Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2024.056073

ARTICLE

# Multiscale Feature Fusion for Gesture Recognition Using Commodity Millimeter-Wave Radar

**Lingsheng Li[1], Weiqing Bai[2] and Chong Han[2,*]**

[1]College of Computer Engineering, Jinling Institute of Technology, Nanjing, 211169, China

[2]College of Computer, Nanjing University of Posts and Telecommunications, Nanjing, 210003, China

*Corresponding Author: Chong Han. Email: hc@njupt.edu.cn

Received: 13 July 2024 Accepted: 23 September 2024 Published: 15 October 2024

**ABSTRACT**

Gestures are one of the most natural and intuitive approach for human-computer interaction. Compared with traditional camera-based or wearable sensors-based solutions, gesture recognition using the millimeter wave radar has attracted growing attention for its characteristics of contact-free, privacy-preserving and less environment-dependence. Although there have been many recent studies on hand gesture recognition, the existing hand gesture recognition methods still have recognition accuracy and generalization ability shortcomings in short-range applications. In this paper, we present a hand gesture recognition method named multiscale feature fusion (MSFF) to accurately identify micro hand gestures. In MSFF, not only the overall action recognition of the palm but also the subtle movements of the fingers are taken into account. Specifically, we adopt hand gesture multiangle Doppler-time and gesture trajectory range-angle map multi-feature fusion to comprehensively extract hand gesture features and fuse high-level deep neural networks to make it pay more attention to subtle finger movements. We evaluate the proposed method using data collected from 10 users and our proposed solution achieves an average recognition accuracy of 99.7%. Extensive experiments on a public mmWave gesture dataset demonstrate the superior effectiveness of the proposed system.

**KEYWORDS**

Gesture recognition; millimeter-wave (mmWave) radar; radio frequency (RF) sensing; human-computer interaction; multiscale feature fusion

## 1 Introduction

Hand gesture recognition is one of the most commonly used human-computer interaction (HCI) [1] methods and is natural, intuitive and effective. Users can control electronic equipment through hand gestures [2], like vehicle equipment control, multimedia equipment regulation, smartphone interactions, sign language translation, etc. Many technologies, such as image vision [3], wearable equipment [4], biological signal-based [5], and wireless radio frequency detection technologies, have been adopted in existing hand gesture recognition methods. In visual gesture recognition, some researchers have used dedicated equipment, such as the Kinect or Leap Motion sensors, and other researchers acquired multiple images to obtain deep measurements by three-dimensional cameras

and multicamera [6]. However, the performance of gesture recognition methods based on image vision is easily affected under low light and no light conditions. At the same time, there are some privacy and power consumption issues [7]. Gesture recognition based on wearable devices captures changes in gesture motion through wearable sensing equipment [8], and commonly used signals include electrocardiography and acceleration [9]. However, this type of identification method can only be used when users wear sensor equipment. The gesture recognition method based on biological signals uses biological signals such as eye movements, brain electrical signals [10], and muscle signals as interaction data [11]. Many studies use electromyograms (EMGs) to classify and identify gestures [12]. However, gesture recognition based on biological signals is not universal, and human biological signals are mostly different. In addition, this method requires the user to wear specialized devices, which is inconvenient.

Wireless radio frequency technology has eliminated the abovementioned problems and attracted the attention of researchers in recent years [13]. Such studies include gesture recognition based on Wi-Fi and radar signals. Gesture recognition based on Wi-Fi signals uses the Doppler effect of Wi-Fi signals to extract gesture signal changes. The disadvantage of the Wi-Fi signal gesture recognition method is that the bandwidth is lower [14], and small gesture movement changes lead to a relatively small Doppler frequency shift, resulting in a lack of gesture recognition effects. In addition, Wi-Fi signal interference occurs extremely easily [15], resulting in a decline in gesture recognition effects.

With the development of mmWave communication and radar technologies, the use of mmWave radar for contactless HCI has gradually become a hotspot in recent years [16]. The advantages of gesture recognition in mmWave radar are particularly prominent [17,18]. A radar signal is not affected by the light environment, and normal signal echo can be observed under dim light conditions. In gesture recognition based on mmWave radar, the data signal is a radar radio frequency signal, not an optical image signal, providing privacy and security.

Radar-based gesture recognition methods collect gesture echo signals and then use signal processing methods to extract gesture features and movement information. Finally, gesture classification and recognition are performed through machine learning or deep learning algorithms. Certain results have been achieved in current technical research and radar gesture recognition applications. For example, Zhang et al. [19] used radar and Kinect sensors to divide the human arm into 7 parts to build a gesture method, which was based on the idea of "puzzle reconstruction". The authors employed a convolutional neural network (CNN) to separate and learn a gesture from users. They also tagged an attention network to enhance the feature components related to the gesture to deepen the learning of each arm part. Liu et al. [20] identified gestures based on relocation learning, which was divided into two parts: source domain recognition and target domain recognition. It effectively solved the problem of position dependence through relocation learning and the robustness of the network. The above two methods are aimed at arm-related action recognition. Arm rotation is used for identification and judgment, but it cannot be used for short-range gesture recognition. Meanwhile, there are many gesture echo signal features are chosen as the deep learning algorithm input for radar gesture recognition, such as range-Doppler map (RDM) [21], range-time map (RTM), Doppler-time map (DTM), angle-time map (ATM) [22], and range-angle map (RAM) [23]. These existing works adopted a single feature [21,24,25] or multidimensional feature [22,26] as the input for radar gesture recognition. These methods are mostly sensitive to palm actions and difficult to extract subtle finger action features, which easily leads to low gesture recognition accuracy. So, research on gesture recognition still has some limitations, e.g., how to use gesture features to improve the gesture recognition accuracy and how to accurately identify micro gestures still need in-depth research. Therefore, in this paper, we

propose a gesture recognition method with small actions and short-range applying assisted driving scenarios in which the driver's hand does not leave the steering wheel area when driving.

As the mainly gesture recognition features, the gesture angle and velocity information contained in the multiangle fused Doppler-time map (DTM) can assist in judging the direction of gestures and enhancing the recognition effect of easily confused gestures. Meanwhile, the range-angle map (RAM) is a two-dimensional heat map generated by the movement of the gesture and can intuitively show gesture movements. Therefore, aiming at the problem of poor classification of tiny finger movements and confusing gestures, this paper proposes a gesture recognition algorithm (named "MSFF") based on multifeature multiscale feature fusion convolutional neural network with multiangle fused DTM and gesture trajectory RAM as the input. The main contributions of our work are summarized as follows:

- To use gesture features to improve the gesture recognition accuracy, this paper adopts multiangle fused DTM and gesture trajectory RAM to comprehensively extract gesture feature information.
- For micro hand gesture recognition, this paper adopts a high-level network fusion and proposes a feature extraction method that is more suitable for radar gesture recognition and pays attention to subtle finger movements while taking the overall action recognition of the palm into account to improve the micro gesture recognition.

The rest of this paper is arranged as follows: Section 2 reviews the related work on gesture recognition. Section 3 introduces the basic principles of mmWave radar. Section 4 describes hand gesture signal model preprocessing, feature extraction, and the design of a multiscale feature fusion network. In Section 5, the experimental results and analysis are presented. The conclusion is drawn in Section 6.

## 2 Related Work

Radar-based gesture recognition involves transmitting and receiving millimeter-wave signals, using various signal processing methods to extract gesture features, and then employing machine learning models to classify the related gestures. So, in the present study, the process of Radar-based gesture recognition mainly consists of two steps: radar signal processing and gestures classification. In the radar signal processing phase, hand gesture features are extracted from the original radar echo. Before the gesture features are input into the network for gesture recognition, they must be extracted along with the motion information from radar signals using signal processing methods. Many researchers have achieved certain results in gesture feature extraction algorithms. For example, Zhang et al. [24] used a 5.8 GHz radar for gesture signal collection and two time-frequency analysis methods, Fourier transformation (FFT) and continuous wavelet transformation, to collect and analyze the received radar signal. However, only the time-frequency information was extracted from the radar signal in this work. The identification effects of some micromovement gestures and easily confused gestures are not good. Molchanov et al. [21] obtained the RDM through a two-dimensional Fourier transformation (2D-FFT). However, because action continuity features are not considered in gesture recognition, the RDM lacks time information and often does not have high recognition accuracy. Zhang et al. [27] used continuous waves, a time reuse method, a single-input Doppler radar sensor and a machine learning algorithm to classify gestures using specific Doppler signals. Although radar multidimensional feature information is used, the mining of the relationship between gesture multidimensional feature spaces is not sufficient.

To better use multidimensional gesture features, Liu et al. [28] processed the collected gesture data into point cloud included $X$-axis, $Y$-axis and $Z$-axis coordinate, range, Doppler, reflection point strength and other information. Sun et al. [29,30] used a multifeature encoder to encode the key 5D gesture feature points. The key points have the largest amplitude in the RDM, and the 5D features include the range, Doppler, time, angle, and amplitude information. These gesture recognition methods are based on point clouds. Although a point cloud contains multidimensional features and the amount of data within it is small, the point cloud resolution range is related to the sensor to detection object range, which is more suitable for long-range gesture recognition. In the point selection process, artificial key point selection is vulnerable to dynamic interference, adapting to complex scenes is difficult, and there is no strong advantage in the driving assistance environment. Gan et al. [31] obtained an RDM through 2D-FFT and proposed a new extraction feature, range-Doppler matrix focus (RDMF). In the matrix, the real and virtual parts are separated, and the two antennas are considered to obtain a three-dimensional quantity to reduce the number of data dimensions. The method used a three-dimensional convolutional network (3D-CNN) and long short-term memory (LSTM) classification framework to classify features. However, RDMF does not consider the gesture angle information, and it is not easy to identify gestures that have the same range and Doppler information. Xia et al. [26] based on a 77 GHz mmWave radar, established a gesture motion model to track gesture movement, determined the best reflection point in the opponent's model to extract the Doppler-time map (DTM), vertical angle-time map (VATM), and horizontal angle-time map (HATM) and then conducted multidimensional feature representation learning (MFRL) based on a three-channel CNN. MFRL aims to solve the robust demand for radar-based gesture recognition, that is, gesture is classified correctly without considering the user's identity, location, and perspective. However, because the mmWave radar platform angle resolution is not high as the range resolution, resulting extracted angle features are not as good in representing micro-motion gestures. Wang et al. [22] adopted the FFT and multiple signal classification (MUSIC) algorithm to measure the range, Doppler, and angle information of a gesture, constructed the corresponding range-time feature diagram through multi-frame accumulation methods to obtain a range-time map (RTM), Doppler-time map (DTM), and angle-time map (ATM), then designed a complementary multidimensional feature fusion (CMFF) approach for gesture recognition. Although CMFF makes use of multidimensional features, it does not dig deeply into the relationship between features and cannot reflect the integrity of gesture actions. Yu et al. [23] fused the gesture range and angle information to form a range-angle map (RAM) and added the RDM feature diagram as the multifeature fusion network input, which fuses RDM and RAM features (FRRF) for gesture recognition. FRRF uses ordinary feature fusion networks without introducing multiscale, and the micro finger movement identification effect is not good.

