APSIPA Transactions on Signal and Information Processing, 2024, 13, e304

Original Paper

A Lightweight Remote Gesture Recognition System with Body-motion Suppression and Foreground Segmentation Using FMCW Radar

Jingxuan Chen[1], Yajie Wu[2], Bo Zhang[2], Shisheng Guo[1,2*] and Guolong Cui[1,2]

[1] *Yangtze Delta Region Institute, University of Electronic Science and Technology of China, Quzhou 324009, China*
[2] *School of Information and Communication Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China*

ABSTRACT

In remote dynamic hand-gesture recognition, uncertainties in timing and distance of gesture occurrences, coupled with the subtle bodily perturbations induced by arm movements, pose substantial challenges to the accurate extraction of gesture features. In this paper, we propose a lightweight real-time gesture recognition system based on support vector machines. By analyzing the Doppler features of different motion states, a Doppler weighting factor was constructed to suppress bodily micro-motion interference in the range-time spectrum, and achieve foreground extraction of gesture signals concurrently. Furthermore, prior to the extraction of HOG features, we employ Gaussian filtering to suppress abrupt transitions and noise inherent in the gesture signals. This preprocessing significantly enhances the stability of feature extraction. Subsequently, the extracted features are input into an SVM for training and classification. Experimental results demonstrate

*Corresponding author: Shisheng Guo, ssguo@uestc.edu.cn. This work was supported in part by the Municipal Government of Quzhou under Grant 2022D008, Grant 2022D005 and Grant 2023D032, and in part by the 111 Project B17008.

Received 31 August 2023; Revised 21 January 2024
ISSN 2048-7703; DOI 10.1561/116.00000061

that, for five distinct gestures exhibited in two different states — standing and seated — within a range of 1 to 5 meters, the recognition accuracy reaches 96%. This proves the feasibility of the proposed methodology, and its potential to realize real-time gesture recognition.

Keywords: FMCW radar, hand gesture recognition, machine learning, lightweight

1 Introduction

As an intuitive human-computer interaction (HCI) technique, remote gesture recognition has great potential in various applications, such as smart home control [17] and human behavior recognition [7], garnering widespread attention in the research field. Various gesture recognition techniques ranging from camera [1], wearable gloves [8], ultrasound [2] and millimeter wave radar [9, 10, 14] have been studied. Compared to traditional wearable gloves, radar has the advantages of non-contact and can recognize gestures without touching the users. Compared with the gesture recognition of the camera, millimeter wave radar has a lower cost and can protect the privacy of users well [15]. Leveraging the azimuth and elevation resolution capabilities of multi-transmit multi-receive radar systems allows for the acquisition of point cloud information from targets. The distinctive characteristics of point cloud features across diverse gestures enable gesture recognition. Literature [16] has proposed a mobile scatter center model to represent three-dimensional point clouds, and devised a multi-channel, three-layer convolutional neural network (CNN) for the learning and classification of multi-dimensional gesture features, ultimately achieving a remarkable peak classification accuracy of 98.9%. In another study, Salami *et al.* [12] point cloud information was employed as input and subjected to message-passing neural networks for classification and learning. This approach attained a classification accuracy of 98.1% for 21 gestures at distances ranging from 1.5 to 5 meters, demonstrating efficacy. The model could be deployed on a Raspberry Pi 4 for real-time recognition, albeit with inference times of 0.3 to 0.4 seconds per instance. Beyond point cloud utilization, some scholars have explored gesture recognition at long distances based on spectrogram data derived from radar signal processing. By capitalizing on variations in distances across distinct gestures, a study by Suh *et al.* [13] employed distance-profile images as classification features. Employing a three-dimensional convolutional neural network (3-D CNN) and Long-Short Term Memory (LSTM), the approach achieved a 96% accuracy rate for eight gestures at a working distance of 1.5 meters. Moreover, Dong *et al.* [5] utilize

short-time Fourier transforms to generate feature spectrograms for six gestures. Classic VGG16 networks were then employed for feature extraction, followed by the application of traditional machine learning methods for classification, yielding accuracy surpassing 96%. In summary, the majority of radar-based remote gesture recognition research has revolved around intricate deep learning algorithms.

Although many attempts have been made to streamline models, achieving real-time performance on compact embedded systems remains a significant challenge. This predicament escalates the practical application costs and hampers deployment within real-world scenarios. Therefore, there is an urgent need to delve into lightweight recognition methods to mitigate the application costs associated with this technology. In response to the constraints posed by processor and memory resources, along with the challenge of extracting angular and Doppler spectral information from low signal-to-noise ratio long-range echoes, this paper introduces a real-time recognition approach that exclusively extracts features from the range-time spectrum. Through the analysis of Doppler features corresponding to distinct motion states, a construction of Doppler weighting factors is proposed to mitigate the influence of bodily micro-movements on gesture actions, concurrently achieving foreground extraction of gesture feature spectra. Subsequently, the gesture foreground images are subjected to Gaussian filtering, then the HOG algorithm is used for feature extraction and Support Vector Machine (SVM) is used for gesture recognition. Ultimately, the feasibility of the proposed algorithm is validated through experimental verification. The contributions of this study can be summarized as follows:

- We propose a machine learning-based lightweight remote gesture recognition system. By segmenting and extracting gesture spectrum, SVM algorithm with smaller model parameters can be used to accurately recognize dynamic gestures of people at different distances and postures.
- Constructing Doppler weighting factors based on distinct motion characteristics enables the suppression of bodily micro-movements while retaining gesture motion information. A foreground extraction algorithm is designed on the basis of the factor, effectively reducing subsequent feature extraction computational complexity while enhancing feature accuracy.
- Gaussian filtering is applied to the extracted range-time Spectrum, curtailing data fluctuations and noise interference. This operation enhances the stability of features extracted through Histogram of Oriented Gradients (HOG), consequently further augmenting the precision of recognition.

The rest of this paper is organized as fallows: Section 2 introduces the details of the proposed method, including FMCW radar signal model, feature extraction and gesture recognition method. Section 3 presents the experimental design, results and analysis. Finally, we provide the conclusion of this paper in Section 4.

2 Methodology

2.1 FMCW Radar Signal Pre-processing

The FMCW radar system utilized in the paper is illustrated in Figure 1. The synth is employed to generate a FMCW signal, which consists of multiple frames, with each frame comprising multiple chirps. Three transmitting antennas are utilized for transmitting the FMCW signal, while four receiving antennas are used to capture the echo signals reflected by the target. The transmitted and received signals are mixed together via a mixer to obtain the analog intermediate frequency (IF) signal. Subsequently, the analog IF signal is processed by an analog-to-digital converter to obtain the digital IF signal.

