors [11-13] does not require the user to wear heavy sensors, and it can quickly achieve human-computer interaction. However, it requires a high-resolution camera and specific working environment requirements. Furthermore, with the widespread use of the Internet, the public is increasingly aware of privacy protection; thus, using the mobile phone camera as a terminal sensor for motion recognition inevitably brings forth hidden dangers of privacy violation, limiting the implementation of this method. Infrared sensor-based gesture recognition does not recognize the infrared light emitted by the human hand. However, it uses a controlled intensity built-in infrared LED light source to illuminate the human hand and then recognizes the gesture by the infrared light reflected from the human hand. When an obstacle is in front of the sensor, the collected information is inaccurate, considerably influencing the recognition performance. Human gesture recognition based on wireless signal sensors most uses WIFI signals, such as in [14-16], which are low-cost and highly popular. These can achieve good gesture recognition in the case of poor light and do not harm user privacy; however, in this case, the multipath effect occurs when receiving the signal, which influences the accurate estimation of parameters.

Compared with sensors based on vision and kinematics, radar as a sensor for human gesture recognition has many advantages. It does not have specific requirements for the working environment, can be used in all weathers, will not harm personal privacy, and can realize recognition through walls.

With the development of monolithic microwave integrated circuit (MMIC) technology, millimeter-wave ra-

---

\* Corresponding Author

dar modules are also widely used in the market due to their high integration, low cost, low power consumption, and millimeter-wave radar-based gesture recognition methods are gradually becoming popular for research in academia and industry. For example, Google proposed various gesture recognition methods based on Soli radar [17-20], and Texas Instruments (TI) introduced a commercial product of radar that can be used for gesture recognition [21]. However, radar-based gesture recognition faces many challenges: 1) The acquisition of radar gesture data is difficult and time-consuming. It is essential to study models that use a small number of samples to solve the gesture recognition problem. 2) When the system receives an unprocessed radar data stream, the gesture should be detected and classified simultaneously to avoid a significant lag between gesture and classification in implementing gesture recognition. 3) Currently, most radar-based gesture recognition models focus on feature extraction and classification of two-dimensional feature images, applying the feature extraction algorithm to each gesture frame independently, ignoring the consistency information between frames. Moreover, the Doppler-time heat map lacks range information, and the range-Doppler heat map ignores the temporal continuity information between frames. Both temporal and spatial information is essential for human gesture recognition. 4) To overcome the challenges posed by inter-frame information of gestures, several algorithms for gesture feature extraction have been proposed successively. For example, the multichannel architecture [22] uses multiple independent convolutional neural networks to extract gesture features. However, this approach achieves high performance at the cost of complexity. Then, a temporal 3DCNN network [23] is designed to process the sequenced 2D images emphasizing their temporal relation. However, due to the high requirements of gesture recognition on classification network complexity and computational speed, it is crucial to design a generic, compact, efficient, and lightweight network architecture.

To solve the above problem, we propose an improved C3D network to address the FMCW millimeter wave radar-based gesture recognition problem. The main contributions of this paper are as follows.

1) IC3D network: we propose a new IC3D network (improved 3D convolutional neural network, IC3D) for feature extraction of dynamic gesture heat map in response to a large number of network parameters of the original C3D convolutional neural network [24] and the problem of improving the gesture recognition rate while compressing the network parameters. We reduce the number of parameters in the convolutional and fully connected layers, replace the original ReLU activation function with a smoother Mish activation function, and replace the traditional SGD algorithm with an Adam optimization algorithm with more stable parameter variations. Finally, the extracted gesture features are spliced and classified in the softmax layer.

2) Gesture collection and processing: The FMCW radar is used to collect human hand gestures. A two-dimensional fast Fourier transform calculates hand gestures' range and velocity information based on the IF signal frequency. The Range-Doppler heat map is generated according to the relationship between range, speed, and the frequency of the IF signal. Finally, we use a multi-frame distance Doppler heat map time series to represent each gesture. The hand gesture data are repeatedly collected and processed to generate a dataset.

3) Experimental validation: We repeatedly collected seven gestures, 200 times each, and divided the data set into training and validation sets in the ratio of 7:3 to validate the effectiveness of the proposed IC3D network and discussed the effect of network parameters on the model. The experimental results show that the average recognition accuracy of the proposed IC3D for each gesture is 99.8%, the computational delay is less than 40 ms, and it has good generalization ability.

The remaining sections of this paper are organized as follows. In Section II, we discuss the related works. Section III introduces the composition of our human gesture recognition system using FMCW radar. Section IV is devoted to pre-processing the baseband signal and generating RDIs. Section V describes the architecture of our proposed IC3D convolutional neural network. Section VI, we perform a parametric optimization of IC3D, demonstrate our results through many experiments, and compare them with other state-of-the-art models. Finally, section VII gives the conclusions, limitations of the work, and future work.

## 2 Related Work

The existing radar sensor-based gesture recognition methods mainly include gesture signal pre-processing, feature extraction, and classification of gestures. This section describes the related works on radar sensor-based gesture recognition.

## 2.1 Gesture Signal Preprocessing

Radar gesture signal preprocessing is mainly to preprocess the received radar gesture echoes, remove the background noise and other interference information from the radar echoes except dynamic gesture information, and extract the practical gesture information. The standard gesture signal preprocessing method is the classical Fourier transform and its derivative algorithm.

Kim et al. [25] obtained the micro-Doppler spectrum of the gesture signal by STFT analysis. They then used the micro-Doppler spectrum image as an input to study the recognition of the measured ten gestures using convolutional neural networks. Although the idea of short-time Fourier transform localization has achieved some achievements in gesture information preprocessing, the fixed invariance of the sliding window function and poor adaptivity drawbacks still exist. Given this, Khaled's team at Carnegie Mellon University [26] used the received signal strength indication information (RSSI) to identify dynamic gestures by wavelet transforming the rising edge, falling edge, and pulse features of the gesture signal, making the gesture recognition system with better adaptivity. LeiWentai [27] et al. used a two-dimensional FFT method with a window function to obtain the RDM, designed an improved wavelet threshold function to remove the noise in the RDM, and used a CA-CFAR detector to suppress the clutter, which better removed the noise and clutter from the gesture radar signal. The experimental results show that the method can better remove the noise and suppress the clutter, which reduces the difficulty of subsequent recognition and helps to identify different hand gestures better. Dekker [28] et al. processed the one-dimensional gesture signal as a Doppler-time spectrum and used the real and imaginary parts of this spectrum as the input to a convolutional neural network, which in turn accomplished the classification and recognition of gestures.

In this paper, the FMCW millimeter wave radar sensor is used to receive the gesture data, and the distance and velocity of the gesture in each frame of the signal are calculated by a two-dimensional fast Fourier to transform to generate a distance-Doppler map for subsequent feature extraction as well as classification recognition.

## 2.2 Gesture Feature Extraction and Classification Recognition Algorithm

After pre-processing, the gesture radar data includes not only human gesture information but also other redundant information, which will increase the difficulty of model recognition and classification and reduce the accuracy and speed of recognition, so the means of feature extraction is taken to remove the redundant information.

Currently, the more mainstream classification recognition algorithms mainly include radar gesture recognition based on template matching, radar gesture recognition based on statistical learning, and radar gesture recognition based on deep learning.