In conclusion, while the existing approaches that use single feature as the input of CNN-based radar gesture recognition are highly sensitive to palm movements and easily lead to reduced accuracy in gesture recognition. Conversely, the current multi-feature gesture recognition methods fail to account for the multi-scale characteristics of gesture features, which limits their generalizability. In this paper, we address these limitations by adopting multiangle DTM and gesture trajectory RAM, enabling multi-feature and multi-scale fusion. The proposed method allows for the comprehensive extraction of gesture features and the integration of deep neural networks, thereby enhancing sensitivity to subtle finger movements and improving overall gesture recognition accuracy and generalization performance.

## 3 MmWave Radar Related Principles

MmWave radar is special radar technology that uses short electromagnetic waves. The electromagnetic wave signal transmitted by the mmWave radar is blocked by an object and reflected. By capturing the radar reflection signal, the range, Doppler, and angle information of the object can be calculated. As a fully digital array variant, multiple-input multiple-output (MIMO) radar cannot only provide the range and velocity of the target but also the angle information.

### *3.1 Ranging Principle*

In this paper, a frequency-modulated continuous-wave (FMCW) mmWave MIMO radar platform with three transmitters and four receivers is used to collect gesture echo signal. The radar operates as follows. First, the frequency synthesizer generates a chirp, which is transmitted by the transmitting antennas (TXs). When the signal encounters an object, the chirp will be reflected back. The receiving antennas (RXs) soon receives the reflected chirp. Next, the signal of TXs and RXs are mixed and filtered by a low-pass filter (LPF) in a mixer, and the output of the mixer is a sinusoid called the intermediate frequency (IF) signal. Finally, the IF signal is sampled and processed to extract gesture features for recognition.

Assume that $t$ is the frequency-adjusted continuous pulse cycle; $S$ is the frequency growth slope; $\tau$ is the signal delay after sudden gesture recognition; and $f$ is the carrier frequency of the radar. The radar transmit signal $X_1$ can be expressed as:

$$X_1 = \sin\left(2\pi f t + \pi S t \cdot t\right) \tag{1}$$

The receiving signal $X_2$ is:

$$X_2 = \sin\left[2\pi f\left(t-\tau\right) + \pi S\left(t-\tau\right)^2\right] \tag{2}$$

Through the mixer and low-pass filter processing, the output IF signal $X$ is:

$$\mathrm{X} = \frac{1}{2}\cos\left(2\pi S\tau t + 2\pi f\tau - \pi S\tau^2\right) \tag{3}$$

For Eq. (3), frequency $f_{IF}$ of the IF signal can be obtained through a one-dimensional FFT. The gesture target to radar range is $d$, and the Doppler of light is $c$. We can obtain Eq. (4):

$$f_{IF} = S\tau = S\frac{2d}{c} \tag{4}$$

Therefore, the detection target range $d$ can be calculated by:

$$d = \frac{cf_{IF}}{2S} \tag{5}$$

The range-FFT results in Eq. (4) show the frequency response in different ranges. Fig. 1: FMCW radar signal processing schematic diagram is the principal diagram of gesture signal processing, and Fig. 1: FMCW radar signal processing schematic diagram (a) reflects the range-FFT raw signal processing result. The adopted FMCW mmWave radar in this paper has a range resolution of 3.75 cm due to the 4 GHz bandwidth [32]. Thus, between finger positions can be detected and micromovements can be distinguished.

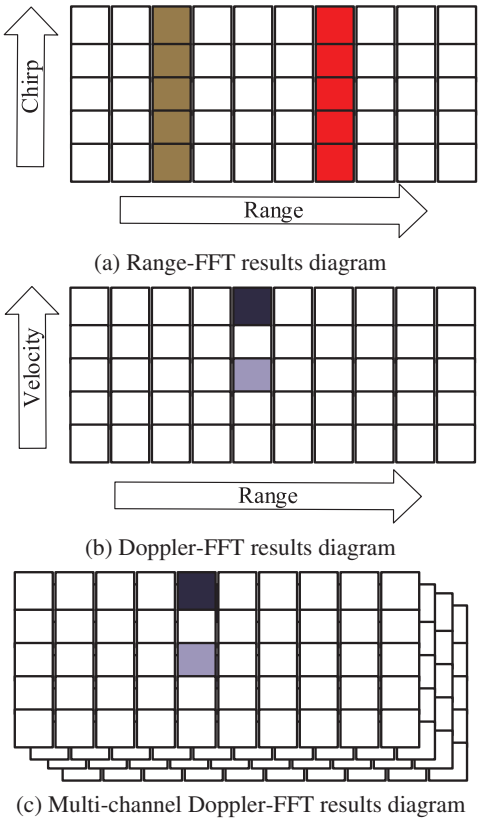



**Figure 1:** FMCW radar signal processing schematic diagram. From range-FFT, Doppler-FFT to multi-channel Doppler-FFT diagram

### *3.2 Doppler Measurement*

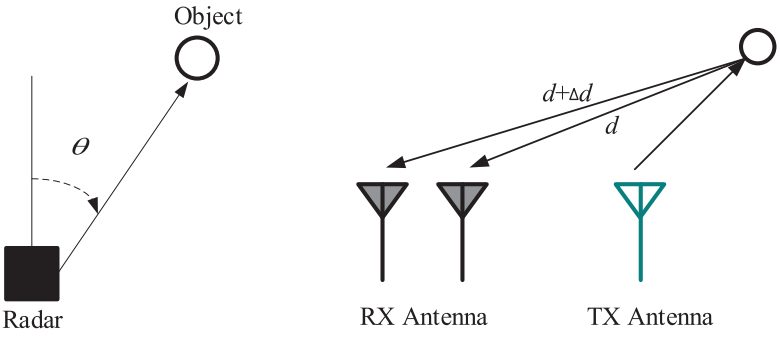
By using Fourier transformation for each chirp signal, a spectrum with different separation peaks will be produced. Each peak represents an object in a specific range. If multiple targets with different velocities are in the same range during a measurement, they will not be distinguished. Therefore, it is necessary to further extract the phase of each chirp signal in the same range and distinguish these targets. The goal of gesture velocity $v$ is to have two adjacent range-FFTs with different ranges, reflecting the target phase difference $\omega$ after the target performs an action during the chirp time $T_c$. Then, the goal is to obtain different velocities. The velocity $v$ is expressed as:

$$v = \frac{\lambda \cdot \omega}{4\pi \cdot T_c} \tag{6}$$

where $\lambda$ is the wavelength. To use the range and velocity information to distinguish multiple finger positions at the same time, another FFT, namely, Doppler-FFT, is performed along the range-FFT column, as shown in Fig. 1: FMCW radar signal processing schematic diagram (a), at the same range and with multiple fingers at different velocities. The different index position colors in Fig. 1: FMCW radar signal processing schematic diagram (b) show the two goals at the same velocity.

### *3.3 Angle Measurement*

Angle information is an important extractable feature for millimeter wave gesture recognition. The horizontal plane angle is further used to estimate the reflex signal to depict the exact position of the target in the Cartesian coordinate system space. As shown in Fig. 2a, this angle $\theta$ is also referred to as the angle of arrival (AOA).




(a) Angel of arrival diagram (b) Radar antenna schematic diagram.

**Figure 2:** Illustration of MIMO radar angle measurement. The angle of arrival can be calculated from the phase difference of multiple receiving antennas

The AOA estimation requires a phase change. The estimation requires a range difference between the object and different antennas, so at least two receiving antennas (RX antennas) are needed, as shown in Fig. 2b. The range between the echo signal received by the two antennas is processed in range-FFT. Because the range between the object and each antenna is different, a phase change will occur at the peak of the FFT.

In this configuration, Eq. (7) can be used to obtained phase change $\omega'$ information of at least two RX antennas:

$$\omega' = \frac{2\pi \Delta d}{\lambda} \tag{7}$$

where $\Delta d = d_{Inter-Rx} sin\theta$. $d_{Inter-Rx}$ is the range between antennas. We can use the FMCW radar with multiple receiving antennas to derive the angle of arrival $\theta$:

$$\theta = \sin^{-1}\left(\frac{\lambda \omega'}{2\pi d_{Inter-Rx}}\right) \tag{8}$$

So, to further distinguish finger overlap in the range-Doppler domain, we could perform the third FFT, that is, angle-FFT, on all receiving channels. For example, after the application angle-FFT, we can capture the AOA with the same range and Doppler, as shown in Fig. 1: FMCW radar signal processing schematic diagram (c).

## 4 Multiscale Feature Fusion Gesture Recognition Algorithm

### *4.1 Overall Framework*

The overall framework of the proposed MSFF gesture recognition algorithm is as follow:

First, we utilize millimeter-wave (mmWave) radar platform to collect user gesture data. From the raw echo data, we extract gesture features, specifically range, Doppler, and angle information. During

the feature extraction process, multiple chirp ranges are employed to enhance the signal-to-noise ratio. Additionally, a Hanning window is applied to mitigate spectral leakage, ensuring the acquisition of clean and distinct gesture feature information. Subsequently, we incorporate both velocity and time data, alongside the capabilities of multiple transmitters and receivers on the mmWave radar platform, to derive a multi-angle fusion Doppler Time Map (DTM).

Second, the MUSIC joint super resolution algorithm is used to estimate the data cube matrix by combining range and angle information, accurately calculating the gesture range and angle information, and generating a gesture trajectory RAM.

Then, these feature images with varied sizes are input into different convolution layers of the multiscale feature fusion network. After convolution layer feature extraction is performed, the features extracted from the DTM and RAM are fused. Finally, the model is trained and predicted using the obtained gesture dataset. The algorithm process is shown in Fig. 3.

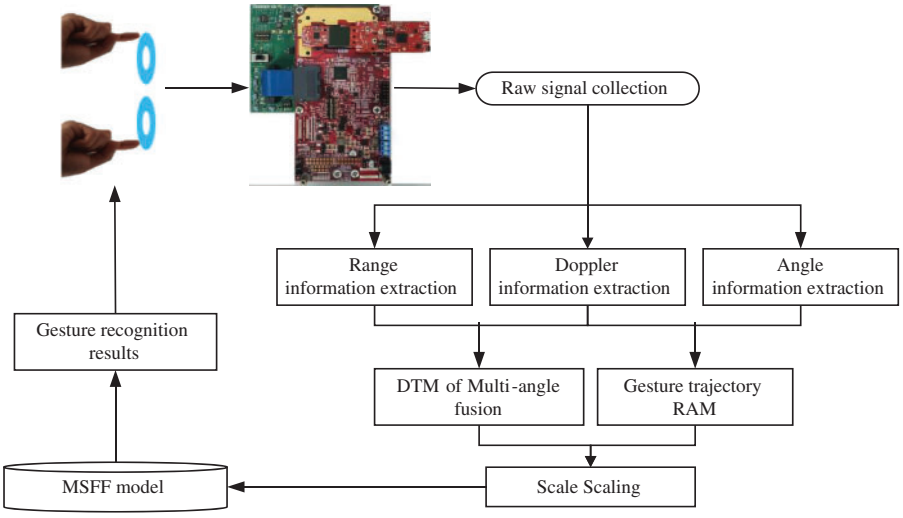



**Figure 3:** Multiscale feature fusion gesture recognition algorithm flow chart. The entire system can be divided into three parts: gesture detection, gesture signal processing, and recognition classification

### *4.2 Extraction of Gesture Range, Doppler, and Angle Information*

Using the collected raw gesture echo data, the gesture range, Doppler, and angle information features are extracted. Then, by combining this information, the multiangle fusion DTM and RAM are processed.