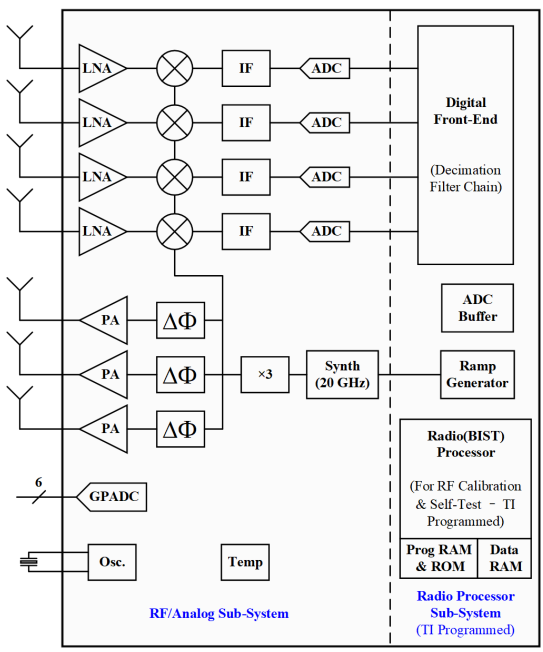


Figure 1: The structure of FMCW radar system.

The FMCW signal from the transmitting antenna can be expressed as

$$S_T(t) = A_T \cos\left(2\pi\left(f_c t + \frac{1}{2}\mu t^2\right) + \varphi_0\right) \tag{1}$$

where A_T is the amplitude of transmitted signal, f_c indicates the start frequency, $\mu = B/T$ denotes the frequency modulation slope, B is the frequency modulation bandwidth, T is the frequency modulation period and φ_0 signifies the initial phase of transmitted signal.

Echo signal received by the receiving antenna can be expressed as

$$S_R(t) = A_R \cos\left(2\pi\left(f_c(t-\tau) + \frac{1}{2}\mu(t-\tau)^2\right) + (\varphi_0 + \Delta\varphi)\right) \tag{2}$$

A_R represents the amplitude of received signal, $\tau = 2(R_0 + v_r t)/c$ indicates the time delay of received signal, R_0 represents the initial range of target to radar. c is the velocity of light and $\Delta\varphi = 4\pi v_r t/\lambda$ is the phase shift.

The transmitted FMCW signal and the received echo signal are passed through a mixer to simplify the sampling operation. Subsequently, the low-frequency signal is selected as the IF signal. This process can be expressed as

$$S_{IF}(t) = A_{IF} \cos\left(2\pi\left(\mu\tau t - \frac{1}{2}\mu\tau^2\right) - \Delta\varphi\right) \tag{3}$$

where A_{IF} is the amplitude of IF signal. Substituting the time delay τ into (3), the IF signal can be expressed as

$$S_{IF} = A_{IF} \cos\left(2\pi\left(\mu\frac{2(R_0 + v_r t)}{c}t - \frac{1}{2}\mu(2(R_0 + v_r t)/c)^2\right) - \Delta\varphi\right) \tag{4}$$

Since τ is very small, the higher power of τ can be ignored. Therefore, the frequency of IF signals can be simplified as

$$f_{IF} = \mu\tau = \mu \times \frac{2(R_0 + v_r t)}{c} = R \times \frac{2\mu}{c} \tag{5}$$

According to (6), the range of target relative to the radar can be calculated by

$$R = f_{IF} \times \frac{c}{2\mu} = \frac{cTf_{IF}}{2B} \tag{6}$$

The range resolution of the radar is defined as the minimum interval between two adjacent targets that can be expressed as

$$r_{res} = \frac{c}{2B} \tag{7}$$

Assuming that a target's radial velocity does not change during a chirp period, the phase difference in each echo caused by the target motion can be expressed as $\omega = (4\pi\Delta r)/\lambda = (4\pi v \cdot T)/\lambda$. To measure an unambiguous velocity, $|\omega| \leq \pi$ should be met, so that the maximum unambiguous velocity is

$$v_{\max} = \frac{\lambda}{4T} \tag{8}$$

where λ is the wavelength corresponding to the start frequency f_c.

Assume that the target and radar equipment are far enough away, the rays of the echo reaching the receiving antenna are parallel to each other. The arrival time difference of two adjacent Rx can be calculated as $\Delta t = d_r \sin\theta/c$, and the phase difference δ can be expressed as

$$\delta = \frac{2\pi d_r \sin\theta}{\lambda} \tag{9}$$

where θ is the azimuth. To measure an unambiguous azimuth, $|\delta| < \pi$ should be met. When $d_r = \lambda/2$, the maximum field of view can be achieved. Then the azimuth of arrival can be expressed as

$$\theta = \sin^{-1}\left(\frac{\delta}{\pi}\right) \tag{10}$$

After mixing the received FMCW signal to obtain the IF signal, FFT is performed in the fast time dimension and the slow time dimension respectively, also known as 2D FFT, as shown in the following equation

$$S(p,q,t) = \sum_{l=0}^{L}\left(\sum_{n=0}^{N} s(n,l,t)\, e^{-j2\pi pn/N}\right) e^{-j2\pi ql/L} \tag{11}$$

where $s(n,l,t)$ is the beat signal, which is first transformed, corresponding to the transmitted chirp signal and this signal is transformed to the frequency domain to obtain Doppler–FFT expressed in $S(p,q,t)$.

Then a Range–Doppler Matrix can be obtained from $S(p,q,t)$, as shown in the following equation

$$RD(r,v,t) = \left|S\left(\frac{r}{\Delta r_f}, \frac{v}{\Delta v_f}, t\right)\right| \tag{12}$$

2.2 Feature Extraction Method

2.2.1 Gesture Recognition Algorithm Overview

The remote gesture recognition process proposed in this paper can be divided into three segments: data preprocessing, feature extraction, model training

and classification, as shown in Figure 2. The following is a comprehensive exposition of each component of the algorithm.

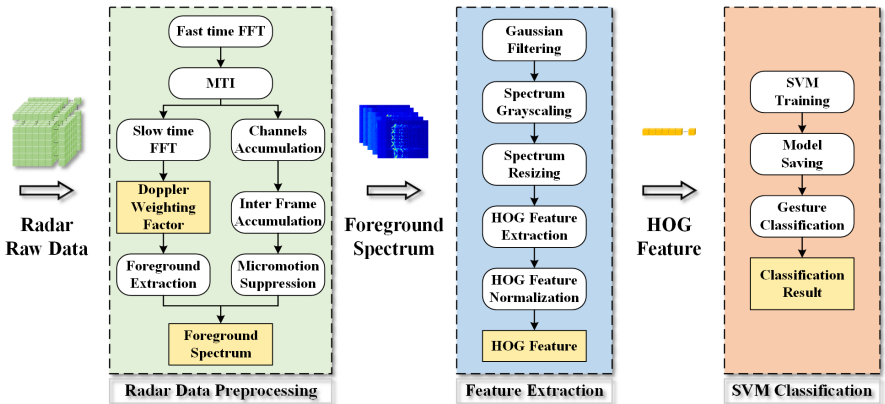


Figure 2: Gesture recognition processing chain.

Data Preprocessing: The data undergoes radar signal processing, encompassing distance and velocity dimension FFT operations. The static clutter suppression is employed to mitigate interference from stationary scenes, and two pulse cancellation is used to achieve Moving Target Indication (MTI). In this paper, we introduce a Doppler weighting technique to counteract bodily micro-movement interference. This method successfully mitigates disturbances caused by bodily micro-movements during gesture execution, consequently enhancing recognition accuracy. Furthermore, Doppler-weighting coefficients are employed to realize dynamic gesture foreground extraction, leading to the segmentation of range-time spectra produced by gesture motions.