**A Template Matching Based Radar Gesture Recognition Algorithm.** Dynamic time warping (DTW) [29-31] is currently the most common template matching algorithm for radar gesture recognition. A reference template set needs to be constructed first when using DTW to process radar data. The gesture data with the slightest difference is calculated as the output result by comparing the similarity between the test data and the reference template. However, the DTW algorithm also has limitations such as high computational complexity and poor stability. Especially in the case of more complex gesture movements and many training samples, the recognition rate will be significantly reduced. Therefore, the focus of gesture feature extraction and classification recognition algorithms has been put on machine learning and deep learning algorithms.

**B Statistical Learning-based Radar Gesture Recognition Algorithm.** Statistical machine learning algorithms construct models for known data to predict and analyze unknown data. Hidden Markov model (HMM), etc.

Liu Zhao et al. [32] used a segmented FFT algorithm to transform radar gesture echoes into two-dimensional images characterizing gesture features and joint SVM to train and classify two-dimensional gesture feature quantities, and its accuracy rate reached 90.25%. Although SVM can effectively solve the small-sample, high-dimensional, nonlinear problem, it is less efficient when the number of training samples is large. Sun et al. [33] extracted the micro-Doppler information of five gestures using 77 GHz FMCW radar. Finally, a KNN classifier was used to classify and recognize these five features, and after experiments, it was shown that the proposed model could guarantee a high recognition accuracy. Although the KNN algorithm is simple and easy to understand, it requires a large amount of space storage and has high time complexity.

Although the statistical machine learning-based algorithm can effectively improve the recognition accuracy, the computation is too large in algorithm execution, which significantly limits the recognition speed.

**C Deep Learning-based Radar Gesture Recognition Algorithm.** Deep learning is an essential branch of machine learning, and its effectiveness in image recognition far exceeds that of previous related technologies. The main algorithms currently applied to radar gesture recognition are recurrent neural networks (RNN) and convolutional neural networks (CNN).

A recurrent neural network (RNN) is a class of neural networks with short-term memory capability. J. Choi [34] successfully recognized ten gestures with 98.48% accuracy using long short-term memory (LSTM) network. However, this method performs best on small data sets, and the computational efficiency decreases significantly for a more significant number of training samples.

CNN has shown a more significant advantage in radar gesture recognition due to its automatic feature extraction and weight-sharing advantages. Karpathy et al. [35] fused the features of each image frame of the gesture and then used CNN to classify the fused gesture features. However, using CNN to extract radar gesture features is not only data-demanding and computationally expensive. Also, the 2D CNN cannot be directly used for 3D data analysis to extract temporal consistency information between frames, making this gesture frames irrelevant to static images. Ji et al. [36] addressed the inability of CNN to extract temporal information of motion, added temporal dimension to the original structure, and proposed 3DCNN for motion recognition. Such a 3D-CNN structure is widely used in various types of behavior recognition. Guiyuan Zhang et al. [37] constructed a 3D-CNN deep neural network to recognize and classify the distance Doppler heat map of hand gestures and experimentally demonstrated that the network achieved 95% accuracy for six common hand gestures. Zhenyuan Zhang et al. [38] combined the distance Doppler information and distance-angle information through a dual-stream recurrent three-dimensional convolutional neural network (R3DCNN) for extraction and fusion and finally achieved a high recognition rate. However, in this scheme, the recognition and classification accuracy of gestures is reduced after the fusion of distance and angle parameters with Doppler parameters for features. The multi-branch structure and complex learning process increase the complexity of the network and reduce the computational speed.

### 3 Composition of the Gesture Recognition System

The proposed human gesture recognition system based on FMCW millimeter-wave radar comprises a personal computer (PC) and millimeter-wave radar, forming a non-contact sensing device. The system is shown in Fig. 1. It includes a collection of human gestures, time-frequency analyses of radar echoes, and gesture feature extraction. Further, we design a 3D CNN to recognize time-series heat maps. The 3D CNN is installed on the PC and can detect human gestures of the test objects in real time after data training.

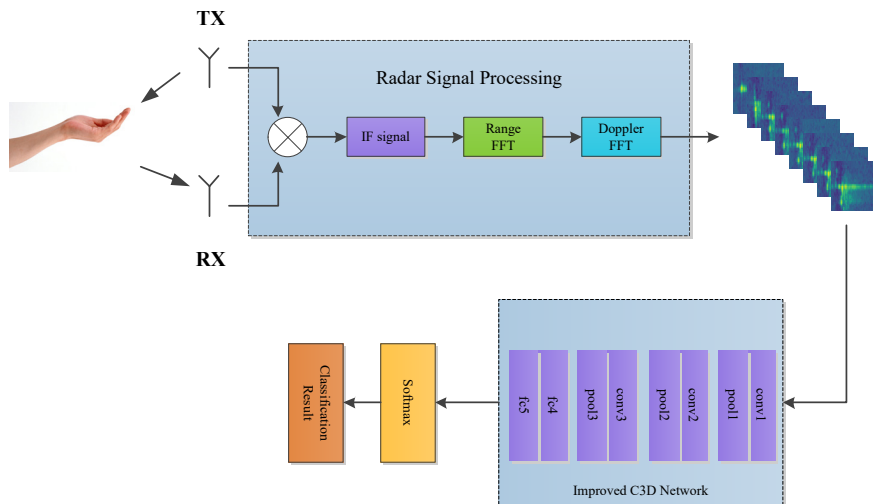


Fig. 1. Human gesture recognition system

## 4 Principle Analysis of FMCW Millimeter-wave Radar

### 4.1 Radar Signal Preprocessing

In the FMCW millimeter-wave radar, the modulation methods typically comprise triangular and sawtooth waves. Herein, the millimeter-wave radar adopted the commonly uses sawtooth wave modulation. After the transmitted signal of the radar encounters the target, it is reflected and reached the receiving antenna after a time delay expressed as  $\Delta t_d$ . The transmitted signal is combined with the echo signal to obtain an intermediate frequency (IF) signal with a constant frequency. Fig. 2 shows the working principle of the FMCW millimeter-wave radar;  $B$  indicates signal bandwidth,  $f_c$  indicates the carrier's center frequency,  $T_c$  indicates the chirp pulse width, and  $S$  indicates the slope of the chirp. The frequency of the transmitted signal increases linearly with time  $\tau$ , which is given as  $f_T(\tau) = S \cdot \tau$ ,  $S = B / T_c$ .  $A_T$  and  $A_R$  represent the amplitudes of the transmitted and received signals.