#### *4.2.1 Gesture Range Information Extraction*

By using the 32-bit unsigned data format, the 12-channel I/Q data are restored to a complex signal from the preserved raw gesture echo signal. First, we analyze the IF signal range spectrum when there is no gesture information. The radar receiving and transmitting signal mixed frequencies will include multiple IF signals with different frequencies. The frequency-filtered signal can be separated from the IF signals through a one-dimensional Fourier transformation. As shown in Fig. 4, for the IF signal range map in the no gesture scenario, a complete gesture (including no gesture) contains 50 data frames. From these data frames, we find that there is a strong peak at 10 cm (1 dm) from the longitudinal radar range. This peak is caused by the DC component of the mmWave radar,

i.e., a radar platform defect. This noise will affect the system's identification of dynamic gestures. Thus, the gesture goal may be misjudged and then must be addressed. In this paper, multiple chirp ranges are superimposed to enhance the signal-to-noise ratio, weakening the movement gesture target interference of the DC component.

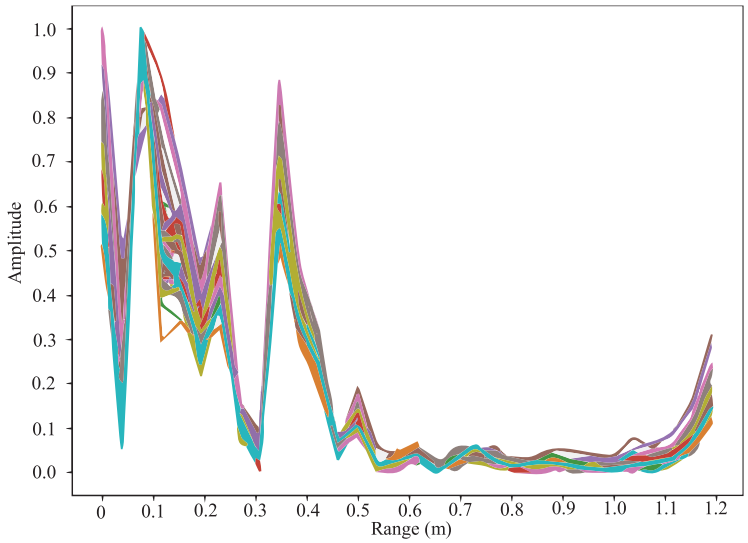
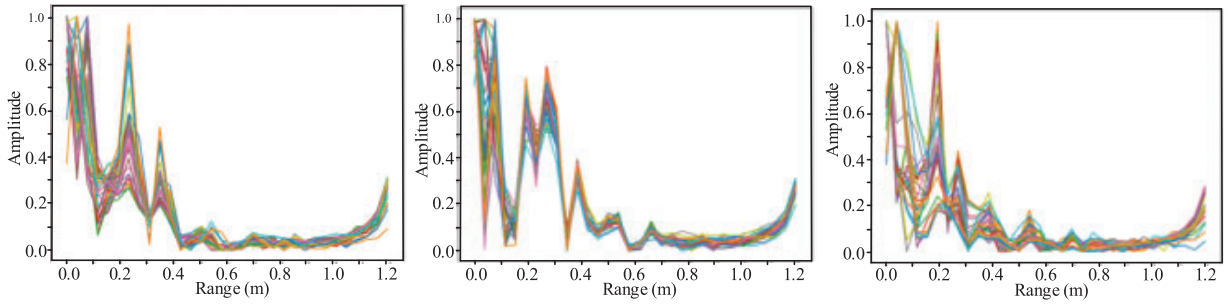



**Figure 4:** Spectrogram in the scene without gestures. Spectrogram reflects the current received radar transmission signal

Fig. 5 shows a dynamic gesture of a finger around a one-dimensional range. The 2nd, 20th, and 49th frames are extracted from the complete 50 frames of data for displaying. Each graph reflects the superposition of 32 chirp ranges in a data frame, and each chirp in a frame is basically the same. This also verifies the feasibility of using multiple chirp ranges to enhance the signal-to-noise ratio. For each frame of superimposed data, the peak search method is used to find the specific location of the gesture goals. By processing all 50 frames of a complete gesture, the gesture length is reduced to 32 sampling points. Thus, the dynamic gesture range is like that of a two-dimensional matrix with a size of $32 \times 50$.




(a) The 2nd frame (b) The 20th frame (c) The 49th frame

**Figure 5:** Gesture wind circle spectrum. The reflection of signals at different times for the same gesture is different

A two-dimensional heat map of a two-dimensional matrix of a finger moving around in a circle is shown in Fig. 6. The starting position is approximately 20 cm from the radar. Then the finger

slowly returns to its original position 20 cm from the radar. In Fig. 5, the 2nd, 20th, and 49th frames correspond to the initial, mid, and ending gestures, respectively, and the one-to-one correspondence with the range time two-dimensional heatmap shows the effectiveness of extracting the gesture range information.

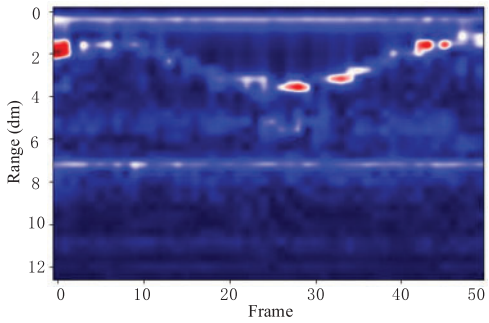



**Figure 6:** Range time two-dimensional heatmap of finger circling. The stronger the reflection of the signal, the redder the color in the heatmap

### *4.2.2 Gesture Doppler Information Extraction*

Doppler information can reflect dynamic hand gesture movements, and the positive and negative Doppler information reflects the gesture movement direction relative to the radar. The Doppler measurement needs to be completed by transmitting multiple chirps. For objects with relative Doppler information, the peak value and the corresponding frequency will not change after FFT in the range direction with different chirp values, but the phase at the peak will change according to the Doppler size rule, that is, rotation occurs. A schematic diagram of the phase change between different chirps is shown in Fig. 7.

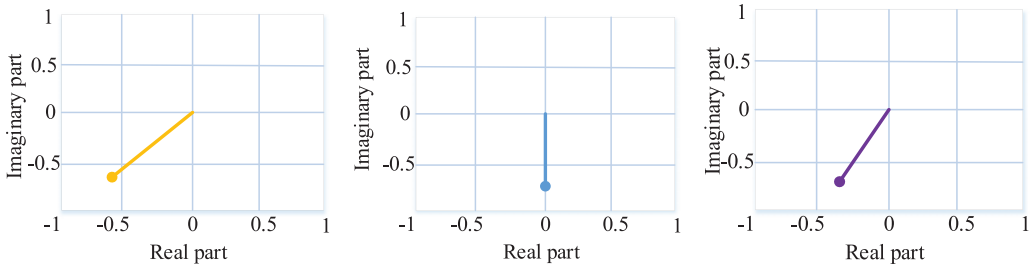



**Figure 7:** Phase change schematic diagram. The different time phase directions of the same gesture will change

Therefore, each corresponding chirp phase value changes. FFT is performed in the chirp dimension for the value obtained after FFT in the range direction to obtain the phase velocity chirp change frequency. In this paper, the abovementioned 2D-FFT is performed on the data matrix of each frame that sliding upward and then sliding down, and the Doppler information of the gesture target at different ranges can be obtained, as shown in Fig. 8.

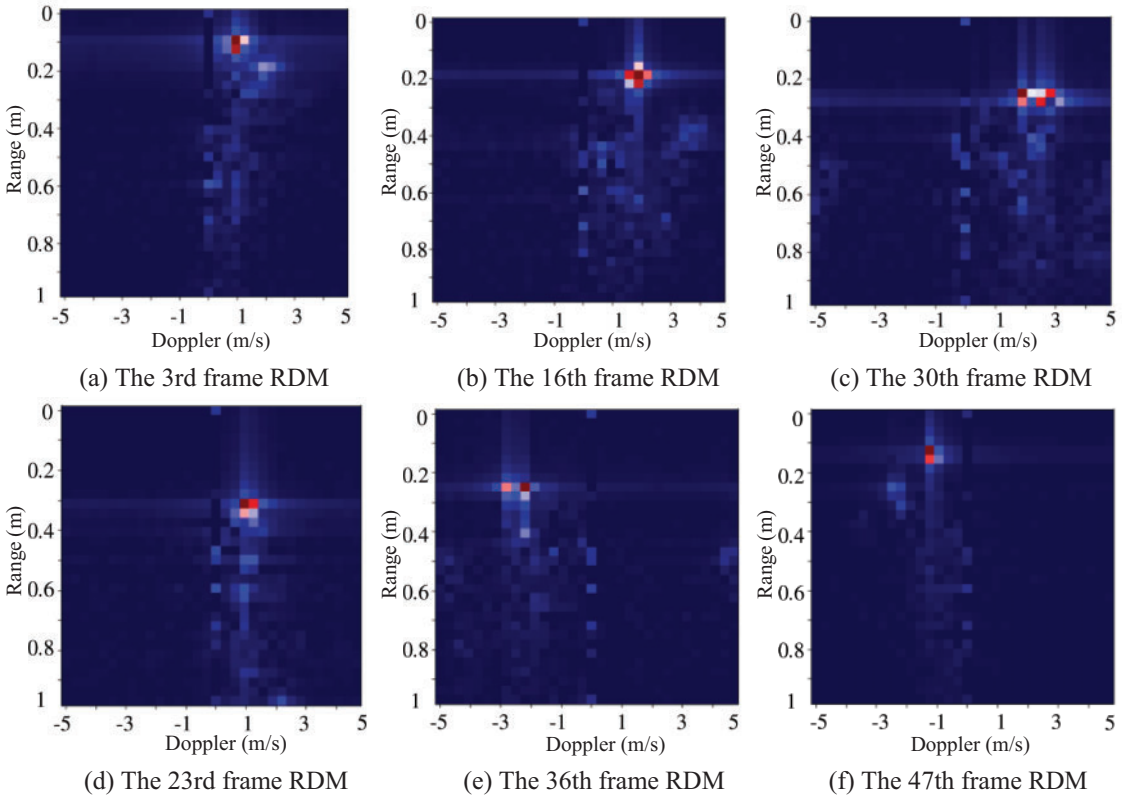



**Figure 8:** Sliding up and sliding down operations on the gesture Doppler symptom ranges. Different gesture heatmaps will present different features

In Fig. 8, the spectrum will leak during the FFT operation. The spectrum leak will reduce the spectral resolution, which makes it difficult to detect real objects. To solve this problem, we consider a window function before the FFT operation to reduce spectrum leakage. The Hanning window [26] can relieve the spectrum leak at a good frequency resolution. Therefore, the Hanning window is applied to the Doppler signal range and the range between each frame and the time domain signal corresponding to the time domain Doppler signal, as shown in Eq. (9):

$$\begin{cases} S_{rw}(N,\ P,\ f) = S(N,\ P,\ f) \times Hanning(N) \\ S_{dw}(N,\ P,\ f) = S_{rw}(N,\ P,\ f) \times Hanning(P) \end{cases} \tag{9}$$

where $S_{rw}(N,\ P,\ f)$ represents the range and Hanning window results, $S_{dw}(N,\ P,\ f)$ indicates the Doppler plus Hanning window result, and the result of using the Hanning window for the range Doppler characteristic diagram in Fig. 8 is shown in Fig. 9.