Feature Extraction: The spectra obtained from data preprocessing are subjected to feature extraction. This involves dimension reduction of image data, yielding one-dimensional feature vectors, thereby reducing the input volume for classification and recognition algorithms. In this study, the incorporation of Gaussian filtering techniques prior to HOG feature extraction significantly enhances the effectiveness of dynamic gesture feature extraction.

Model Training and Recognition: Leveraging the SVM algorithm from the realm of machine learning, the feature vectors obtained from preprocessing undergo model training, subsequently facilitating gesture recognition and classification.

2.2.2 Body-motion Suppression Algorithm

When an individual performs gesture actions in a standing posture, the swinging of the arms induces slight bodily oscillations, generating a low-frequency interference signal. Due to the significant difference in scattering area between the body and the arms, the resulting echoes from the body overpower those from the arms. This exerts considerable influence on subsequent feature extraction. While two-pulse cancellations can suppress static clutter and exhibit certain inhibitory effects on low-frequency signals, the motion of continuous waving gestures at far distances is characterized by relatively slow radial velocities in relation to the radar. Although slightly higher in frequency than bodily shaking, these gestures still constitute low-frequency signals. Consequently, designing Moving Target Indication (MTI) filters for both suppressing bodily micro-movements and retaining gesture signals proves challenging.

In this paper, we propose a Doppler weighting algorithm to suppress bodily micro-movements while retaining echoes generated by waving gestures. The underlying principle of this algorithm is rooted in the disparate Doppler frequency characteristics corresponding to body and gesture motions. Specifically, gesture actions result in higher Doppler frequencies owing to their rapid radial velocities, whereas bodily micro movements engender lower Doppler frequencies due to slower velocities. By leveraging this inherent distinction, a weighting factor is formulated. This factor magnifies the distance image associated with the gesture component and diminishes the distance image linked to bodily micro movements. As a consequence, interference generated by bodily micro movements is mitigated. Upon analyzing collected gesture signals, as shown in Figure 3, the following observations are made:

(1) Gesture actions entail higher speeds, leading to Doppler amplitude curves with peak positions skewed toward the ends, exhibiting asymmetric distributions.

(2) Due to the relatively low velocities of bodily micro-movements, Doppler amplitude curves feature larger values around zero, showcasing more pronounced symmetrical distributions.

Based on the above analysis, by incorporating the peak position of the Doppler amplitude curve and the cumulative values around zero into the weighting factor formula, we introduce the Doppler weighting factor calculation formula in this paper:

$$W_d(m) = k_p \cdot W_{\max}(m) + (1 - k_p) \cdot W_{amp}(m) \tag{13}$$

$W_d(m)$ denotes the value of the weighting factor corresponding to the m distance bin. k_p represents an empirical parameter constrained within the range of $[0, 1]$. It serves as a proportionality factor employed to harmonize

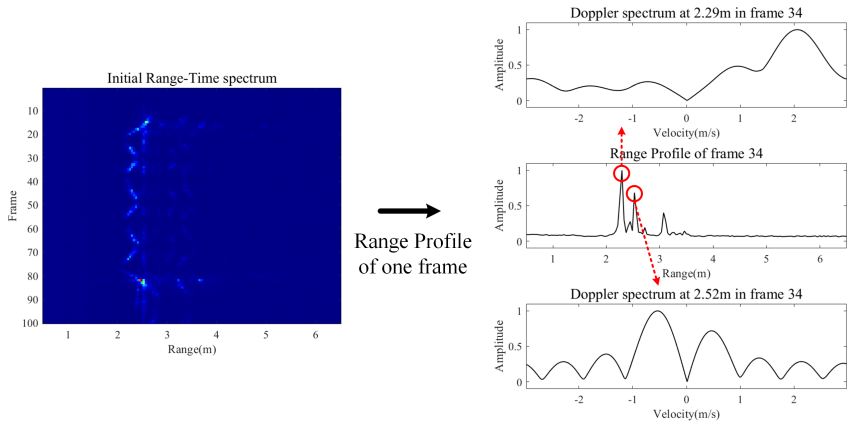


Figure 3: The difference between Doppler generated by bodily micro-movements and gesture movements.

the weighting of peak position and amplitude accumulation. In this study, in order to ensure equal consideration of the impact on the two weighting factors, we set the k_p value to 0.5. $W_{\max}(m)$ and $W_{amp}(m)$ denote the weighting of peak position and amplitude accumulation, respectively, for the m distance bin. The explicit formulations for these parameters are provided as follows:

$$\begin{cases} W_{\max}(m) = \dfrac{2|v_{\max p}|}{v_{\max}} \cdot \dfrac{\max(z) \cdot I}{\sum_{i=0}^{I-1} z_i} \\ W_{amp}(m) = \max\left(\dfrac{\sum_{i=0}^{I/2-1} z_i}{\sum_{i=I/2+1}^{I-1} z_i}, \dfrac{\sum_{i=I/2+1}^{I-1} z_i}{\sum_{i=0}^{I/2-1} z_i} \right) \end{cases} \tag{14}$$

For the m distance bin, a fast Fourier transform(FFT) to the slow time dimension is performed, resulting in the Doppler amplitude sequence z. $v_{\max p}$ signifies the velocity corresponding to the peak of the Doppler amplitude sequence, $v_{\max}$ represents the maximum unambiguous velocity, and $\max(\cdot)$ signifies the chosen maximum. The formulation of $W_{\max}(m)$ is primarily grounded in the swift nature of gesture motion, where larger values of $v_{\max p}$ correspond to faster velocities. Furthermore, to account for the influence of noise, the ratio of the maximum value to the mean is introduced to prevent noise amplification. Regarding the formulation of $W_{amp}(m)$, it is mainly based on the uneven distribution of the Doppler amplitude curve of gesture movements around the zero point, while the bodily micro-motion Doppler amplitude curve does not have the above characteristics, so the gesture part can also be enhanced. As shown in Figure 4, the bodily micro- motion interference in the processed distance time spectrum has been effectively suppressed.