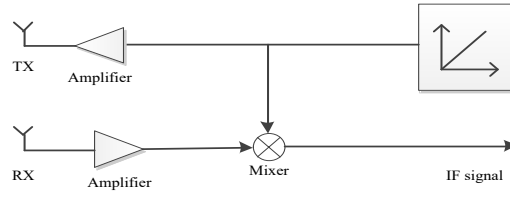


Fig. 2. Schematic representation of the working principle of the radar

The chirp signal transmitted by FMCW radar is expressed using Eq. (1):

$$S_T(t) = A_T \cos[2\pi(f_c t + \int_0^t f_T(\tau) d\tau)] \quad (0 \leq t \leq T_c) . \quad (1)$$

The signal is reflected when it encounters the target and is received by the receiving antenna. The received signal is expressed using Eq. (2):

$$S_R(t) = A_R \cos\{2\pi[f_c(t - \Delta t_d) + \int_0^{t - \Delta t_d} [S(t - \Delta t_d) + \Delta f_d] d\tau]\} . \quad (2)$$

Where  $\Delta f_d = -\frac{2f_c v}{c}$  represents the Doppler shift due to object motion. The received signal  $S_R(t)$  mix with the transmitted signal  $S_T(t)$ , and this mixed signal is passed through a low-pass filter to obtain the IF signal:

$$S_{IF}(t) = S_R(t) \cdot S_T(t) = \frac{1}{2} A_R A_T \cos\{2\pi[(f_c \cdot \Delta t_d) + (S \cdot \Delta t_d - \Delta f_d)t]\} . \quad (3)$$

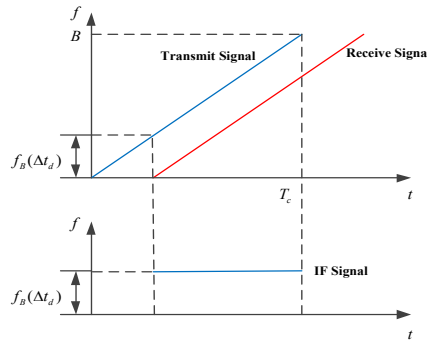


Fig. 3. IF signal formation

Fig. 3 illustrates the principle of IF signal generation.

The distance and velocity of the target is determined via time-frequency analysis on the IF signal.

The frequency of the IF signal generated by the target is different at different distances. The distance to the target is calculated by analyzing and calculating the frequency  $f_B(\Delta t_d)$  of the IF signal. FFT is used on the IF time domain signal to determine the amplitude corresponding to different frequencies, and the distance to the target is calculated using Eq. (4):

$$R(\Delta t_d) = \frac{c \cdot T_c \cdot f_B(\Delta t_d)}{2B} . \quad (4)$$

Where  $c$  is the speed of light,  $T_c$  is the pulse width and duration, and  $f_B(\Delta t_d)$  is the frequency of the IF signal at this moment.

The velocity of the target is estimated according to the accumulation of multiple pulse sweeps. Different sweep pulse signals at the same distance have different phases. The relationship between the phase difference and velocity between two sweep pulses is given using Eq. (5). Where  $\Delta w$  is the phase difference.

$$v = \frac{\lambda \Delta w}{4\pi T_c} . \quad (5)$$

## 4.2 Radar Parameter Settings

Herein, we use IWR1642BOOST and DCA1000EVM millimeter-wave radars from Texas Instruments to collect the echo data of human gestures. IWR1642BOOST operates in the 76–81 GHz frequency band with continuous chirps up to 4 GHz. It is a monolithic implementation comprising a 2TX-4RX system with a built-in PLL, a digital-to-analog converter, and a DSP subsystem. The proposed human gesture recognition system adopts the 1TX-1RX antenna mode. Fig. 4 illustrates the chirp signal transmitted by IWR1642BOOST.

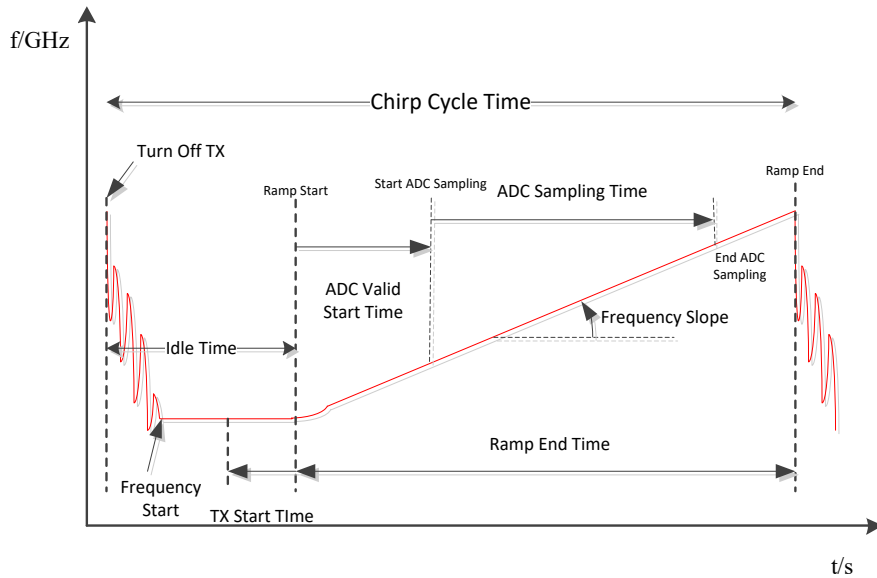


Fig. 4. Chirp signal waveform

In the radar parameter setting, to ensure the integrity of human gesture radar information extraction, it is necessary to consider the range resolution,  $R_{res}$ , velocity resolution,  $v_{res}$ , and the maximum measurable velocity  $v_{max}$ . These parameters are expressed using Eqs. (6)–(8), where  $N_c$  is the number of chirps transmitted in one frame:

$$R_{res} = \frac{c}{2B} . \quad (6)$$

$$v_{max} = \frac{\lambda}{4T_c} . \quad (7)$$

$$v_{res} = \frac{\lambda}{2N_c T_c} . \quad (8)$$

As shown in these equations, the bandwidth of the radar signal should be sufficiently large to have a small range resolution. The speed resolution depends on the frame period, which should satisfy the tiny gestures that can be collected. In the parameter setting of sampling time, it should be taken into account that a short sampling time will lead to inaccurate human gesture collection, and a very long sampling time will lead to redundant information in addition to gesture information. The complete gesture action echo collected by radar is divided into eight frames to get the radar data time series of human gestures. Table 1 presents the parameter settings of the chirp. Table 2 presents the data sampling parameter settings.

**Table 1.** Chirp parameter settings

Parameter	Value
Start frequency (GHz)	77.000
Frequency slope (MHz/ $\mu$ s)	49.970
Idle time ( $\mu$ s)	20.00
ADC start time ( $\mu$ s)	6.00
Ramp end time ( $\mu$ s)	80.00
RX gain (dB)	30

**Table 2.** Sampling parameter settings

Parameter	Value
ADC samples	256
Sample rate (ksps)	10000
Number of chirps every frame	128
Periodicity (ms)	112
Number of frames	8

### 4.3 Principle of Gesture Imaging

The millimeter-wave radar measures the distance by detecting the IF frequency and the velocity by detecting the Doppler frequency shift of the measured target. The binary echo data collected by the radar arrange in a two-dimensional matrix shown in Fig. 5. As described in Table 2, there are 128 chirp signals in one frame, and the number of ADC sampling points for each chirp signal is 256. The dimension of the radar data matrix for one frame is thus  $256 \times 128$ .

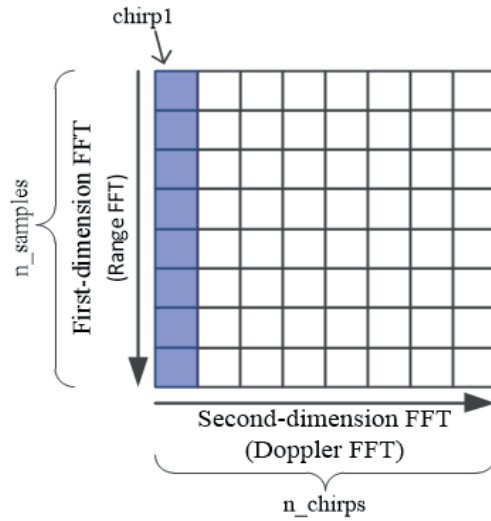


Fig. 5. Radar data matrix

The 256 ADC samples of each chirp process with a windowed FFT called distance dimension FFT. Then, another FFT performs on the result of distance dimension FFT in the chirp dimension, called velocity dimension FFT. Then a distance-Doppler heat map of one data frame is obtained, as shown in Fig. 6.