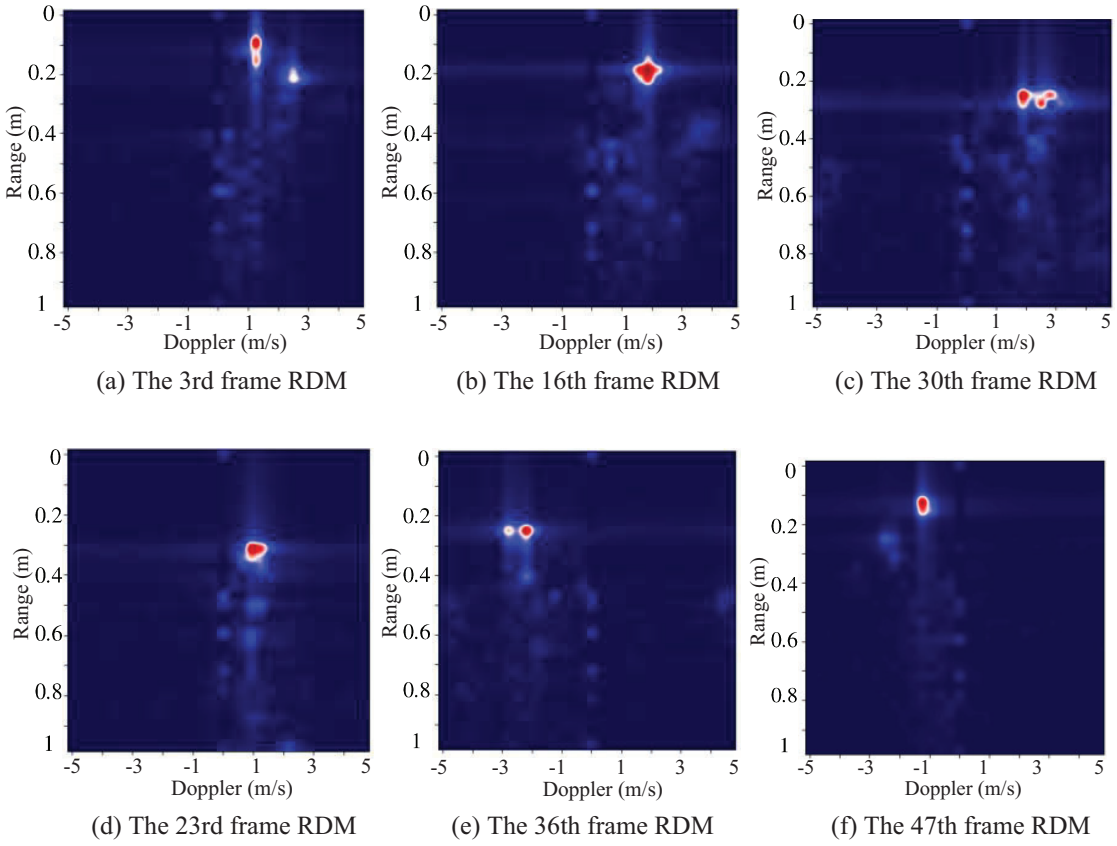



**Figure 9:** RDM after adding a window. As the gesture progresses, RDM will display a certain trajectory

Thus far, the gesture range and Doppler information have been extracted, and a two-dimensional thermogram of the range Doppler has been obtained.

#### *4.2.3 Gesture Angle Information Extraction*

The gesture angle information contains the azimuth information and elevation information of the hand. These two pieces of angle information reflect the gesture movement trajectory. To extract the gesture elevation corner information, we use the target range information corresponding to the gesture angle information from the one-dimensional range sequence. The angle can be calculated by using the range difference between the target and different receiving antennas. After the range-FFT performed, a search for the peak value is implemented. The phase value corresponding to the antenna is taken for each data point where the Doppler information is not zero at the peak value, the same operation is performed for multiple receiving antennas, and then the values are summed and averaged. The phase information is used to obtain the angle information. Finally, the gesture angle information map size is $32 \times 50$, as shown in Fig. 10.

### *4.3 Doppler-Time Feature Diagram Processing of Multiangle Fusion*

After extracting the Doppler information, there are many noise and miscellaneous waves in the RDM, which affects the DTM for accurate gesture extraction. We consider scattering object detection in the range-Doppler domain. Based on hand movements, the gesture movement Doppler information

is linked to its position changes, and the constraints of some prior parameters are given to establish a model for gesture tracking.

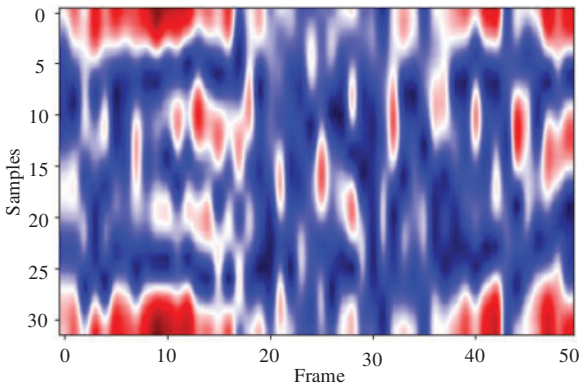



**Figure 10:** ATM of a finger circle example. ATM with different gestures will exhibit different characteristics

Because the number of motion scattering centers is not fixed in the RDM, we select a key movement scattering center as the gesture goal point to track the gesture motion trajectory. Suppose the Doppler information is constant during a frame cycle, and the movement is expressed as follows:

$$r\left(\vec{T}\right) = \overrightarrow{r_0} + \sum_{t=1}^{T} V_t \cdot \overrightarrow{T_f} \tag{10}$$

where $r\left(\vec{T}\right)$ represents the movement scattering center position at the $T^{th}$ frame, $\overrightarrow{r_0}$ is the starting position of the movement scattering center, and $V_t$ indicates the movement scattering center during the $T$ frame. The motion equation is described by displacement and Doppler information, so the key to determining the movement scattering center is to determine the target position and Doppler information of each frame. The specific steps are described below:

Firstly, select the first frame in the gesture motion RDM; filter all moving target points under the condition of satisfying the target detection; select a point with the minimum radial range and maximum radial velocity as the key point; obtain its corresponding two-dimensional azimuth pitch angle graph; and take the motion scattering center with the maximum amplitude in the two-dimensional angle graph as the initial target point. The initial target point has the minimum radial range, the maximum radial velocity and maximum energy amplitude.

Secondly, calculate the radial range $r_1$, radial velocity $v_1$, azimuth $\theta_1$ and elevation $\varphi_1$ of the initial target point in the first frame.

Finally, detect the motion target point in the $T$th frame RDM, and then calculate the radial range $r_T$, radial velocity $v_T$, azimuth $\theta_1$ and elevation $\varphi_1$ of all target points. This paper determines the best motion scattering center of the $T$th frame as that with the smallest displacement center tracked in all target points and the previous frame. The displacement size $d$ can be represented by Eq. (11):

$$d = min\left(\sqrt{(x_T - x_{T-1})^2 + (y_T - y_{T-1})^2 + (z_T - z_{T-1})^2}\right) \tag{11}$$

where $(x_T,\ y_T,\ z_T)$ represents the coordinate position of the $T$th frame motion target point in the three-dimensional space and $(x_{T-1},\ y_{T-1},\ z_{T-1})$ means that the coordinate position of the $T-1$th frame motion target point in the three-dimensional space can be represented by the radial range $r_{T-1}$, azimuth

$\theta_{T-1}$, and elevation $\varphi_{T-1}$:

$$\begin{cases} x_{T-1} = r_{T-1} \cdot \cos(\theta) \cdot \sin(\varphi) \\ y_{T-1} = r_{T-1} \cdot \cos(\theta) \cdot \cos(\varphi) \\ z_{T-1} = r_{T-1} \cdot \sin(\varphi) \end{cases} \tag{12}$$

Thus far, we have constructed a hand-scattered motion model, which can ensure the spatial continuity of time-varying features. Hand tracking based on the motion equation can be used to accurately extract effective targets, suppress the invalid noise interference, and provide important screening work for accurate Doppler information extraction. We extract the coordinate positions ($R_T$, $C_T$) of the key points in the RDM, where $R_T$ represents the coordinate positions in the range dimension, and $C_T$ represents the coordinate positions in the chirp dimension. We use the $C_T$ information to generate the Doppler spectrum and splice consecutive gesture frames to form the DTM, as shown in Fig. 11.

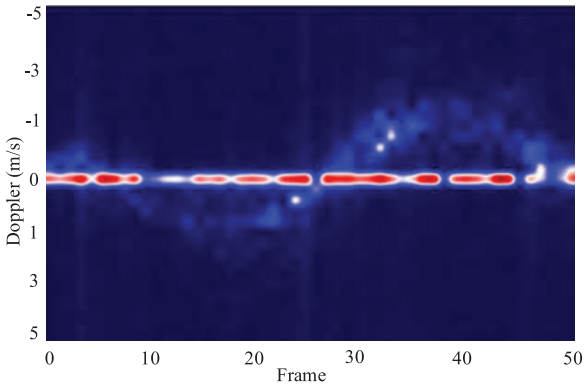
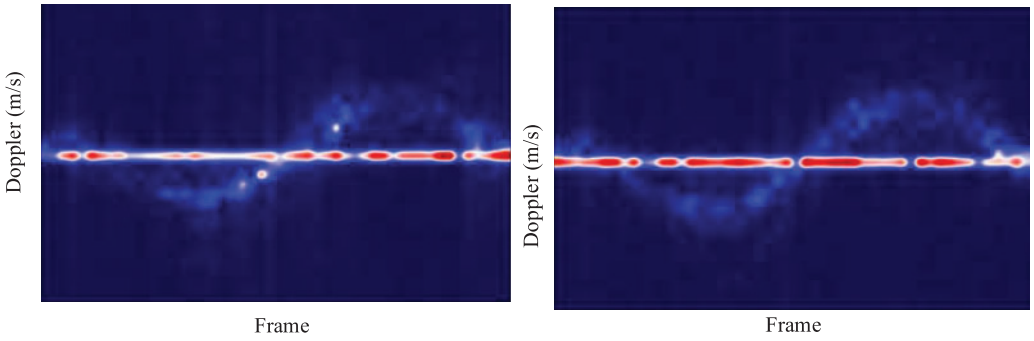



**Figure 11:** Average full channel DTM for a finger circle example. The DTM map integrates the feature information of HDTM and VDTM

The traditional DTM accumulate and average the full mmWave channel to enhance the signal-to-noise ratio. The mmWave platform used in this paper is a MIMO fully digital array. The two vertical angles are extracted from azimuth and elevation channels. The azimuth and elevation information are used to generate HDTM and VDTM, respectively, as shown in Fig. 12.