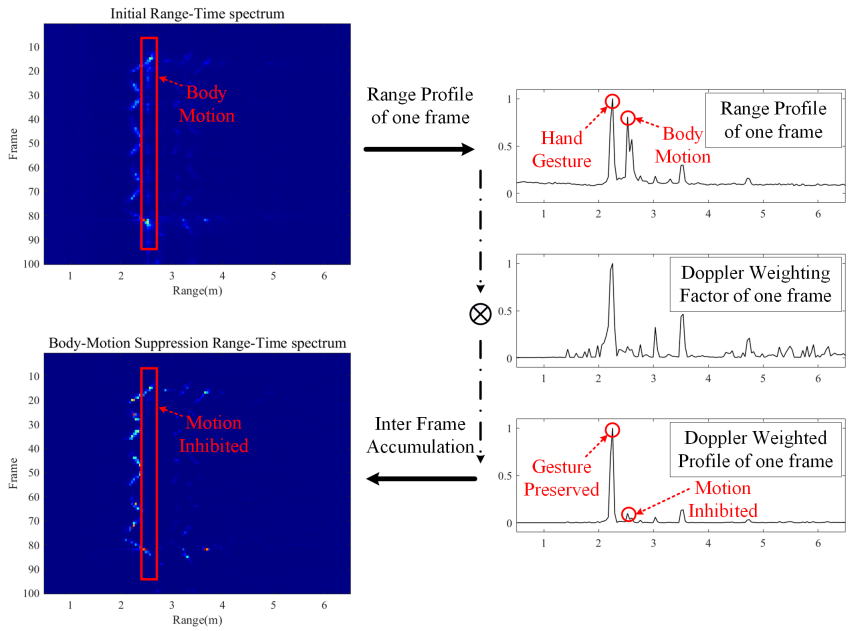


Figure 4: Doppler weighting algorithm processing flow.

2.2.3 Foreground Segmentation Algorithm

In remote gesture recognition, the time and distance of gesture actions are uncertain, which leads to significant differences in the range-time(RT) spectrum of the same gesture in both time and distance dimensions. The design of recognition algorithms must address this diversity to accurately extract distinct gesture features. One approach involves collecting a substantial number of samples and performing feature extraction on the entire image. This strategy aims to ensure that gesture distance-profile images, regardless of variations in time and distance, are correctly recognized. However, this method presents challenges. On one hand, it increases the length of feature vectors required to represent a broader range of images, thereby augmenting the number of parameters in the final recognition model, which is not conducive to the lightweight of the model. On the other hand, due to the presence of features that do not belong to gestures, it will inevitably increase the recognition difficulty of the classifier, resulting in a decrease in recognition accuracy.

To address the aforementioned challenges, this paper introduces a foreground segmentation algorithm based on Doppler weighting factors. Through statistical analysis of the Doppler amplitude accumulation values within the weighting factor, this approach facilitates the determination of the starting

and ending positions of dynamic gestures. Additionally, during the course of a gesture, continual detection of the foremost edge of the gesture enables the spatial and temporal segmentation and extraction of gesture motion, thereby ameliorating the identified issues.

Taking into consideration the Doppler amplitude accumulation coefficient constructed in the preceding section, denoted as $W_{amp}(m)$, which characterizes the Doppler spectral distribution of the distance bin, it can be observed that when this distance bin exhibits substantial amplitude motion, the ratio of $W_{amp}(m)$ will surpass 1 due to the unevenness of the Doppler spectral distribution. Conversely, when only bodily micro-movements or noise is present, the ratio of $W_{amp}(m)$ will tend to approach 1. As a result, by conducting statistical analysis on the Doppler amplitude accumulation coefficient, it becomes feasible to ascertain the presence of gesture actions. The algorithm proposed is delineated as shown in Algorithm 1.

Algorithm 1 Foreground Extraction Algorithm.

Input: Range Profile R_i of size $w \times 1$, Doppler factor W_{amp}, Arm length in Range Profile bins n_{arms}.
Output: Gesture Foreground Spectrum RT_{ext}

Record the W_{amp} difference values of the last 10 frames as W_{dif}.
Calculate standard deviation of W_{dif} and sum up as w_{gate}.
$G_{start} = 0$; $G_{end} = 0$; $G_{front} = w$.
while $G_{start} = 0$ *or* $G_{end} = 0$ **do**
  if $w_{gate} > 1.2$ *and* $G_{start} = 0$ **then**
    Get current frame f_{start}, $G_{start} = f_{start}$.
    Start accumulating Range Profile R_i as RT Spectrum.
  else if $G_{start} > 0$ *and* $G_{end} = 0$ *and* $w_{gate} \geq 1$ **then**
    Find the max value of the R_i position n_{max}.
    Find the front position of gesture n_{front}.
    if $|n_{front} - n_{max}| < n_{arms}$ **then**
      $G_{front} = min(G_{front}, n_{front})$
    end if
  else if $G_{start} > 0$ *and* $G_{end} = 0$ *and* $w_{gate} < 1$ **then**
    Get current frame f_{end}, $G_{end} = f_{end}$.
    End accumulating Range Profile R_i.
  end if
end while
$RT_{ext} = RT[G_{start} : G_{end}, G_{front} : G_{front} + n_{arms}]$.
return RT_{ext}

2.2.4 HOG Feature Extraction With Gauss Filtering

The HOG algorithm, introduced by Dalal and Triggs [4], is an image feature extraction method designed to effectively capture and extract edge contour information from images. It has found wide applications in fields such as object detection. When combined with traditional machine learning methods such as SVM, it enables rapid and accurate classification and recognition of simple images.

Due to the substantial influence of Radar Cross Section (RCS) variations on millimeter-wave radar detection results, significant fluctuations and random fluctuations arise in detection outcomes during gesture motion due to changing body RCS. When the HOG algorithm is directly applied to feature extraction from the distance-time spectrogram obtained from radar signal processing, these intense changes lead to unstable gradient information extraction, consequently yielding subpar classification results.

In classical edge detection algorithms like the Canny algorithm for image edge detection, Gaussian filtering is applied to images as a preliminary step [3]. In image processing, classic filtering operations include mean filtering, Gaussian filtering, and median filtering [6]. Due to the possibility of blurred image details caused by mean filtering, while median filtering is mainly used to remove salt and pepper noise, which is not common in radar signal processing. Recognizing the analogous nature of feature extraction and edge detection from the distance-time spectrogram, this paper employs Gaussian filtering on the distance-time spectrogram obtained from foreground segmentation before conducting HOG feature extraction, enhancing the stability of feature extraction.

As the features along the distance dimension lose distinctiveness following foreground segmentation, the HOG feature extraction method employed in this study partitions only the temporal dimension direction in terms of cells and blocks. Furthermore, acknowledging the potential inconsistency in input feature image sizes due to various algorithmic processes, this paper applies the HOG algorithm with a fixed number of cells, adaptively computing the number of image rows contained in each cell. The workflow of the feature extraction algorithm is depicted below.

2.3 Classification Method

The classifier employed in this study is SVM. SVM is a binary linear classification model that determines a separating hyperplane by solving for the parameters corresponding to the maximum margin between two classes of feature vectors in the feature space, thereby achieving classification [11]. The conceptual illustration of SVM classification is presented as shown in Figure 5, with vectors closest to the hyperplane referred to as support vectors. The

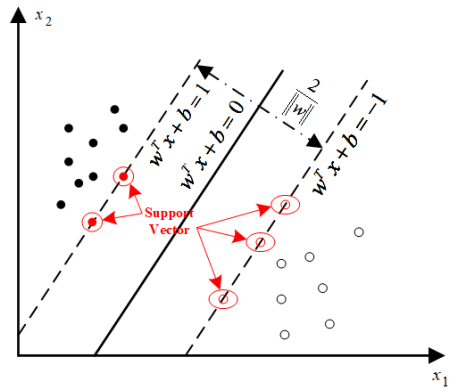


Figure 5: Schematic diagram of SVM.

training process of SVM involves the pursuit of parameters that define the hyperplane.