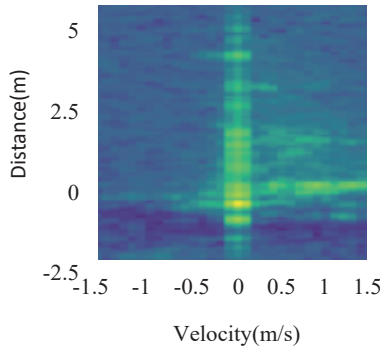
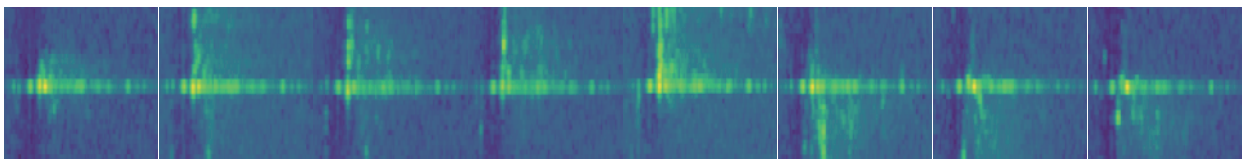


Fig. 6. The range-Doppler heat map of one frame

The 8-frame distance-Doppler heat maps of a sweeping hand gesture is obtained by frame-by-frame processing; Table 3 shows the 8-frame heat map of a hand wave. The horizontal coordinate is the distance, and the vertical coordinate is the velocity. In this study, the person in front of the radar is also taken into account. The distance-invariant horizontal line is considered in the distance-Doppler heat map, where the upper part of the horizontal line is the negative motion velocity characteristic toward the radar, and the lower part of the horizontal line is the positive motion velocity characteristic backward from the radar.

Table 3. The range-Doppler heat map sequence of waving





## 5 Convolutional Neural Network

### 5.1 C3D Convolutional Neural Network

In 2010, to overcome the inability of two-dimensional neural networks for the time-series information of motion extraction, Ji et al. added the time dimension to the original structure and proposed 3DCNN for motion recognition. In 2015, Du et al. [24] improved the existing 3D CNN and proposed the C3D CNN. The original C3D CNN is presented in Fig. 7; the network comprises eight convolutional layers, five pooling layers, two fully connected layers, and one softmax classification layer. The number of convolution kernels in each layer was “64”, “128”, “256, 256”, “512, 512”, and “512, 512”. Appropriate padding was set in space and time dimensions, and the stride was set to one. Based on the results of several experiments, a  $3 \times 3 \times 3$  convolution kernel was considered optimal. All pooling layers used the maximum pooling method. The kernel size of the first layer of the pooling layer was  $1 \times 2 \times 2$ , and the stride was  $1 \times 2 \times 2$ , where one means depth. The kernel size and stride of the remaining maximum pooling layers were  $2 \times 2 \times 2$  each.



Fig. 7. Structure and size of the original C3D CNN

### 5.2 IC3D Convolutional Neural Network

Owing to the large number of parameters of the original C3D CNN, it is necessary to further improve the recognition rate of human gestures in the radar data set while decreasing the number of network parameters. The improvement strategies are as follows, (1) first, the number of convolution and pooling layers is reduced; (2) then, the number of input parameters is decreased so that the original input decreased from 16 frames of  $112 \times 112$  images to 8 frames of  $40 \times 40$ ; (3) in order to reduce the network parameters and to solve the problem that the stochastic gradient descent (SGD) optimization algorithm tends to introduce more random noise, another improved gradient descent algorithm, the Adam algorithm, is used in this paper; (4) The ReLU activation function causes “necrosis” of neurons, and eventually, the corresponding parameters are never updated. We optimize the ReLU activation function to a smoother Mish activation function.

**A Structure of the IC3D CNN.** In this study, the input size of the network is  $12 \times 3 \times 8 \times 40 \times 40$ , which can be written in the general form as  $N \times C_{in} \times D_{in} \times H_{in} \times W_{in}$ , where  $N$  is the batch size,  $D_{in}$  is the number of input frames, and  $C_{in}$ ,  $H_{in}$ , and  $W_{in}$  are the number of channels, height, and width, respectively, of the input image. The formula for convolution is expressed using Eq. (9):

$$H = \frac{H + 2 \times padding - kernelsize}{stride} + 1 . \quad (9)$$

The final network architecture comprises three convolutional layers, three pooling layers, two fully connected layers, and one softmax layer. The kernel size of the convolution is  $3 \times 3 \times 3$  in all cases, and the stride is  $1 \times 1 \times 1$ . To not prematurely reduce the length in the time dimension, the first pooling layer has its size and step size set to  $1 \times 2 \times 2$ . It ensures that the images’ input and output sizes before and after convolution are the same. The kernel size of the pooling layer in the second layer and the stride are  $2 \times 2 \times 2$ . The kernel size and stride of the third pooling layer are both  $4 \times 4 \times 4$ . Table 4 presents the network structure.

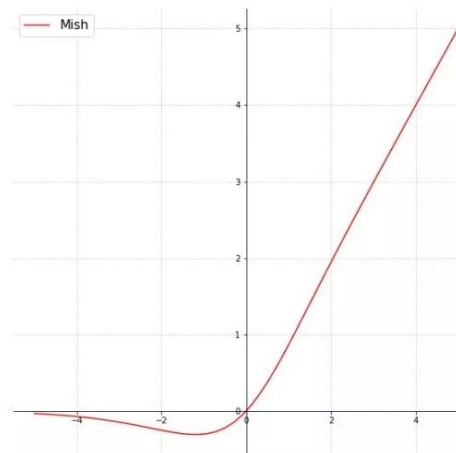
**B Optimization Algorithm.** In the original network, SGD is used as the optimization algorithm. SGD randomly optimizes the loss function on a certain piece of training data in each round of iteration, and it increases the parameter update speed of each round; however, it also introduces random noise, decreasing the training accu-

racy and possibly converging to a local optimum. To mitigate the above problems, in this study, the Adam algorithm [39] is adopted, which is an improved SGD method. The Adam algorithm combines the advantages of the AdaGrad and RMSProp algorithms, which can iteratively updates neural network weights based on training data.