To maintain the advantages of the high signal-to-noise ratio CA-DTM, the MIMO array angle information is used to improve the pair recognition rate of confusion. We integrate three feature maps into the three RGB color image channels to form a multiangle fusion DTM, as shown in Fig. 13. The DTM obtained from three different angles shows the difference in radar signals from different angles of the same gesture in terms of time and Doppler information. Because of this subtle difference, the DTM has good feature differences that can be used to recognized easily confused gestures and subtle actions, which can improve the accuracy and precision of gesture recognition.




(a) Finger Wrap HDTM (b) Finger Wrap VDTM

**Figure 12:** Different finger circle angles in the DTM. The characteristics of HDTM and VDTM are slightly different

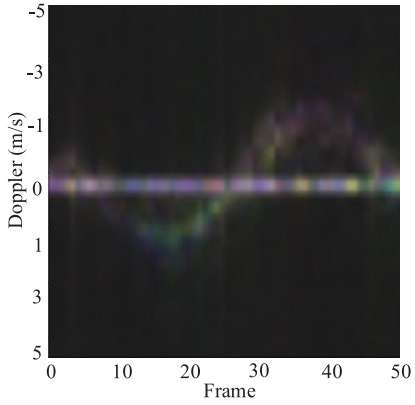



**Figure 13:** Multiangle DTM fusion. Multiangle DTM fusion contains richer information and more distinct features

### *4.4 Range-Angle Gesture Trajectory Feature Processing*

The generation of a range-angle gesture trajectory feature map requires the joint estimation of the gesture range and azimuth information. We use the MUSIC joint super resolution algorithm to estimate the range and azimuth information based on range-FFT to obtain the radial as well as lateral information of a gesture. In this paper, the received raw data need to be reconstructed to make them suitable for processing by MUSIC algorithm, and the collected data will be reshaped into an $N$ number of sampling points $\times$ $P$ number of chirp $\times$ $L$ number of virtual receiving antennas, which can be expressed by Eq. (13):

$$s(n,\ p,\ l) = \exp\left(j2\pi\left[\left(\frac{2KR}{C} + f_d\right)\frac{n-1}{N}T + \frac{2f_cR}{C} + f_d pT + \frac{(l-1)\,d_r \sin(\theta)}{\lambda}\right]\right) \tag{13}$$

where the $n = 1,\ 2, \ldots, N$; $p = 1,\ 2, \ldots, P$; $l = 1,\ 2, \ldots, L$. The three-dimensional space matrix contains range, Doppler, and azimuth information, as shown in Fig. 14.

In this paper, the cube matrix $S(N,\ P,\ L)$ in Eq. (13) uses the MUSIC combined superresolution algorithm to make estimations, and the three-dimensional matrix $S$ is only adjusted to the $S_{N\times L}$ two-dimensional matrix in $P$ to generate an RAM for each frame signal. The finger circle gesture data was

processed as described above, an RAM of 50 frames was obtained, and the 2nd, 13rd, 30th, and 42nd frames circle RAM were drawn, and the results are shown in Fig. 15.

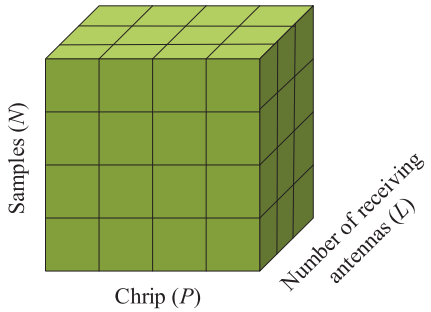



**Figure 14:** Schematic diagram of the three-dimensional space matrix, which is a data cube

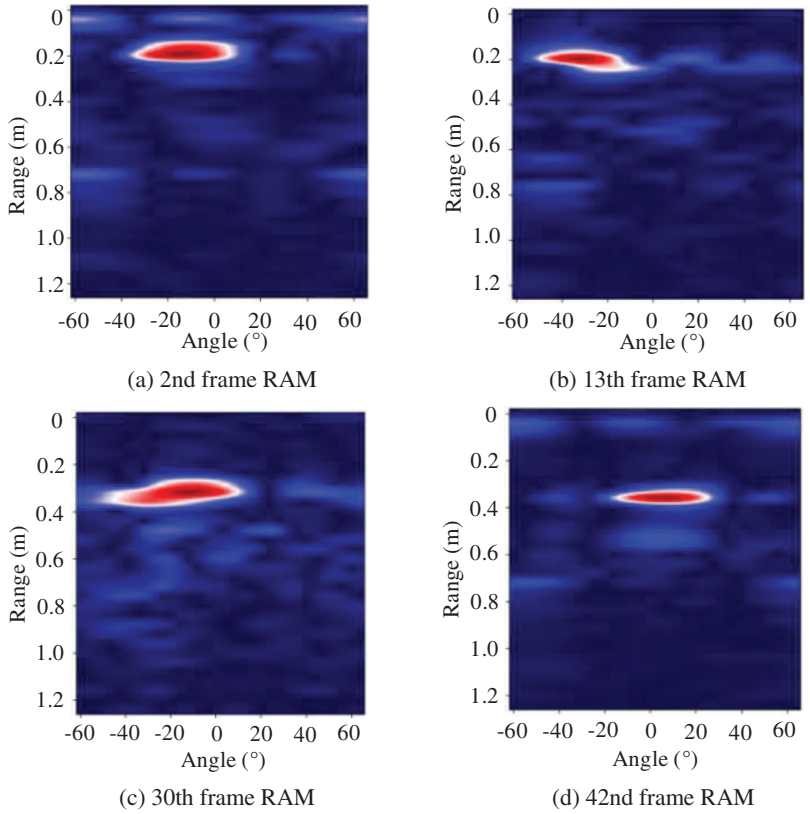



**Figure 15:** Finger circle RAM. The RAM characteristics of the same gesture at different times are different

Then, the abovementioned feature processing method is applied to each 50-frame of complete gesture data, and the RAM of 50 consecutive frames is accumulated into a graph constituting a gesture trajectory map to show the lateral and radial information of the gesture. As shown in Fig. 16, the high energy area forms a circle-like shape, and the finger circle gesture is clearly identified.

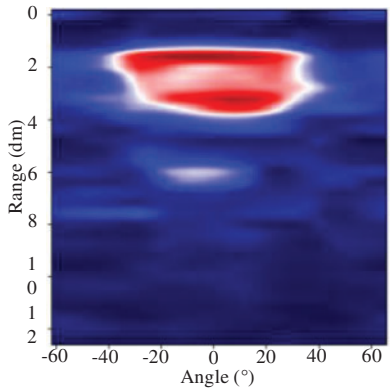



**Figure 16:** Accumulation of 50 frames of an RAM. The accumulated RAM after multiple frames contains the trajectory information of the gesture

### *4.5 Multiscale Feature Fusion Gesture Recognition Network Architecture*

After extracting the multiangle fusion DTM and the gesture motion trajectory RAM, we adopt them as the input of multiscale feature fusion networks for gesture recognition. It is first necessary to scale the two feature images from their original size of 64 × 64 to sizes of 32 × 32 and 16 × 16. Then, these feature images of varied sizes are input into different convolutional layers in the multiscale feature fusion network. After feature in the convolutional layer is extracted, the features extracted from DTM and RAM are fused. The multiscale network architecture diagram is shown in Fig. 17.

The multiscale feature fusion network contains two sets of inputs with multiangle fusion DTM and gesture trajectory RAM. The convolution operations of the two groups of inputs are the same. Taking the multiangle fusion DTM as an example:

(1) First, the multiangle fusion DTM with the original size of 64 × 64 is convolved with a convolution kernel of size 3 × 3. In the following, it is summed with a multiangle fusion DTM of size 32 × 32, scaled to half of its original size and convolved with a convolution kernel of size 1 × 1. Then, it is summed with a convolution kernel of size 3 × 3 with a multiangle fusion DTM of size 16 × 16, scaled to one-quarter of its original size and convolved with a convolution kernel of size 1 × 1. The batch normalization (BN) layer uses the convolution data to generate a more stable distribution, accelerate the training and convergence of the network, and slightly reduce the strong initialization dependence. We use a nonlinear activation function (ReLU) to increase the nonlinear properties in convolutional operations and apply this function in all elements of the input tensor without changing its space or depth information.

A dropout layer is added after the second and third convolutional layers, and the dropout rate is set to 0.4. The addition of dropout layer can temporarily remove the neural network training units from the network according to a certain probability during the training process because it is randomly discarded. Thus, a different network is trained in each epoch, which can play a role in reducing overfitting.

(2) After feature extraction is performed in the CNN, the flatten layer flattens the input data, that is, the multidimensional input becomes one-dimensional in the transition from the convolution layer to the fully connected layer.

(3) In the fusion process, multiple inputs are set up with the different dataset features. In this paper, multiangle fusion DTM and gesture trajectory RAM are used as the two inputs to the fusion network. The features from the above two feature maps are fused after the convolution and flattening operations, and gesture recognition is performed by using the feature complementarity of the multilayer convolutional neural network, i.e., the advantages of a strong representation of semantic information in the higher layer network and geometric detail information in the lower layer network are fully utilized, and the prediction is performed in the last feature fusion layer to recognize different gestures.

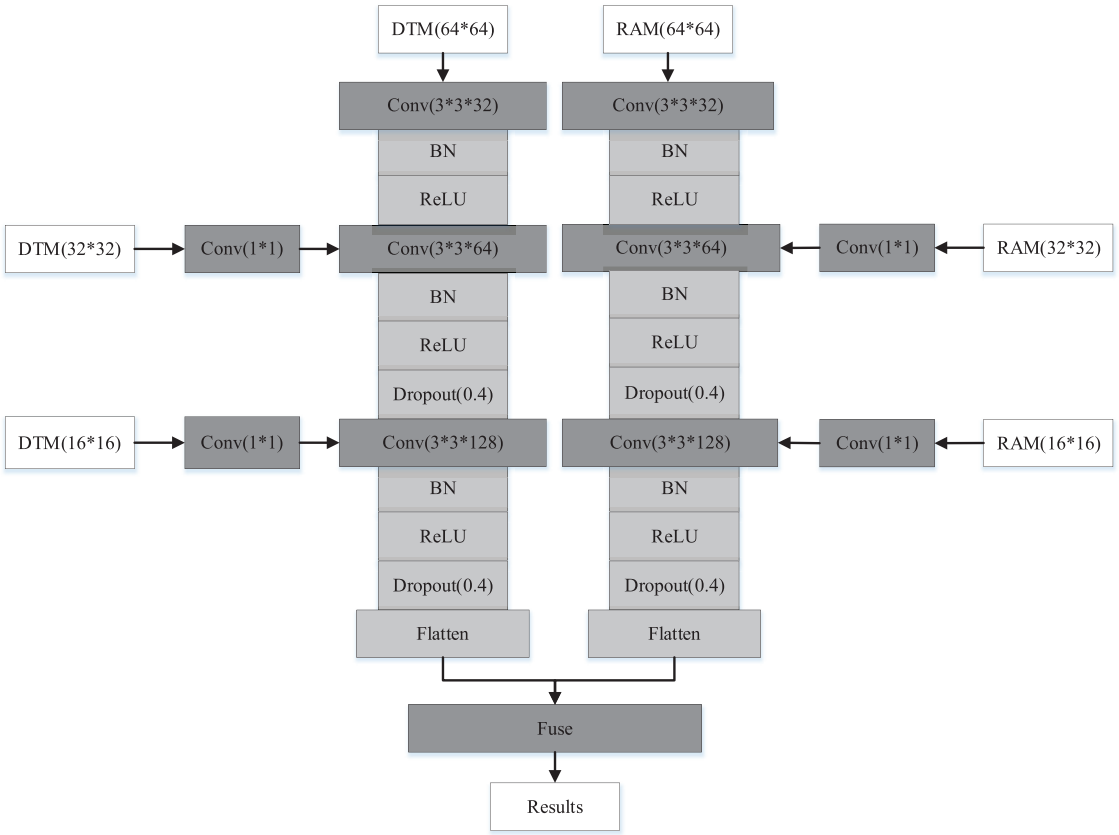



**Figure 17:** Multiscale feature fusion network architecture diagram. The multiscale feature fusion network contains two sets of inputs with multiangle fusion DTM and gesture trajectory RAM