Suppose w be the normal vector of the hyperplane, and b represent the bias term. Through the constraints imposed by support vectors, the optimal parameters can be determined, maximizing the classification margin. The equation for the SVM decision hyperplane can be expressed as follows:

$$w^T x + b = 0 \tag{15}$$

w is an n dimensional vector parameter, b is a real number, and x represents a feature vector. Based on the above equation, the distance l from an n dimensional space point to a line is given by:

$$l = \frac{\left| w^T x + b \right|}{\| w \|} \tag{16}$$

The objective is to find a set of parameters (w, b) that the distances from the support vectors of the two classes to the hyperplane are maximized. The hyperplane equation passing through the classes y_1 and y_2 can be constrained as:

$$\begin{cases} w^T x + b = 1 & , y = y_1 \\ w^T x + b = -1 & , y = y_2 \end{cases} \tag{17}$$

Subsequently, the classification interval can be calculated as $r = 2/\|w\|$. This further leads to the classification equation:

$$\begin{cases} w^T x + b \geq 1 & , y = 1 \\ w^T x + b \leq -1 & , y = -1 \end{cases} \tag{18}$$

The aim of solving for the optimal hyperplane parameters is to maximize r^2, which is equivalent to minimizing w. This is subject to the constraint:

$$\begin{cases} \min\limits_{w,b} \frac{1}{2}\|w\|^2 \\ \text{s.t.} y\left(w^T x + b\right) \geq 1 \end{cases} \tag{19}$$

Introducing Lagrange multipliers gives rise to the dual problem, through the utilization of the Sequential Minimal Optimization(SMO) algorithm, the hyperplane parameters (w, b) are determined.

From the above analysis, it is obvious that the SVM classifier is applicable solely to the optimization of the optimal separating hyperplane in linearly separable two-class problems. In the case of non-linear problems, the SVM algorithm employs kernel functions to map the data into higher-dimensional spaces, thereby transforming them into linearly separable instances. Commonly used kernel functions include linear, polynomial, and Gaussian kernels. For addressing multi-class problems, SVM resorts to multiple binary classifications and ultimately utilizes a voting mechanism to generate the final classification outcome.

In this paper, the Gaussian kernel is utilized as the kernel function within the SVM. The Gaussian kernel's ability to map original data to an infinite-dimensional feature space is pivotal in effectively capturing intricate data relationships. Additionally, the Gaussian kernel, characterized by a singular parameter, simplifies the model adjustment process. However, the selection of this hyperparameter is critical to the model's performance. Consequently, a five-fold cross-validation method is employed for optimal parameter determination. This approach is instrumental in preventing overfitting and enhancing the model's generalization capacity.

3 Experiment

3.1 Experimental Setup

3.1.1 Radar Parameters

Considering the authentic indoor gesture recognition scenarios and aiming for comprehensive radar coverage, the radar is mounted on a tripod at a height of 2.5 meters with an inclination angle of 30 degrees for data collection. The data collection scenarios are depicted in Figure 6, where the collection actions are performed at distances of 1 meter, 3 meters, and 5 meters ahead to acquire feature spectra at varying distances.

In the radar system, the design of parameters such as ADC sampling rate, number of samples, pulse repetition period, frame period, and chirp slope is

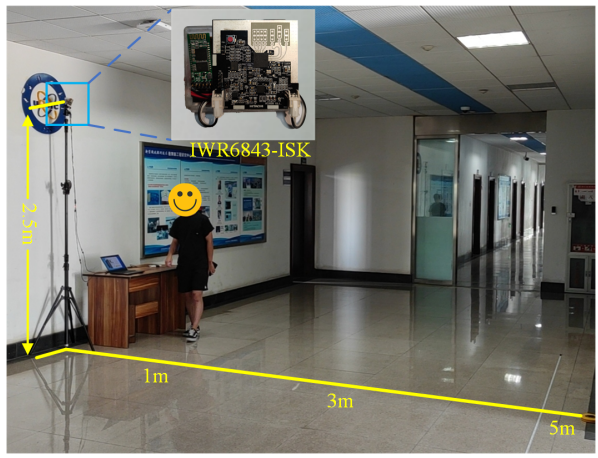


Figure 6: Remote gesture data collection scenario.

crucial to ensure both practicality and fulfillment of the application scenario demands. The considerations are as follows:

(1) The effective range for recognizing remote gestures is within 1 to 5 meters, taking into account the influence of multipath interference.

(2) Based on empirical tests, certain gesture movements exhibit radial velocities in the range of ±2m/s. Consequently, the velocity parameter design necessitates considerations $v_{\max} \geq 4$m/s.

(3) To enhance the detection of subtle arm movements, a higher distance resolution is essential to provide more detailed information.

Considering the above factors, the parameters for remote gesture classification are presented in Table 1. From the parameter indicators in Table 2, it can be seen that the radar parameter is reasonable and can meet the needs of remote application scenarios.

3.1.2 Datasets Introduction

In the context of remote application scenarios, taking into account the practical usage of gestures in real-world settings and easy to distinguish from radar signals, the following five gesture actions were designed. The specific descriptions of these five gestures are as follows:

(1) Cross Hands: Both hands move upward together to the chest level and then cross.

Table 1: Radar parameters for data acquisition.

Parameters	Values
Number of TX antennas	3
Number of RX antennas	4
Frequency modulation rate	30 MHz/us
ADC sampling rate	2000 Ksps
ADC samples per chirp	256
Sampling Bandwidth	3840 MHz
Number of chirps per frame	16
Time duration of per frame	40 ms

Table 2: Remote gesture recognition parameter indicators.

Parameters	Values
Maximum unambiguous Range	10 m
Range resolution	0.0391 m
Maximum unambiguous Velocity	$\pm$ 2.98 m/s
Velocity resolution	0.37 m/s

(2) Wave Hands: One hand is lifted to the head's height, and then it rapidly swings left and right for about two cycles before returning to the hanging position.

(3) Click Three Times: One hand is lifted to the head's height, and then it quickly moves forward and backward three times in front of the radar before returning to the hanging position.

(4) Click Two Times: Similar to the triple tap, but the forward-backward motion is performed only twice.

(5) Click One Time: One hand is lifted to the head's height and then quickly returned to the hanging position. The raising and lowering motions should be coherent.

To capture the genuine distribution of data as effectively as possible, this datasets was collected from 9 participants in two postures, standing and sitting. In total, 800 sets of experimental data were collected, forming the datasets for remote gesture recognition in various poses. The quantities of each gesture category's data are depicted in Table 3.

Table 3: Remote Gesture Datasets overview.