**Table 4.** Structure of the IC3D CNN

Module	Layer name	Filter size	Input size	Input channels	Output channels
Layer 1 Convolution-pooling module	conv1 pool1	$3 \times 3 \times 3$ $1 \times 2 \times 2$	$40 \times 40$	3	64
Layer 2 Convolution-pooling module	conv2 pool2	$3 \times 3 \times 3$ $2 \times 2 \times 2$	$20 \times 20$	64	128
Layer 3 Convolution-pooling module	conv3 pool3	$3 \times 3 \times 3$ $4 \times 4 \times 4$	$10 \times 10$	128	256
Layer 4 Fully connected layer	fc4	-	$3 \times 3$	256	2304
Layer 5 Fully connected layer	fc5	-	$2304 \times 1$	1	4096
Layer 6 Softmax layer	softmax	-	$4096 \times 1$	1	7

**C Activation Functions.** The activation function used in the original C3D network is the rectified linear unit (ReLU), which is simple to compute and solves the problem of gradient disappearance in the backpropagation process. However, ReLU does not adapt to large gradient inputs during training, which can cause some neurons to “die”, and the corresponding parameters cannot be updated. It is not derivable and smooth at the origin, thus leading to some unexpected problems during gradient optimization. Therefore, a new activation function, namely the Mish activation function, is used in this paper. This activation function does not have an upper bound, so it does not have a zero gradient due to saturation, which allows for better parameter updating. Thus, the recognition accuracy of the model can increase. The Mish activation function has a lower bound and a smaller weight in the negative half-axis to prevent neuron necrosis in the ReLU function. The Mish activation function is a smooth function with a better generalization ability and practical optimization ability of the results, which can improve the quality of the results. Fig. 8 shows the image of Mish’s function.

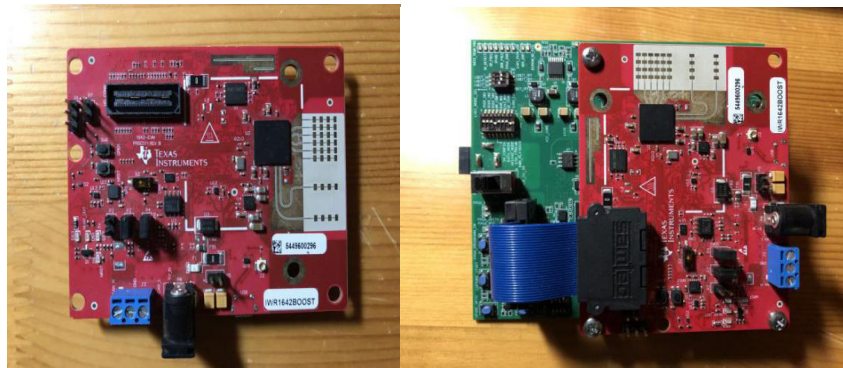


**Fig. 8.** Mish activation functions

## 6 Experimental Verification and Comparative Analysis

### 6.1 Dataset Construction

During the experiments, seven types of gestures are considered: still, push, pull, wave, circle, cross, and tick. Data of a single action contained eight distance-Doppler heat maps, and each action was collected 200 times, making a total of 1400 samples. We got six people to collect the data to improve the generalizability of the data. Moreover, we use the different test backgrounds when the data are collected, and some static objects are placed around as a distraction. Fig. 9 shows images of IWR1642BOOST during the experiment.



(a) IWR1642BOOST

(b) DCA1000EVM data acquisition kit

**Fig. 9.** Radar platform used for experiments

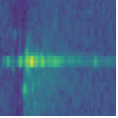
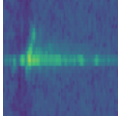
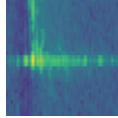
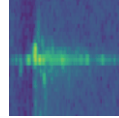
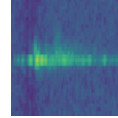
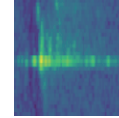
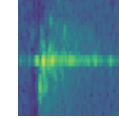
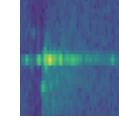
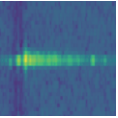
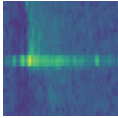
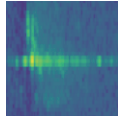
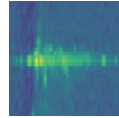
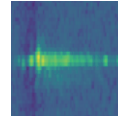
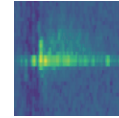
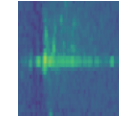
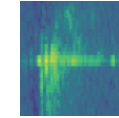
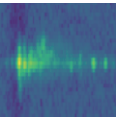
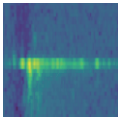
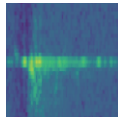
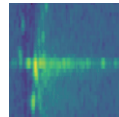
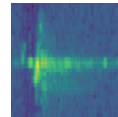
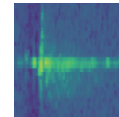
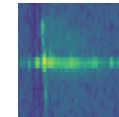
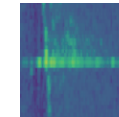
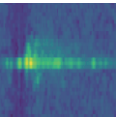
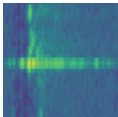
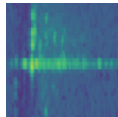
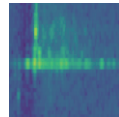
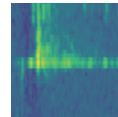
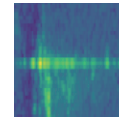
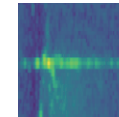
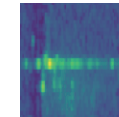
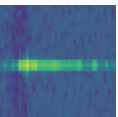
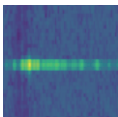
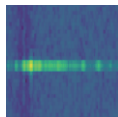
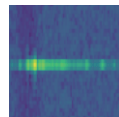
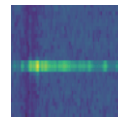
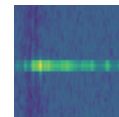
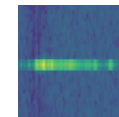
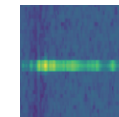
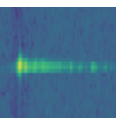
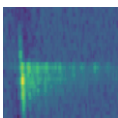
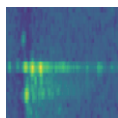
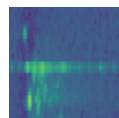
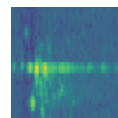
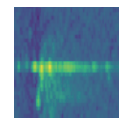
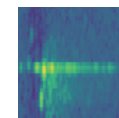
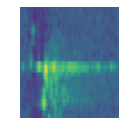
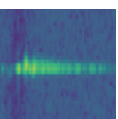
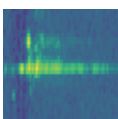
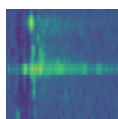
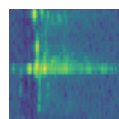
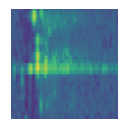
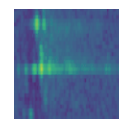
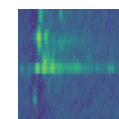
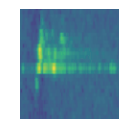
The IWR1642BOOST device supports a 77–81 GHz maximum frequency modulation bandwidth of 4 GHz; the horizontal radar field of view is set to  $\pm 60^\circ$ , and the elevation field of view is set to  $22^\circ$ . The radar erection height is set to one meter to achieve sufficient detection area. Fig. 10 presents the experimental setup.



**Fig. 10.** Experimental setup

Table 5 shows the captured 8-frame range-Doppler time-series heat maps for seven types of gestures.

**Table 5.** Range-Doppler heat map sequence

Gesture	8-Frame range-Doppler time-series heat maps							
Cross								
Tick								
Circle								
Wave								
Still								
Pull								
Push								

## 6.2 Parameter Optimization

**A Impact of the Training to Testing Dataset Ratio.** We conducted experiments using different training to test dataset ratios, with the ratios set to 3:7, 5:5, and 7:3. Fig. 11 gives the results of accuracy varying with the number of the epoch. From the figure, we can see that IC3D has a poor fitting ability of the network, and the recognition accuracy decreases due to the small training dataset. When increasing the ratio of the training set to 7:3, the network generalization ability of IC3D is higher, and the recognition accuracy is also higher. In the following experiments, we conduct experiments with the ratio of 7:3.