## 5 Experiment and Analysis

### *5.1 Experimental Setup and Parameter Settings*

To evaluate the design of the gesture recognition method in this paper, a mmWave radar testbed for collecting gesture data is designed. As shown in Fig. 18, the system consists of two functional modules: the IWR6843BOOST-ODS [32] and the real-time high-speed data acquisition adapter. The data capture adapter captures raw ADC data from the radar chip via a low-voltage differential signal interface and outputs the raw data for further processing. Based on the time-division multiplexing MIMO scheme, three transmit antennas and four receive antennas are employed. The radar system uses a two-dimensional virtual antenna array including 12 data channels. There are up to 4 virtual

channels in the horizontal and vertical directions, corresponding to angular resolutions of 29° and 29°, respectively. The sliding window length of the gesture features is set to 50 frames or 2 s based on the characteristics of the gesture movements.

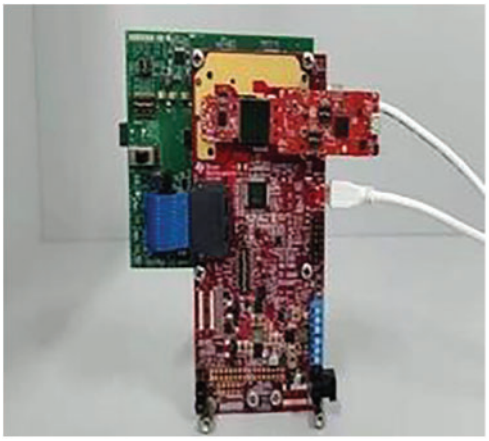


**Figure 18:** FMCW mmWave radar platform. The system consists of two functional modules: mmWave chipset and data acquisition adapter

The application scenario in this paper is vehicle driver assistance, that is, the driver makes gestures to manipulate the multimedia equipment in the car. Taking the driver's posture into account when driving the car and the small space in the car, the radar sensor device is placed flat on a table and gestures are performed within 20 cm of the radar device, while the gesture movement range is controlled within 1 m to simulate the manipulation of multimedia equipment in the car, as shown in Fig. 19.

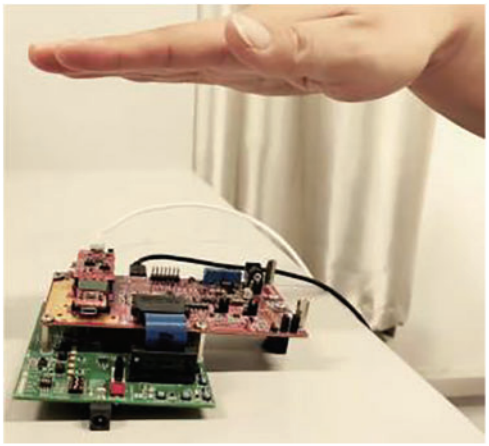


**Figure 19:** Gesture data acquisition. This deployment is used to simulate gesture operations in driving scenarios

In the data acquisition process, a 12-channel radar system, with 3 transmitting and 4 receiving channels, was turned on, and 8 gestures are designed as the target gestures: double click, finger wrap, right sweep, left sweep, slide up, slide down, slide up then slide down and no gesture, as shown in Fig. 20.

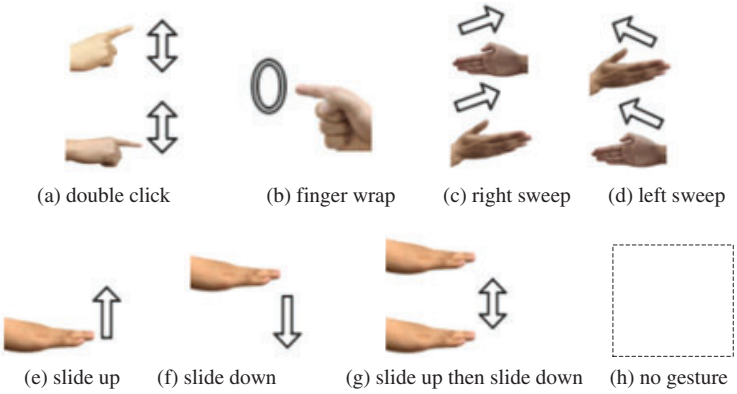



**Figure 20:** Gesture categories. These gestures are designed to simulate relevant operations in driving scenarios

When collecting radar gesture data, users are allowed to perform hand gestures within a longitudinal range of 1 m from the radar. Thus, lateral distance changes will lead to angle and Doppler measurement changes to enrich the data diversity, and at the same time, the robustness of the algorithm can be verified. In addition, we recruited 10 users (7 males and 3 females) to perform 8 gesture actions for strong generalization capability. Each action is performed at a distance of 1 m relative to the radar, and each action is repeated 30 times. The ten users are divided into 8 training users and 2 test users, and the dataset corresponding to the test user data is not included in the training. According to the characteristics of gesture movement, the effective gesture range is intercepted as 1 m $\times$ 1 m, while these data are collected in different situations, such as different time points, different movement speeds, and different posture standards.

### *5.2 Experimental Comparison and Results*

We divide the 8 training users into a training set and test set with 8:2 to construct the gesture dataset and conduct multiple sets of comparison experiments using different algorithms, including the RDM+RDM-based 3DCNN+LSTM gesture recognition algorithm proposed by Gan et al. [31], CMFF gesture recognition algorithm based on RTM+DTM+ATM proposed by Wang et al. [22], RDM-and RAM-based feature fusion FRRF gesture recognition algorithm proposed by Yu et al. [23], RDM-based end-to-end (ETE) gesture recognition algorithm experiments proposed by the Soli team [25], the DTM+HATM+VATM-based MFRL gesture recognition algorithm MSFF proposed by Xia et al. [26]. Ablation experiments are conducted to verify the necessity of the gesture feature preprocessing and multiscale feature fusion networks.

From Tables 1 and 2, we can see that the average recognition precision and the overall recognition accuracy of the model using the MSFF algorithm reached 99.7% on 8 gestures, reflecting the better performance of this model compared with the other 7 gesture recognition algorithms studied. The differences between the MSFF algorithm and several other algorithms will be compared by combining the recognition accuracy and precision of different gestures.

**Table 1:** Recognition precision of different gesture recognition algorithms on the self-constructed dataset

| Algorithms | Input features | Gesture recognition precision (%) | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | (a) | (b) | (c) | (d) | (e) | (f) | (g) | (h) | Avg |
| CMFF [22] | RTM+DTM+ATM | 83.3 | 84.3 | 85.1 | 90.7 | 97.9 | 100 | 97.8 | 100 | 92.4 |
| MFRL [26] | DTM+HATM+VATM | 84.0 | 93.8 | 85.4 | 87.2 | 100 | 100 | 97.9 | 98.0 | 93.3 |
| RDMF [31] | RDM+RDM | 96.0 | 100 | 89.6 | 87.5 | 100 | 100 | 100 | 100 | 96.6 |
| FRRF [23] | RDM+RAM | 98.5 | 100 | 100 | 99.3 | 97.8 | 99.2 | 97.8 | 100 | 99.1 |
| ETE [25] | RDM | 100 | 100 | 100 | 96.0 | 100 | 100 | 97.9 | 100 | 99.2 |
| MSFF w/o AF | DTM+RAM | 100 | 100 | 94.1 | 100 | 100 | 100 | 100 | 100 | 99.2 |
| MSFF w/o AF+MS | DTM+RAM | 100 | 98.0 | 94.0 | 97.8 | 100 | 100 | 100 | 100 | 98.7 |
| **MSFF (Ours)** | **DTM+RAM** | **100** | **100** | **98.0** | **100** | **100** | **100** | **100** | **100** | **99.7** |

**Table 2:** Recognition accuracy of different gesture recognition algorithms on the self-constructed dataset

| Algorithms | Input features | Recognition accuracy (%) |
|---|---|---|
| CMFF [22] | RTM+DTM+ATM | 92.19 |
| MFRL [26] | DTM+HATM+VATM | 93.23 |
| RDMF [31] | RDM+RDM | 96.61 |
| FRRF [23] | RDM+RAM | 99.07 |
| ETE [25] | RDM | 99.21 |
| MSFF w/o AF | DTM+RAM | 99.22 |
| MSFF w/o AF+MS | DTM+RAM | 98.70 |
| **MSFF (Ours)** | **DTM+RAM** | **99.74** |

CMFF uses the RTM, DTM and ATM as three independent inputs in the CMFF network, which does not make full use of the connection between the gesture range, Doppler and angle information and does not incorporate the "multiscale" idea. Therefore, a high recognition rate for small gestures is not observed by fusing these three independent features.

MFRL uses DTM, HATM and VATM fused information as the data input into the network. Multiangle fusion can further improve the recognition of confusing gestures (e) and (f). However, it does not fully allow the network to learn gesture features better due to the lack of range and Doppler information.

RDMF uses two-channel RDM as the input of the network, and the feature map does not contain the multiangle gesture information, which has a certain impact on the recognition accuracy of confusing gestures and micro gestures. The RDM and RAM are used by FRRF as the network input, and the added angle information makes the network model further improve the recognition rate of confusing gestures, but there is still a certain disadvantage in the recognition of micro gestures.

The ETE uses a single-channel RDM input in a network composed of 2D-CNN and LSTM, which improves the recognition of micro gestures and is not much different from the method proposed in this paper in terms of gesture recognition accuracy. However, this model lacks generalization ability.

MSFF w/o AF and MSFF w/o AF+MS reflect the ablation validation experiments. MSFF w/o AF indicates that the DTM without angle fusion preprocessing is input to the multiscale feature fusion network. It is found that the DTM without multiangle fusion leads to a lower recognition rate when recognizing the confusing gestures in groups (c) and (d) because of the lack of multiangle fusion information, illustrating the necessity of gesture feature multiangle fusion preprocessing before network input to recognize confusing gestures. MSFF w/o AF+MS indicates that multiangle fusion is not used and no multiscale transformation is performed to input DTM and RAM into the fusion network. The experimental results show that not only is the recognition of confusing gestures low, but in (b), the recognition rate of subtle finger gestures is also reduced, which indicates that the incorporation of multiscale transformations improves the attention to subtle gestures.