Number	Gesture type	Perform mode	Data size
1	Cross Hands	sit	80
		stand	80
2	Wave Hands	sit	80
		stand	80
3	Click Three Times	sit	80
		stand	80
4	Click Two Times	sit	80
		stand	80
5	Click One Time	sit	80
		stand	80
	Total Data size		800

3.2 Experimental Results

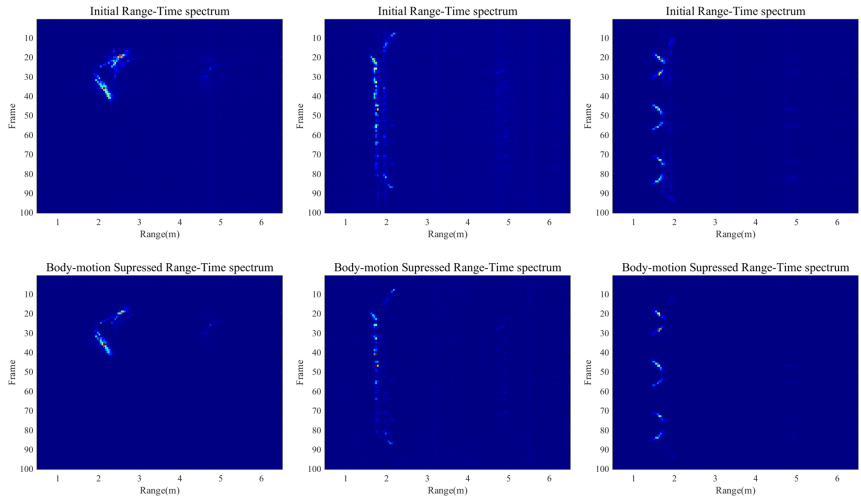
3.2.1 Body-motion Suppression and Foreground Segmentation Results

To validate the effectiveness of the micro-motion suppression algorithm and foreground extraction algorithm, this section analyzes the impact of micro-motion suppression on the R-T spectrum and the subsequent effect of foreground extraction on HOG feature extraction.

When comparing Figure 7, it is evident that the Doppler weighting factors calculated using Equation (13) exhibit substantial magnitudes at positions corresponding to gesture actions and considerably reduced values at locations corresponding to bodily micro-movements. Consequently, the micro-motion interference is effectively suppressed through weighted computation.

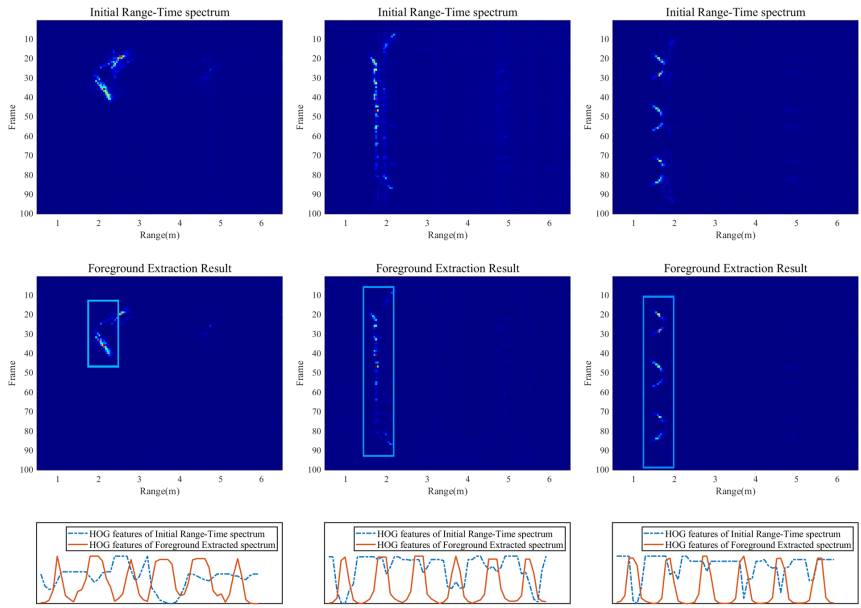
Figure 8 displays the extracted HOG features from range-Doppler spectrum of different gestures. Throughout these feature extraction processes, consistent HOG feature extraction algorithm parameters are employed, with a cell size of 64×256, a block size of 2×1, and an image size of 256×256.

Figure 9 reveals that even for the same gesture, HOG features vary across different distances. Conversely, the HOG features of the foreground images obtained through the Doppler factor foreground extraction algorithm maintain a high level of consistency. Consistency of features for the same gesture across different distances is pivotal for enhancing gesture recognition accuracy.



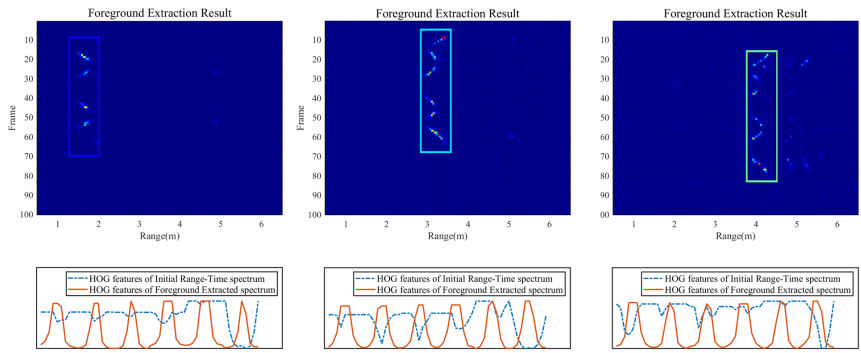
(a) Cross Hands (b) Wave Hands (c) Click three times

Figure 7: Comparison of bodily micro-motion suppression algorithms.



(a) Cross Hands (b) Wave Hands (c) Click three times

Figure 8: Comparison of foreground extraction effect and HOG feature extraction.



(a) Gesture 4 performs at 2m (b) Gesture 4 performs at 3.7m (c) Gesture 4 performs at 4.6m

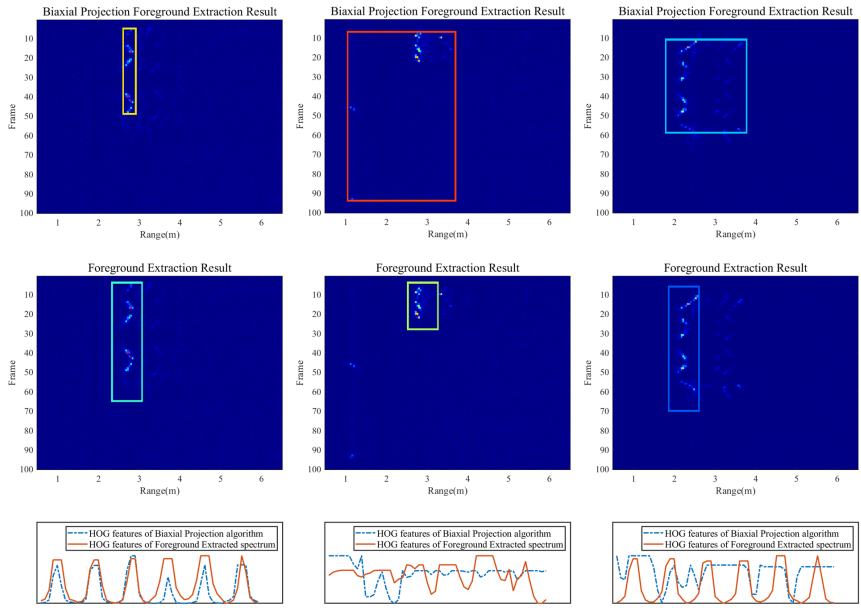
Figure 9: Comparison of foreground extraction effects at different distances.