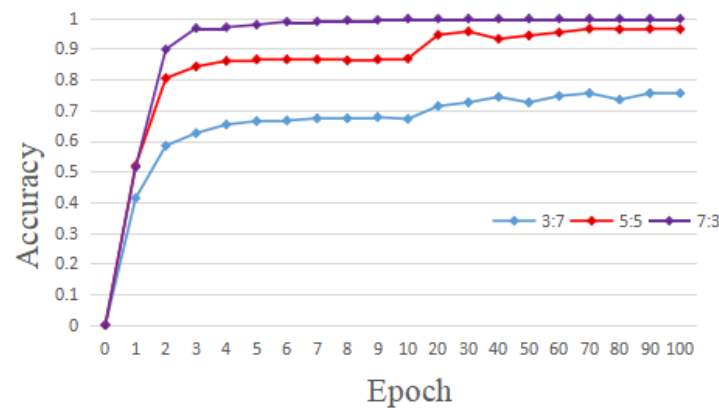


Fig. 11. Accuracy of different dataset proportions

**B Impact of Learning Rate.** Learning rate is an essential hyperparameter in deep learning; it determines whether the target can converge and when it converges to the minimum value. For a very high learning rate, the magnitude of the parameter update will be significant, which would fail to connect and increase the error. For a minimum learning rate, the error convergence speed will be plodding, which would lead to a local minimum and failure to obtain an ideal network model. Herein, different learning rates, namely, 0.001, 0.003, 0.009, 0.0001, 0.0003, and 0.0009, are selected for comparison. Fig. 12 presents the accuracy curve on the validation set without a learning rate.

**C Impact of Fully Connected Layer.** The size of the fully connected layer is also crucial for training the model. In a CNN, a fully connected layer plays the role of a “classifier.” A weighted summation performs on the features of the previous layer, and the feature space is mapped to the sample label space through a linear transformation. A single fully connected layer has many neurons, which may easily cause overfitting, increase the operation time, and decrease efficiency. However, if the size of the fully connected layer is too small, it isn’t easy to effectively fit the ideal model. In this study, the layer size of fc4 is set to 2304 and 4096, and that of fc5 is set to 2304, 4096, and 8192. Fig. 13 presents the experimental results.

Fig. 12 shows that the training accuracy is the highest with an initial learning rate of 0.003. Fig. 13 shows that for a constant width of the fc5 layer, a smaller width for fc4 led to better accuracy. For a constant width of the fc4 layer, training accuracy increases with increasing width of the fc5 layer. Thus, selecting an optimal large-size fully connected layer can enhance the fitting ability of the network. For fc5 layer sizes of 4096 and 8192, the training accuracies are slightly similar; thus, the sizes of the fully connected layer in this paper are selected as fc4 (2304, 2304) and fc5 (2304, 4096). By testing different learning rates and different fully connected layer sizes, the accuracy of the optimal combined structure and the loss are shown in Fig. 14 and Fig. 15.

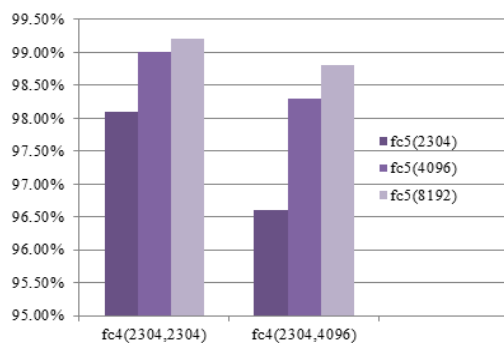


Fig. 12. Different learning rates

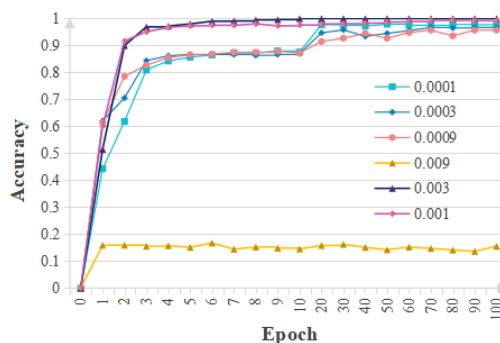


Fig. 13. Different fully connected layer widths

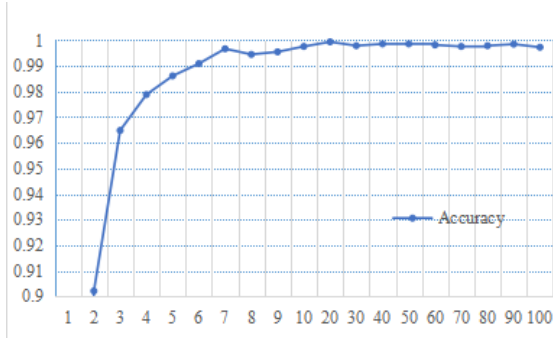


Fig. 14. The training accuracy curve of the network

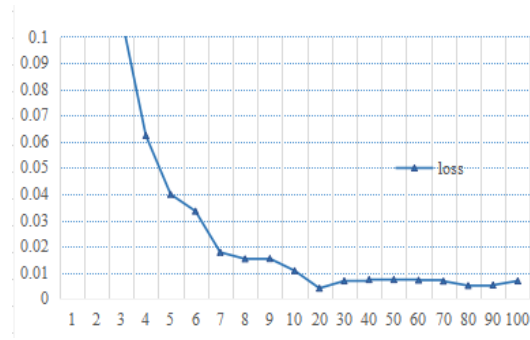


Fig. 15. The training error curve of the network

### 6.3 Performance Comparison with Different Radar-Based Approaches

To further verify the effectiveness of the IC3D network, we select some behavioral recognition deep learning networks for comparison: Dual-channel CNN [40], RDA-T [41], VGGNet [42], LRACN [19], 3DCNN+LSTM, CNN+LSTM [43], C3D. CNN uses five layers of convolutional networks; the convolutional kernel size is 3x3. The number of CNN kernels in each layer is 64, 128, 256, 512, and 512, respectively. 3DCNN and C3D kernels are set to 3x3x3, the number of training epochs is set to 100, and the initial learning rate is set to 0.003.

We compared IC3D with other gesture recognition methods regarding action type, data volume, time complexity, space complexity, number of iterations, and accuracy. Table 6 shows the comparison results.