Table 3 shows the recall and F1-score of 8 different gesture recognition algorithms on the self-constructed dataset. It can be seen from Table 3 that the R and F1 values obtained by the MSFF algorithm proposed in this paper are 99.74% higher than those of other algorithms, reflecting the better classification performance of the network model proposed in this paper.

**Table 3:** Recall and F1-score of different gesture recognition algorithms on the self-constructed dataset

| Algorithms | Recall | F1-score |
|---|---|---|
| CMFF [22] | 92.19 | 92.20 |
| MFRL [26] | 93.23 | 93.25 |
| RDMF [31] | 96.61 | 96.62 |
| FRRF [23] | 99.07 | 99.07 |
| ETE [25] | 99.22 | 99.22 |
| MSFF w/o AF | 99.21 | 99.22 |
| MSFF w/o AF+MS | 98.69 | 98.70 |
| **MSFF (Ours)** | **99.74** | **99.74** |

In summary, among the 8 gestures in our self-built dataset, (a) and (b) are subtle finger movements used to verify the algorithm's recognition performance for small motion gestures. (c), (d), (e) and (f) are common palm movement gestures, including multiple pairs of easily confused gestures, used to verify the classification performance of the algorithm on gestures with high similarity. Through the comparison of existing algorithms and ablation experiments, it can be seen that using only single or multi-dimensional features as inputs for CNN for radar gesture recognition is only sensitive to palm movements and does not pay attention to local finger movements. It is difficult to extract subtle finger movement features, which can easily lead to low accuracy in gesture recognition. The idea of multi-scale and multi-dimensional feature fusion in our proposed method MSFF, combined with the strong semantic information representation ability and the strong geometric detail information representation ability of CNN, has shown superiority in experimental results of small motion gestures. In addition, the gesture trajectory, velocity, and angle information contained in DTM and RAM after gesture feature preprocessing can enhance the recognition effect of easily confused gestures.

In order to verify the computational efficiency of the algorithm MSFF proposed in this paper, the spatial and temporal complexity of the MSFF was compared with MFRL, CMFF, FRRF, and ETE. The experimental results are shown in Table 4. From Table 4, it can be seen that the size of the algorithm parameter model in this article is about 10 MB, which can be embedded into portable

devices. The feature processing and classification time consumption of MSFF is 97 ms, and can achieve a detection rate of 10 frames per second. The time complexity basically meets the requirements of real-time performance. Based on the comparative experiments with other algorithms, it can be concluded that the MSFF has higher computational efficiency while ensuring accuracy.

**Table 4:** Comparison of time and space complexity of different gesture recognition algorithms

| Algorithms | MFRL [26] | CMFF [22] | FRRF [23] | ETE [25] | MSFF (Ours) |
|---|---|---|---|---|---|
| Parameter model size (KB) | 6458 | 86030 | 30235 | 27952 | **10176** |
| Time consumption for feature processing and classification (ms) | 228 | 118 | 136 | 113 | **97** |

### *5.3 Generalization Ability Comparison on the Self-Constructed Dataset*

We verify the generalization performance of the proposed algorithm on the self-constructed dataset and compare the results with those of 5 existing algorithms. We test 2 untrained users in the self-constructed dataset to compare the generalization performance of the 6 algorithms. As shown in Table 5, the recognition rate of the proposed algorithm based on the MSFF algorithm reaches 87.29% on 8 gestures, which is higher than the other 5 gesture recognition algorithms.

**Table 5:** Recognition effect of users who did not participate in training in the self-built dataset

| Algorithms | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|
| CMFF [22] | 66.25 | 70.67 | 66.25 | 64.11 |
| MFRL [26] | 85.63 | 85.55 | 85.63 | 85.13 |
| RDMF [31] | 70.63 | 68.58 | 70.63 | 68.07 |
| FRRF [23] | 74.17 | 80.44 | 74.17 | 72.57 |
| ETE [25] | 52.71 | 65.75 | 52.71 | 53.10 |
| **MSFF (Ours)** | **87.29** | **88.40** | **87.29** | **86.20** |

From the results shown in Table 5, we can conclude that MSFF further improves the generalization ability of the network model after performing multiangle fusion and adding multiscale information. Compared with the ETE and FRRF algorithms, whose gesture recognition accuracies are not much different from the algorithm proposed in this paper, our algorithm improves the generalization ability by 10%, further proving the importance of multiangle and multiscale information.

### *5.4 Recognition Performance Comparison on the Google Soli Dataset [25]*

In this paper, with reference to the feature image RDM of the ETE network input, the RDM of different channels are selected as the two inputs of the multiscale feature fusion network. By using this setup, the same dataset and input feature map are obtained, and only the network architecture is different. Thus, the comparison experiment results can effectively illustrate the multiscale feature fusion network performance. The experimental parameters are set the same as those in ETE, the number of iteration epochs is 50, and the dataset is divided into 50% training and 50% testing, with 238 images each. The experimental results in Table 6 show that the average accuracy of MSFF in terms of

classifying these 11 gestures is 98%, which is higher than the experimental result of ETE, which is 94%, and the recognition results of each specific gesture are shown in Table 6. There is a large gap between MSFF and ETE in terms of the recognition results of two gestures, Gesture 0 (pinch index finger) and Gesture 1 (palm flip). The MSFF recognition rates for these two gestures are 95.4% and 88.2%, respectively, while for ETE, these values are only 79.2% and 74.4%, respectively. The main reason is that MSFF incorporates a multiscale feature fusion method, which increases the attention to the micro gesture of pinching the index finger and improves the recognition rate. We also compare two other gesture recognition methods: EGR [33] and STDP-CSNN [34]. EGR is early recognition framework to achieve reliable accuracy in the early stages of gesture movement. In STDP-CSNN, Radar range-Doppler image data are encoded into spike sequences with CSNN, and the leaky integrate-and-fire (LIF) model is employed as a neuron in the network nodes. We can see that the algorithm proposed in this paper still has better performance. In addition, the selection of different RDM channels increase in the angular information of the gesture, which is helpful for the symmetric gesture of palm flip. Moreover, we also conducted ablation experiments with MSFF w/o MS, that is, the feature fusion algorithm without multiscale transformation, and the experimental results show that the recognition effect of MSFF w/o MS is slightly lower than that of MSFF for micro gestures. Table 7 shows the comparison of the accuracy, recall and F1-score values for different gestures on the Soli dataset, which also shows that the proposed algorithm has better performance.

**Table 6:** Recognition precision of different gesture recognition algorithms on the Google Soli dataset

| Algorithms | Gesture recognition precision (%) | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | (0) | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) | Avg |
| ETE [25] | 79.2 | 74.4 | 95.6 | 100 | 97.6 | 94.8 | 100 | 100 | 100 | 100 | 94.1 | 94.2 |
| EGR [33] | 70 | 100 | 76 | 72 | 86 | 100 | 100 | 99 | 97 | 99 | 95 | 90.36 |
| STDP-CSNN [34] | 86 | 100 | 95 | 83 | 91 | 98 | 98 | 99 | 84 | 92 | 99 | 93.18 |
| MSFF w/o MS | 93.7 | 100 | 91.6 | 87.0 | 100 | 98.9 | 98.9 | 100 | 100 | 93.7 | 96.8 | 96.4 |
| **MSFF (Ours)** | **95.4** | **88.2** | **98.7** | **100** | **100** | **100** | **100** | **100** | **100** | **99.2** | **96.2** | **98.0** |

**Table 7:** Accuracy, recall and F1-score of different gesture recognition algorithms on the Google Soli dataset

| Algorithms | Accuracy | Recall | F1-score |
|---|---|---|---|
| ETE [25] | 95.15 | 93.81 | 93.78 |
| MSFF w/o MS | 96.38 | 96.43 | 96.42 |
| **MSFF (Ours)** | **97.98** | **97.98** | **97.97** |

### *5.5 Recognition Performance Comparison on the Fudan University Gesture Dataset [26]*

We compare the multiscale feature fusion gesture recognition network architecture (MSFF) proposed in this paper with MFRL, a multidimensional feature representation and learning gesture recognition network model proposed by Xia et al. [26] on the Fudan University gesture dataset. The number of epochs are both set to 50, and the dataset is divided in the ratio of training set: test set = 8:2. The experimental results show that the average accuracy of MSFF in terms of classifying the eight

gestures is 99.1%, which is slightly higher than the experimental result of MFRL (98.9%), and the recognition results of each specific gesture are shown in Table 8. There are some differences between MSFF and MFRL in the two gesture recognition results of Gesture 1 (slide down) and Gesture 5 (finger circle), and the rest of the gesture recognition rates are the same. The main reason is that MFRL fuses DTM, HATM and VATM into a whole input, which makes full use of the gesture angle information, and the addition azimuthal information can better distinguish confusing gestures such as upward and downward gestures, and the recognition effect is slightly higher than that of MSFF. For the recognition effect of micro gestures such as double finger snaps, MSFF is better than MFRL because the multiscale feature fusion method increases the attention to micro gestures. The MSFF w/o MS is also validated on this dataset, and the experimental results show that the recognition effect of MSFF is slightly lower than that of MSFF for micro gestures, which again validates the effectiveness of the multiscale feature fusion idea for micro gesture action recognition.

**Table 8:** Recognition precision of different gesture recognition algorithms on the Fudan dataset

| Algorithms | Gesture recognition precision (%) | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | (0) | (1) | (2) | (3) | (4) | (5) | (6) | (7) | Avg |
| ETE [25] | 96.3 | 100 | 98.8 | 100 | 100 | 98.8 | 100 | 97.5 | 98.9 |
| MSFF w/o MS | 98.8 | 98.8 | 100 | 95.0 | 95.0 | 98.8 | 96.2 | 97.5 | 97.5 |
| **MSFF (Ours)** | **98.7** | **97.5** | **100** | **100** | **100** | **100** | **100** | **97.4** | **99.1** |

The comparison results of the accuracy, recall and F1-scores on the Fudan dataset is shown in Table 9, from which the MSFF gesture recognition algorithm proposed in this paper also has superior performance.

**Table 9:** Accuracy, recall and F1-score of different gesture recognition algorithms on the Fudan dataset

| Algorithms | Accuracy | Recall | F1-score |
|---|---|---|---|
| ETE [25] | 98.90 | 98.91 | 98.91 |
| MSFF w/o MS | 97.50 | 97.50 | 97.51 |
| **MSFF (Ours)** | **99.06** | **99.06** | **99.06** |

In summary, the comparison of all algorithms reveals that only single features or multidimensional features are used as the input of CNN for radar gesture recognition, which is only sensitive to palm motions and does not focus on local finger movements. In addition, it is difficult to extract subtle finger movement features, which easily leads to low gesture recognition accuracy. The multiscale feature fusion idea in this algorithm combines the advantages of the strong semantic information characterization ability of high-level networks and the geometric detail information characterization ability of low-level networks in CNNs. The experimental results of micro gestures show the superiority of this method. In addition, the DTM and RAM containing gesture trajectory, velocity and angle information after gesture feature preprocessing can enhance the recognition of confusing gestures, and the generalization of the network model also yields advantageous results.

## 6 Conclusion

In this paper, we propose a multiscale feature fusion-based gesture recognition algorithm. Specifically, we use multiangle fused DTM and gesture trajectory RAM as the input of the multiscale feature fusion network. We combine the advantages of CNN, i.e., the high-level network with strong semantic information representation capability and the low-level network with strong geometric detail information representation capability and use the multiangle fused DTM and RAM for scaling size as the feature input of the CNN network to make the network achieve better performance. In the future, we will consider using the cross-supervised network model to build a motion model for accurate gesture tracking and trying to implement practical applications of millimeter wave gesture recognition.