3.2.2 Comparative Analysis of Foreground Segmentation Algorithm

Existing foreground extraction algorithms for long-range gestures often employ the biaxial projection method, which segments the accumulated range-time spectrum by projecting it onto different dimensions, thereby achieving a relatively stable detection outcome [18]. In this section, a comparative analysis is conducted between the proposed method and biaxial projection method in terms of their efficacy for foreground extraction from range images, as depicted in Figure 10.

While the biaxial projection method exhibits stable gesture motion detection, it necessitates a longer temporal window to encompass the entire gesture motion signal. This compromises real-time performance and renders it more sensitive to noise. The foreground extraction algorithm proposed in this study, based on the Doppler factor, enables real-time assessment of each frame's signal. Moreover, leveraging the constructed Doppler factor for signal processing renders the entire algorithm less susceptible to noise, resulting in more accurate foreground extraction from captured imagery. The accuracy of gesture classification processed by these two methods are shown in Table 4.

Compared with the biaxial projection method, the proposed method has an overall improvement of about 5% in recognition accuracy, but still has lower recognition accuracy for Click Three Times and Click Two Times gestures. Considering the similarity between these two gestures, directly using the features extracted from the foreground for classification can lead to significant confusion. Therefore, in subsequent signal processing, Gaussian filtering is used to further improve the discrimination of different features.



(a) Cross Hands (b) Wave Hands (c) Click three times

Figure 10: Comparison of proposed method and the biaxial projection method.

Table 4: Comparison of recognition accuracy between two methods.

Gesture type	Biaxial Projection	Proposed Method
Cross Hands	85.6%	96.1%
Wave Hands	93.1%	96.2%
Click Three Times	81.2%	86.2%
Click Two Times	91.2%	86.9%
Click One Time	86.2%	95.6%
Average Accuracy	87.5%	92.2%

3.2.3 Effectiveness Analysis of Gauss Filtering

Gaussian filtering induces image blurring, thereby facilitating the attenuation of subtle details. This smoothing effect directs edge detection to focus on prominent variations within the image while being less affected by minor fluctuations. To assess the role of Gaussian filtering in feature extraction, this section undertakes a comparative analysis between before and after Gaussian-

filtered feature maps along with their corresponding classification outcomes.

Given that edges and noise both manifest as high-frequency components within an image, Gaussian blurring suppresses noise by attenuating high-frequency information. While this procedure also affects edges, the continuity inherent to edges is enhanced to some extent due to Gaussian filtering as shown in Figure 11. Conversely, isolated noise components are further suppressed. Consequently, calculating gradients using the smoothed image yields more stable and continuous gradient information, thereby leading to more precise edge detection outcomes.The accuracy of gesture classification processed by these two methods is shown in Table 5.

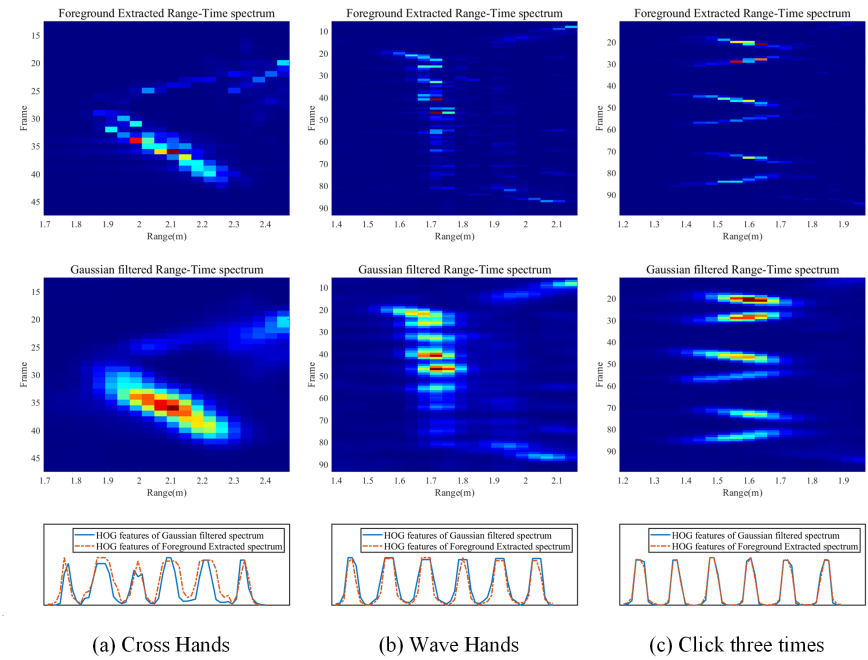


Figure 11: Comparison of Gauss filtering effect and HOG feature extraction.

3.2.4 Algorithm Effectiveness Evaluation through Ablation

In this section, a series of ablation experiments are conducted on the algorithms employed within the entirety of the signal processing pipeline, aiming to validate the efficacy of each algorithm and assess the consistency of their contributions to the improvement of recognition outcomes.

Table 5: Comparison of Gaussian filter for improving recognition accuracy.

Gesture type	Raw Spectrum	Gauss Filtered Spectrum
Cross Hands	89.0%	89.6%
Wave Hands	78.6%	84.9%
Click Three Times	65.0%	80.0%
Click Two Times	72.5%	82.5%
Click One Time	86.2%	91.9%
Average Accuracy	78.2%	85.8%

Experiment 1 shows the recognition accuracy of removing bodily micro-motion suppression, Experiment 2 shows the recognition accuracy without using foreground extraction algorithm, and Experiment 3 shows the recognition accuracy without Gaussian filtering operation. The accuracy of gesture classification using SVM is shown in Table 6.

Table 6: Results of ablation experiment.

Gesture type	Experiment 1	Experiment 2	Experiment 3
Cross Hands	99.4%	89.6%	97.4%
Wave Hands	95.6%	84.9%	88.1%
Click Three Times	88.8%	80.0%	75.0%
Click Two Times	93.8%	82.5%	91.2%
Click One Time	96.2%	91.9%	94.4%
Average Accuracy	94.7%	85.8%	89.2%

Meanwhile, the extracted features are input into a 6-layer LSTM network for analysis, thereby validating that the obtained characteristics encompass temporal information regarding the gesture movements. The recognition accuracy is depicted in Table 7. The initial accuracy represents the classification accuracy of the features obtained from the original spectra, and the last accuracy represents the classification accuracy of the features obtained after all signal processing.

From the results in Table 7, the HOG feature extracted from the raw spectra fails to induce convergence in the LSTM network, which suggests a lack of discernible temporal patterns within the extracted features. Conversely, HOG feature processed through the proposed methodology achieves a testing accuracy of 91.6%, indicating that the extracted features already contain temporal information of the gestures.