**Table 6.** Comparing the results of different models

Network structure	Action type	Amount of data	Accuracy (%)	Capture frames	Time complexity ( $10^9$ FLOPs)	Space complexity ( $10^6$ byte)
Dual-channel CNN	7	2240	99	-	2.01	22.4
RDA-T	6	5600	95.3	32	2.11	89.6
VGGNet	8	1400	96.32	8	15.5	136
LRACN	5	2000	93.75	100	0.87	-
CNN+LSTM	8	1200	94.75	100	2.15	-
3DCNN+LSTM	8	-	97.17	32	30.83	-
C3D	7	1400	98.25	16	34.75	90.4
Proposed method	2	400	100	-	-	-
	5	1000	99.92	8	1.15	85.9
	7	1400	99.8	-	-	-

The proposed IC3D method requires only approximately 350 iterations to converge. In comparison, other network require >500 iterations to converge owing to their complex network structures, numerous network parameters, and a large amount of data. Compared with the traditional spectral input network model (dual-channel CNN), the proposed C3D model has approximately 42.78% lower time complexity. Compared with RDA-T, the proposed model has more downtime and space complexity. Compared with the traditional VGGNet, the proposed model reduces the time complexity to 1/10 and the space complexity to 3/5. Compared with LRACN, the proposed model has a 6.05% higher accuracy; however, it also has higher time complexity. In this study, two, five, and seven types of gestures are trained and recognized, and high recognition accuracies are obtained. The comparison shows that the proposed model has better generalizability for multitype action detection. The processing time of a single sample is *ms*, providing theoretical insights for real-time gesture recognition applications.

After the model training process is completed, the test dataset is input to each network for gesture recognition. Fig. 16 shows the recognition accuracy of each gesture on our self-built dataset. The figure shows that the VGGNet, 3D-CNN, and C3D networks are slower to fit due to their more complex structures, but the final recognition accuracy is higher than that of the RDA-T, LRACH, and CNN+LSTM models. The IC3D model network is faster to fit due to its simple structure and higher recognition accuracy than the other networks.

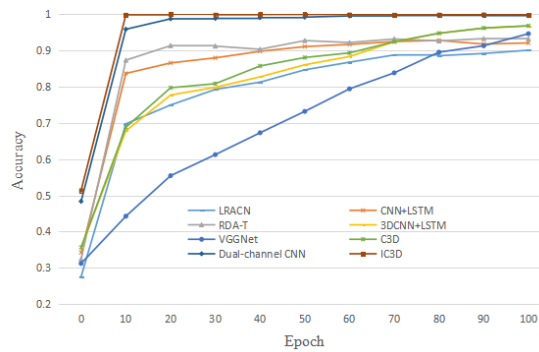


Fig. 16. Training process comparison of each networks.

To verify the online recognition capability of IC3D, 700 groups of data that are not used for training are selected for evaluation. The gesture classification results and the confusion matrix are presented in Fig. 17.

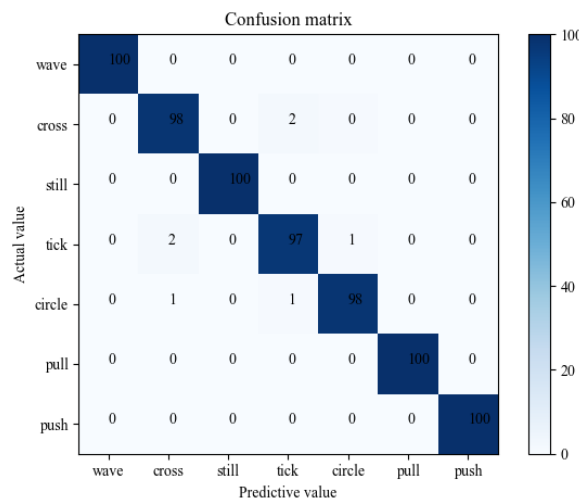


Fig. 17. Gesture classification confusion matrix

The confusion matrix shows cross, tick, and circle have strong coupling. Due to similar movement speeds, their features are not evident and are challenging to identify. Waving, pushing, pulling, and standing have a relatively small misjudgment due to apparent differences in motion characteristics.

## 7 Conclusion and Future Work

In this study, a dynamic human gesture recognition method based on millimeter-wave radar is proposed; it works effectively in cases of privacy protection and in dark environments. An improved C3D convolutional neural network is proposed for the dynamic sequence characteristics of human hand gestures. We performed parameter optimization through several experiments, which better solved the problem of many parameters of the original network, reduced the phenomenon of overfitting of the neural network, and increased the generalization ability of the neural network. A radar data set of human gestures is established, and the neural network is trained using this data set and evaluated using the validation set, achieving high accuracy of 99.8%. Thus, the proposed human gesture recognition system can improve detection accuracy, reduce computational delay, and have good gesture detection generalizability.

The experimental scenario of the current study is relatively homogeneous compared to the actual complex application environment. There are few interfering objects for gesture detection using millimeter wave radar. The background environment is ideal. In future work, we plan to add more complex experimental scenarios or artificially add interference information that can be used to train a general and more robust learning model. To ensure that the gesture recognition system can maintain a high recognition speed and accuracy in complex environments.

## 8 Acknowledgement

This work was supported by Natural Science Foundation of Beijing Municipality, Grant/Award Number: 4202019. The authors would like to thank the reviewers for the careful review and valuable suggestions.

## References

- [1] H. Yang, A. Park, S. Lee, Gesture Spotting and Recognition for Human-Robot Interaction, *IEEE Transactions on Robotics* 23(2)(2007) 256-270.
- [2] S.-J. Mambou, O. Krejcar, P. Maresova, A. Selamat, K. Kuca, Novel Hand Gesture Alert System, *Applied Sciences* 9(16)(2019) 3419.
- [3] A. Mishra, J. Kim, J. Cha, D. Kim, S. Kim, Authorized Traffic Controller Hand Gesture Recognition for Situation-Aware Autonomous Driving, *Sensors* 21(23)(2021) 7914.
- [4] J. Kim, H. Jung, M. Kang, K. Chung, 3D Human-Gesture Interface for Fighting Games Using Motion Recognition Sensor, *Wireless Personal Communications* 89(3)(2016) 927-940.
- [5] P. Nallabolu, Z. Li, H. Hong, C. Li, Human Presence Sensing and Gesture Recognition for Smart Home Applications With Moving and Stationary Clutter Suppression Using a 60-GHz Digital Beamforming FMCW Radar, *IEEE Access* 9(2021) 72857-72866.
- [6] Y. Lou, W. Wu, R. Vatavu, W.-T. Tsai, Personalized gesture interactions for cyber-physical smart-home environments, *Science China Information Sciences* 60(7)(2017) 1-15.
- [7] J. Cifuentes, M.-T. Pham, R. Moreau, P. Boulanger, F.-A. Prieto, Medical gesture recognition using dynamic arc length warping, *Biomedical signal processing and control* 52(2019) 162-170.
- [8] N.-M. Mahmoud, H. Fouad, A.-M. Soliman, Smart healthcare solutions using the internet of medical things for hand gesture recognition system, *Complex & Intelligent Systems* 7(3)(2020) 1253-1264.
- [9] G. Khodabandelou, P.-G. Jung, Y. Amirat, S. Mohammed, Attention-Based Gated Recurrent Unit for Gesture Recognition, *IEEE Transactions on Automation Science and Engineering* 18(2)(2021) 495-507.
- [10] M.-S. Lee, K.-W. Kim, M.-H. Ryu, J.-N. Kim, Hand Gesture Recognition with Inertial Sensors and a Magnetometer, *Sensors & Materials* 28(6)(2016) 655-660.
- [11] D. Jiang, Z. Zheng, G. Li, Y. Sun, J. Kong, G. Jiang, H. Xiong, B. Tao, S. Xu, H. Yu, H. Liu, Z. Lu, Gesture recognition based on binocular vision, *Cluster Computing* 22(Suppl. 6)(2019) 13261-13271.
- [12] O.-K. Oyedotun, A. Khashman, Deep learning in vision-based static hand gesture recognition, *Neural Computing & Applications* 28(12)(2017) 3941-3951.