**Acknowledgement:** The authors would like to thank the editors and reviewers for the valuable comments and suggestions.

**Funding Statement:** This work is partly supported by the National Natural Science Foundation of China under grant no. 62272242.

**Author Contributions:** The authors confirm contribution to the paper as follows: Study conception and design: Lingsheng Li, Chong Han; data collection: Weiqing Bai; analysis and interpretation of results: Lingsheng Li, Weiqing Bai; draft manuscript preparation: Lingsheng Li; supervision: Chong Han. All authors reviewed the results and approved the final version of the manuscript.

**Availability of Data and Materials:** The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

**Ethics Approval:** Not applicable.

**Conflicts of Interest:** The authors declare that they have no conflicts of interest to report regarding the present study.

## References

[1] W. Xu, "Toward human-centered AI: A perspective from human-computer interaction," *Interactions*, vol. 26, no. 4, pp. 42–46, 2019. doi: 10.1145/3328485.

[2] L. Guo, Z. Lu, and L. Yao, "Human-machine interaction sensing technology based on hand gesture recognition: A review," *IEEE Trans. Hum. Mach. Syst.*, vol. 51, no. 4, pp. 300–309, 2021. doi: 10.1109/THMS.2021.3086003.

[3] C. Osimani, J. J. Ojeda-Castelo, and J. A. Piedra-Fernandez, "Point cloud deep learning solution for hand gesture recognition," *Int. J. Interact. Multimed. Artif. Intell.*, vol. 8, no. 4, pp. 78–87, 2023. doi: 10.9781/ijimai.2023.01.001.

[4] R. Tchantchane, H. Zhou, S. Zhang, and G. Alici, "A review of hand gesture recognition systems based on noninvasive wearable sensors," *Adv. Intell. Syst.*, vol. 5, no. 10, 2023, Art. no. 2300207. doi: 10.1002/aisy.202300207.

[5] Z. Xu *et al.*, "A novel SE-CNN attention architecture for sEMG-based hand gesture recognition," *Comput. Model. Eng. Sci.*, vol. 134, no. 1, pp. 157–177, 2023. doi: 10.32604/cmes.2022.020035.

[6] J. Lin *et al.*, "Overview of 3D human pose estimation," *Comput. Model. Eng. Sci.*, vol. 134, no. 3, pp. 1621–1651, 2023. doi: 10.32604/cmes.2022.020857.

[7] A. Coppens, J. Hermen, L. Schwartz, C. Moll, and V. Maquil, "Supporting mixed-presence awareness across wall-sized displays using a tracking pipeline based on depth cameras," *Proc. ACM Hum.-Comput. Interact.*, vol. 8, pp. 1–32, 2024, Art. no. 260. doi: 10.1145/3664634.

[8] J. Shin, M. Hasan, M. Maniruzzaman, T. Watanabe, and I. Jozume, "Dynamic hand gesture-based person identification using leap motion and machine learning approaches," *Comput. Mater. Contin.*, vol. 79, no. 1, pp. 1205–1222, 2024. doi: 10.32604/cmc.2024.046954.

[9] S. Lee, N. Kini, W. Peng, C. Ma, and J. Hwang, "HuPR: A benchmark for human pose estimation using millimeter wave radar," in *2023 IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Waikoloa, HI, USA, Jan. 3–7, 2023, pp. 5704–5713. doi: 10.1109/WACV56688.2023.00567.

[10] D. Kajiwara and K. Murao, "Gesture recognition method with acceleration data weighted by sEMG," in *Adjunct Proc. 2019 ACM Int. Joint Conf. Pervasive Ubiquitous Comput. Proc. 2019 ACM Int. Symp. Wearable Comput.*, London, UK, Sep. 11–13, 2019, pp. 741–745. doi: 10.1145/3341162.3345589.

[11] J. Qi, G. Jiang, G. Li, Y. Sun, and B. Tao, "Intelligent human-computer interaction based on surface EMG gesture recognition," *IEEE Access*, vol. 7, pp. 61378–61387, 2019. doi: 10.1109/ACCESS.2019.2914728.

[12] M. F. Qureshi, Z. Mushtaq, M. Z. Rehman, and E. N. Kamavuako, "Spectral image-based multiday surface electromyography classification of hand motions using CNN for human-computer interaction," *IEEE Sens. J.*, vol. 22, no. 21, pp. 20676–20683, 2022. doi: 10.1109/JSEN.2022.3204121.

[13] J. Xiao, H. Li, M. Wu, H. Jin, M. J. Deen and J. Cao, "A survey on wireless device-free human sensing: Application scenarios, current solutions, and open issues," *ACM Comput. Surv.*, vol. 55, no. 5, pp. 1–35, Dec. 2022. doi: 10.1145/3530682.

[14] F. Zhang *et al.*, "Embracing consumer-level UWB-equipped devices for fine-grained wireless sensing," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 6, no. 4, pp. 1–27, 2023. doi: 10.1145/3569487.

[15] D. Salami, R. Hasibi, S. Palipana, P. Popovski, T. Michoel and S. Sigg, "Tesla-Rapture: A lightweight gesture recognition system from mmWave radar sparse point clouds," *IEEE Trans. Mob. Comput.*, vol. 22, no. 8, pp. 4946–4960, 2022. doi: 10.1109/TMC.2022.3153717.

[16] B. Jin, X. Ma, B. Hu, Z. Zhang, Z. Lian and B. Wang, "Gesture-mmWAVE: Compact and accurate millimeter-wave radar-based dynamic gesture recognition for embedded devices," *IEEE Trans. Hum. Mach. Syst.*, vol. 54, no. 3, pp. 337–347, 2024. doi: 10.1109/THMS.2024.3385124.

[17] S. Ahmed, K. D. Kallu, S. Ahmed, and S. H. Cho, "Hand gestures recognition using radar sensors for human-computer-interaction: A review," *Remote Sens.*, vol. 13, no. 3, 2021, Art. no. 527. doi: 10.3390/rs13030527.

[18] A. Alteaimi and M. T. B. Othman, "Robust interactive method for hand gestures recognition using 639 machine learning," *Comput. Mater. Contin.*, vol. 72, no. 1, pp. 577–595, 2022. doi: 10.32604/cmc.2022.023591.

[19] Y. Zhang, Z. Yang, G. Zhang, C. Wu, and L. Zhang, "XGest: Enabling cross-label gesture recognition with RF signals," *ACM Trans. Sens. Netw.*, vol. 17, no. 4, pp. 1–23, 2021. doi: 10.1145/3458750.

[20] H. Liu *et al.*, ""mTransSee: Enabling environment-independent mmWave sensing based gesture recognition via transfer learning," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 6, no. 1, pp. 1–28, 2022. doi: 10.1145/3571588.

[21] P. Molchanov, S. Gupta, K. Kim, and K. Pulli, "Multi-sensor system for driver's hand-gesture recognition," in *2015 11th IEEE Int. Conf. Workshops Automatic Face Gesture Recognit. (FG)*, Ljubljana, Slovenia, IEEE, May 4–8, 2015, pp. 1–8. doi: 10.1109/FG.2015.7163132.

[22] Y. Wang, Y. Shu, X. Jia, M. Zhou, L. Xie and L. Guo, "Multifeature fusion-based hand gesture sensing and recognition system," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2021. doi: 10.1109/LGRS.2021.3086136.

[23] J. Yu, L. Yen, and P. Tseng, "mmWave radar-based hand gesture recognition using range-angle image," in *2020 IEEE 91st Veh. Technol. Conf. (VTC2020-Spring)*, IEEE, May 25–31, 2020, pp. 1–5. doi: 10.1109/VTC2020-Spring48590.2020.9128573.

[24] J. Zhang, J. Tao, and Z. Shi, "Doppler-radar based hand gesture recognition system using convolutional neural networks," in *Int. Conf. Commun., Signal Process., Syst.*, Harbin, China, Springer, Jul. 14–16, 2017, pp. 1096–1113. doi: 10.1007/978-981-10-6571-2_132.

[25] 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 *Proc. 29th Annu. Symp. User Interface Softw. Technol.*, 2016, pp. 851–860. doi: 10.1145/2984511.298456.

[26] Z. Xia, Y. Luomei, C. Zhou, and F. Xu, "Multidimensional feature representation and learning for robust hand-gesture recognition on commercial millimeter-wave radar," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 6, pp. 4749–4764, 2020. doi: 10.1109/TGRS.2020.3010880.

[27] Y. Zhang, S. Dong, C. Zhu, M. Balle, B. Zhang and L. Ran, "Hand gesture recognition for smart devices by classifying deterministic doppler signals," *IEEE Trans. Microw. Theory Tech.*, vol. 69, no. 1, pp. 365–377, 2020. doi: 10.1109/TMTT.2020.3031619.

[28] H. Liu *et al.*, "*M-Gesture*: Person-independent real-time in-air gesture recognition using commodity millimeter wave radar," *IEEE Internet Things J.*, vol. 9, no. 5, pp. 3397–3415, 2021. doi: 10.1109/JIOT.2021.3098338.

[29] Y. Sun, T. Fei, X. Li, A. Warnecke, E. Warsitz and N. Pohl, "Real-time radar-based gesture detection and recognition built in an edge-computing platform," *IEEE Sens. J.*, vol. 20, no. 18, pp. 10706–10716, 2020. doi: 10.1109/JSEN.2020.2994292.

[30] Y. Sun, T. Fei, X. Li, A. Warnecke, E. Warsitz and N. Pohl, "Multi-feature encoder for radar-based gesture recognition," in *2020 IEEE Int. Radar Conf. (RADAR)*, Washington, DC, USA, IEEE, Apr. 28–30, 2020, pp. 351–356. doi: 10.1109/RADAR42522.2020.9114664.

[31] L. Gan, Y. Liu, Y. Li, R. Zhang, L. Huang and C. Shi, "Gesture recognition system using 24 GHz FMCW radar sensor realized on real-time edge computing platform," *IEEE Sens. J.*, vol. 22, no. 9, pp. 8904–8914, 2022. doi: 10.1109/JSEN.2022.3163449.

[32] TI, "IWR6443 Single-Chip 60- to 64-GHz mmWave Sensor," 2021. Accessed: Jul. 1, 2024. [Online]. Available: https://www.ti.com/product/IWR6843

[33] R. Min, X. Wang, J. Zou, J. Gao, L. Wang and Z. Cao, "Early gesture recognition with reliable accuracy based on high-resolution IoT radar sensors," *IEEE Internet Things J.*, vol. 8, no. 20, pp. 15396–15406, 2021. doi: 10.1109/JIOT.2021.3072169.

[34] Y. Wu, L. Wu, Z. Xiao, and T. Hu, "Spiking-timing-dependent plasticity convolutional spiking neural network for efficient radar-based gesture recognition," in *2023 Int. Conf. Image Process., Comput. Vis. Mach. Learn. (ICICML)*, Chengdu, China, Nov. 3–5, 2023, pp. 3–5. doi: 10.1109/ICICML60161.2023.10424838.