Table 7: Temporal feature validation using LSTM network.

Gesture type	Initial Accuracy	Experiment 3	Last Accuracy
Cross Hands	0.0%	97.8%	100.0%
Wave Hands	0.0%	80.4%	92.3%
Click Three Times	20.0%	77.1%	84.2%
Click Two Times	0.0%	75.6%	87.5%
Click One Time	0.0%	87.8%	94.9%
Average Accuracy	Not Convergent	83.6%	91.6%

Then, the impact of different filters on the recognition results before HOG feature extraction was compared, as shown in Table 8. It can be seen that due to the ability of Gaussian filtering to retain edge information, the detection accuracy of 'Click Three Times' has been significantly improved.

Table 8: The recognition accuracy using different filters.

Gesture type	Mean Filter	Median Filter	Gauss Filter
Cross Hands	99.3%	98.7%	98.7%
Wave Hands	94.3%	91.2%	96.9%
Click Three Times	87.5%	85.0%	92.5%
Click Two Times	93.8%	93.1%	93.8%
Click One Time	98.8%	97.5%	98.1%
Average Accuracy	94.7%	94.0%	96.0%

Table 9: The recognition accuracy of initial accuracy and after processed.

Gesture type	Initial Accuracy	Last Accuracy
Cross Hands	89.0%	98.7%
Wave Hands	78.6%	96.9%
Click Three Times	65.0%	92.5%
Click Two Times	72.5%	93.8%
Click One Time	86.2%	98.1%
Average Accuracy	78.2%	96.0%

Through comparative experimentation, it is evident that micro-motion suppression, foreground extraction, and Gaussian filtering each contribute to the enhancement of the final recognition accuracy. Moreover, these enhancement effects exhibit an accumulative nature, thereby attesting to the effectiveness of the entire signal processing pipeline. Ultimately, the employment of the SVM algorithm achieves a classification accuracy up to 96%, as shown in Table 9.

4 Conclusion

This paper introduces a novel remote dynamic gesture recognition system utilizing a 60GHz FMCW radar. By constructing Doppler factors, we achieved micro-motion suppression of the body and foreground extraction of gestures. These processes eliminate interference from the background, enabling more precise feature extraction that accurately targets the gesture segment. As a result of the combined effects of these two algorithms, there is a noticeable increase in gesture recognition accuracy. Furthermore, Gaussian filtering applied prior to feature extraction significantly enhances the effectiveness of the HOG feature extraction. Ultimately, employing the SVM algorithm yielded a recognition accuracy of 96% for the five gestures in both standing and seated positions within the 1-5m detection range. As a continuation of this research, we will encompass the evaluation of the negative class impact on gesture recognition, with a particular focus on distinguishing other hand-gesture actions occurring within the detection range that are not included in the predefined gesture categories.

References

[1] M. R. Abid, E. M. Petriu, and E. Amjadian, "Dynamic Sign Language Recognition for Smart Home Interactive Application Using Stochastic Linear Formal Grammar", *IEEE Transactions on Instrumentation and Measurement*, 64(3), 2015, 596–605.

[2] Y. Bai, I. Shahid, H. Takawale, and N. Roy, "WhisperWand: Simultaneous Voice and Gesture Tracking Interface", 2023, arXiv: 2301.10314.

[3] J. Canny, "A Computational Approach to Edge Detection", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-8(6), 1986, 679–98.

[4] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection", in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, Vol. 1, 2005, 886–93.

[5] Y. Dong, W. Qu, P. Wang, H. Jiang, T. Gao, and Y. Shu, "Radar Gesture Recognition Based on Lightweight Convolutional Neural Network", in *Seventh Asia Pacific Conference on Optics Manufacture and 2021 International Forum of Young Scientists on Advanced Optical Manufacturing (APCOM and YSAOM 2021)*, Vol. 12166, 2022, 221–8.

[6] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*, Second, Upper Saddle River, New Jersey 07458: Prentice Hall, 2001, ISBN: 0-201-18075-8.

[7] S. Gupta, S. Bagga, and D. K. Sharma, "Hand Gesture Recognition for Human Computer Interaction and Its Applications in Virtual Reality", in, ed. D. Gupta, A. E. Hassanien, and A. Khanna, Cham, 2020, 85–105.

[8] S. Jiang, B. Lv, W. Guo, C. Zhang, H. Wang, X. Sheng, and P. B. Shull, "Feasibility of Wrist-Worn, Real-Time Hand, and Surface Gesture Recognition via sEMG and IMU Sensing", *IEEE Transactions on Industrial Informatics*, 14(8), 2018, 3376–85.

[9] B. G. Lee and S. M. Lee, "Smart Wearable Hand Device for Sign Language Interpretation System With Sensors Fusion", *IEEE Sensors Journal*, 18(3), 2018, 1224–32.

[10] D. G. León, J. Gröli, S. R. Yeduri, D. Rossier, R. Mosqueron, O. J. Pandey, and L. R. Cenkeramaddi, "Video Hand Gestures Recognition Using Depth Camera and Lightweight CNN", *IEEE Sensors Journal*, 22(14), 2022, 14610–9.

[11] E. Osuna, R. Freund, and F. Girosi, "An Improved Training Algorithm for Support Vector Machines", in *Neural Networks for Signal Processing VII. Proceedings of the 1997 IEEE Signal Processing Society Workshop*, 1997, 276–85.

[12] 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 Transactions on Mobile Computing*, 2022.

[13] J. S. Suh, S. Ryu, B. Han, J. Choi, J.-H. Kim, and S. Hong, "24 GHz FMCW Radar System for Real-Time Hand Gesture Recognition Using LSTM", in *2018 Asia-Pacific Microwave Conference (APMC)*, 2018, 860–2.

[14] M. Ur Rehman, F. Ahmed, M. Attique Khan, U. Tariq, F. Abdulaziz Alfouzan, N. M. Alzahrani, and J. Ahmad, "Dynamic Hand Gesture Recognition Using 3D-CNN and LSTM Networks", *Computers, Materials & Continua*, 70(3), 2022, 4675–90.

[15] Z. Wang, F. Liu, X. Li, M. Ma, X. Feng, and Y. Guo, "A Survey of Hand Gesture Recognition Based on FMCW Radar", in *Proceedings of the 8th International Conference on Communication and Information Processing*, *ICCIP '22*, New York, NY, USA, 2023, 73–9.

[16] 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 Transactions on Geoscience and Remote Sensing*, 59(6), 2021, 4749–64.

[17] Z. Xia and F. Xu, "Time-Space Dimension Reduction of Millimeter-Wave Radar Point-Clouds for Smart-Home Hand-Gesture Recognition", *IEEE Sensors Journal*, 22(5), 2022, 4425–37.

[18] B. Zhang, H. Luo, C. Tang, G. Wang, Y. Zhang, S. Guo, and G. Cui, "Long-Range Real-Time Gesture Recognition for Millimeter Wave Radar", in *2022 2nd International Conference on Frontiers of Electronics, Information and Computation Technologies (ICFEICT)*, 2022, 298–303.