- [13] K. Liu, N. Kehtarnavaz, Real-time robust vision-based hand gesture recognition using stereo images, *Journal of Real-Time Image Processing* 11(1)(2016) 201-209.
- [14] Z. Tang, Q. Liu, M. Wu, W. Chen, J. Huang, WiFi CSI gesture recognition based on parallel LSTM-FCN deep space-time neural network, *China Communications* 18(3)(2021) 205-215.
- [15] S. Tan, J. Yang, Fine-grained gesture recognition using WiFi, in: *Proc. 2016 IEEE Conference on Computer Communications Workshops*, 2016.
- [16] R.-G. Sonkamble, A.-K. Rathod, G.-G. Macchale, P. Rajamanickam, R.-Y. Mulla, WiFi Gesture Recognition System, *ICATSA: Techno-Societal 2016*, International Conference on Advanced Technologies for Societal Applications, 2021.
- [17] J. Lien, N. Gillian, M.-E. Karagozler, P. Amihhood, C. Schwesig, E. Olson, H. Raja, I. Poupyrev, Soli: Ubiquitous gesture sensing with millimeter wave radar, *ACM Transactions on Graphics* 35(4)(2016) 1-19.
- [18] S. Wang, J. Song, J. Lien, I. Poupyrev, O. Hilliges, Interacting with soli: Exploring fine-grained dynamic gesture recognition in the radio-frequency spectrum, in: *Proc. 2016 Symposium on User Interface Software & Technology ACM*, 2016.
- [19] S. Hazra, A. Santra, Robust gesture recognition using millimetric-wave radar system, *IEEE Sensors Letters* 2(4)(2018) 1-4.
- [20] K.-A. Smith, C. Csech, D. Murdoch, G. Shaker, Gesture recognition using mm-Wave sensor for human-car interface, *IEEE Sensors Letters* 2(2)(2018) 1-4.
- [21] M.-J. Cheok, Z. Omar, M.-H. Jaward, A review of hand gesture and sign language recognition techniques, *International Journal of Machine Learning and Cybernetics* 10(1)(2019) 131-153.
- [22] Y. Wang, Y. Shu, X. Jia, M. Zhou, L. Xie, L. Guo, Multifeature Fusion-Based Hand Gesture Sensing and Recognition System, *IEEE Geoscience and Remote Sensing Letters* 19(2022) 1-5.
- [23] D. Yao, Y. Wang, W. Nie, L. Xie, M. Zhou, X. Yang, A Multi-feature Fusion Temporal Neural Network for Multi-hand Gesture Recognition using Millimeter-wave Radar Sensor, in: *Proc. 2021 IEEE Asia-Pacific Microwave Conference*, 2021.
- [24] D. Tran, L. Bourdev, R. Fergus, L. Torresani, M. Paluri, Learning Spatiotemporal Features with 3D Convolutional Networks, in: *Proc. 2015 IEEE International Conference on Computer Vision (ICCV)*, 2015.
- [25] Y. Kim, B. Toomajian, Hand Gesture Recognition Using Micro-Doppler Signatures With Convolutional Neural Network, *IEEE Access* 4(2016) 7125-7130.
- [26] H. Abdelnasser, K.-A. Harras, M. Youssef, A Ubiquitous WiFi-based Fine-Grained Gesture Recognition System, *IEEE Transactions on Mobile Computing* 18(11)(2019) 2474-2487.
- [27] W. Lei, X. Jiang, Q. Tan, L. Xu, Y. Zhao, T. Xu, Y. Li, Q. Gu, G. Liu, Y. Zhao, W. Li, A TD-CF preprocessing method of FMCW radar for Dynamic Hand Gesture Recognition, in: *Proc. 2019 IEEE International Conference on Signal, Information and Data Processing*, 2019.
- [28] B. Dekker, S.-A. Jacobs, A.-S. Kossen, M. Kruithof, A. Huizing, M. Geurts, Gesture recognition with a low power FMCW radar and a deep convolutional neural network, in: *Proc. 2017 European Radar Conference*, 2017.
- [29] Z. Zhou, Z. Cao, Y. Pi, Dynamic Gesture Recognition with a Terahertz Radar Based on Range Profile Sequences and Doppler Signatures, *Sensors* 18(1)(2018) 10.
- [30] A. Ren, Y. Wang, X. Yang, M. Zhou, A Dynamic Continuous Hand Gesture Detection and Recognition Method with FMCW Radar, in: *Proc. 2020 IEEE/CIC International Conference on Communications in China (ICCC)*, 2020.
- [31] Y. Wang, A. Ren, M. Zhou, W. Wang, X. Yang, A Novel Detection and Recognition Method for Continuous Hand Gesture Using FMCW Radar, *IEEE Access* 8(2020) 167264-167275.
- [32] Z. Liu, H.-R. Cui, J.-L. Wu, J.C. Xie, T. Liu, Gesture recognition algorithm based on continuous wave doppler radar, *Information Technology and Network Security* 38(3)(2019) 30-34.
- [33] Y. Sun, T. Fei, F. Schliep, N. Pohl, Gesture Classification with Handcrafted Micro-Doppler Features using a FMCW Radar, in: *Proc. 2018 IEEE MTT-S International Conference on Microwaves for Intelligent Mobility*, 2018.
- [34] J.-W. Choi, S.-J. Ryu, J.-H. Kim, Short-Range Radar based Real-Time Hand Gesture Recognition Using LSTM Encoder, *IEEE Access* 7(2019) 33610-33618.
- [35] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, F.-F. Li, Large-scale video classification with convolutional neural networks, in: *Proc. 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.
- [36] S. Ji, W. Xu, M. Yang, K. Yu, 3D Convolutional Neural Networks for Human Action Recognition, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35(1)(2013) 221-231.
- [37] G. Zhang, S. Lan, K. Zhang, L. Ye, Temporal-Range-Doppler Features Interpretation and Recognition of Hand Gestures Using mmW FMCW Radar Sensors, in: *Proc. 2020 14th European Conference on Antennas and Propagation*, 2020.
- [38] Z. Zhang, Z. Tian, M. Zhou, SmartFinger: A Finger-Sensing System for Mobile Interaction via MIMO FMCW Radar, in: *Proc. 2019 IEEE Globecom Workshops*, 2019.
- [39] D. Kingma, J. Ba, Adam: A Method for Stochastic Optimization. <<https://arxiv.org/abs/1412.6980>>, 2014 (accessed 30.01.17).
- [40] T. Chen, F.-T. Zhang, Z.-M. Liu, Gesture recognition based on FMCW millimeter-wave radar, *Applied Science and Technology* 48(6)(2021) 23-27.
- [41] Y. Wang, J.-J. Wu, Z.-S. Tian, M. Zhou, S.-S. Wang, Gesture Recognition with Multi-dimensional Parameter Using FMCW Radar, *Journal of Electronics and Information Technology* 41(4)(2019) 822-829.
- [42] K. Simonyan, A. Zisserman, Very Deep Convolutional Networks for Large-Scale Image Recognition. <<https://arxiv.org/>

abs/1409.1556>, 2014 (accessed 10.04.15).

- [43] S. Hazra, A. Santra, Radar Gesture Recognition System in Presence of Interference using Self-Attention Neural Network, in: Proc. 2019 18th IEEE International Conference On Machine Learning And Applications, 2019.