Domain-Independent Gesture Recognition Using Single-Channel Time Modulated Array

Lei Guan, Xiaodong Yang, Nan Zhao, Akram Alomainy, Muhammad Ali Imran, Qammer H. Abbasi

***Abstract*—In recent years, gesture recognition system based on radio frequency (RF) sensing has a wide application prospect and attraction in non-contact electronic interaction with its advantages of privacy security, lighting independence, and wide sensing range. The traditional RF sensing system depends on the environment and the subject, and the multi-channel sensing equipment is expensive, which brings great challenges to the practical application. To address the above issues, a single-channel, low-cost, and domain-independent gesture recognition system is proposed. Specifically, the time modulation technology is adopted to expand the number of antennas of the sensing device. The time modulation array (TMA) is converted into a traditional array through harmonic recovery technology. 2D-FFT, moving target indication filter, and data normalization are used to extract domain-independent Angle-Doppler Maps (ADMs) gesture features. In order to ensure recognition accuracy, we propose a lightweight neural network with attention mechanism, which only needs one training and can be applied to different data domains. The experimental results show that the accuracy of in-domain recognition of the proposed system is 98.9%, and the accuracy of cross-domain (i.e. new environments, new users, and new positions) recognition is 85.6%-97.4% without model retraining.**

***Index Terms*—Gesture Recognition, TMA, Neural Network**

I. Introduction

With the rapid development of the Internet of Things and the continuously increasing requirement of human-machine interaction (HMI). Hand gesture recognition (HGR), as an HCI mode with a high frequency of use and strong expression ability, has attracted wide attention. HGR provides convenience for human life and has important applications in smart home, sign language interaction, virtual reality, vehicle-assisted driving, and other fields.

Currently, several technologies have been used to implement gesture recognition systems, e.g., computer vision (CV), infrared sensors and wearable devices. CV-based methods [1-2] support contactless gesture recognition, but these methods face the problems of privacy disclosure and light sensitivity. The low-cost infrared sensors can obtain more fine-grained palm [3] and thumb-tip [4] gestures. The detection range of infrared sensors are generally only within 35 cm, and their performance is affected by temperature and light. While wearable devices such as inertial sensors [5-6] and smart watches [7] can achieve very high precision in detail, they limit the flexibility of user gestures and reduce the immersive experience. The need for contactless, privacy-preserving triggers extensive research on radio frequency (RF) sensing.

With the development of RF microelectronics technology, ubiquitous wireless signals and widely available radio frequency sensors are used in sensing tasks. Doppler radar [8-9] has been widely studied in the field of gesture recognition due to its high sensitivity to small-scale motion and its excellent ability to distinguish non-stationary objects from static backgrounds. In WiSee [10], the authors use universal software radio peripheral (USRP) to extract the doppler information of orthogonal frequency division multiplexing (OFDM) signals to enable whole-home sensing and recognition of human gestures. The single channel doppler radar can only measure the radial velocity of the target. The single dimensional doppler information makes it difficult to distinguish multiple similar gestures, which leads to system performance degradation. In [11], the dual-antenna doppler radar system improves the accuracy of single gesture recognition by combining doppler spectrum with angle of arrival (AOA) spectrum. In [12], the author uses four continuous wave (CW) radars to form an array, which can obtain more doppler in the direction and provide more abundant gesture spatial information. Besides, researchers try to introduce the information of distance domain and angle domain to improve the robustness of the system. In [13], the authors input raw data from multi-channel frequency-modulated continuous-wave (FMCW) radars into a neural network to fully extract gesture features. In [14] and [15], the authors use time division multiplexing (TDM) multiple-input-multiple-output (MIMO) FMCW to package features from multiple domains into feature blocks to improve the accuracy of gesture recognition. Li *et al*. proposed a virtual array configuration strategy to achieve adaptive gesture recognition at different distances [16]. However, the high price is a great challenge for the practical deployment of MIMO millimeter wave radar in the home. The ubiquitous WiFi signal is not only used to transmit information, but also has the sensing ability. The received signal strength indicator (RSSI) [17-19] and channel state information (CSI) are widely used in sensing applications. The sensing system based on RSSI will suffer from the same frequency interference. Besides, the coarse-grained RSSI is difficult to handle complex tasks and small-scale perception. The amplitude [20] and phase difference [21-22] of fine-grained CSI can describe the influence of gesture on

Lei Guan, Xiaodong Yang, Nan Zhao are with the School of Electronic Engineering, Xidian University, Xi’an, Shaanxi, China, 710071.

Akram Alomainy is with the School of Electronic Engineering and Computer Science, Queen Mary University of London, London, E1 4NS, UK.

Muhammad Ali Imran and Qammer H. Abbasi are with the James Watt School of Engineering, University of Glasgow, Glasgow, G12 8QQ, UK.

the current channel. Although these studies have achieved high recognition accuracy in-domain dataset, they ignore the role of signal models, which makes these systems difficult to apply in dynamic environments. In other words, when the trained model is deployed in a new scenario, the performance of the system will decline significantly. In widar3.0 [23] and [24], the authors explored the cross-domain generalization ability of WiFi signals. By generating domain-independent information through doppler, excellent recognition accuracy is achieved in different environments, different directions and different subjects. However, these studies need to deploy multiple WiFi nodes, which makes it difficult to apply in the home scenario.

Machine learning and deep learning are widely used in gesture recognition based on RF sensing. Dynamic time warping (DTW) [25], Support vector machine (SVM)[26] random forest classifier [27] and Hidden Gauss–Markov model [28] are used for gesture recognition. The conventional machine learning method has low computational complexity and high interpretability. However, traditional machine learning relies on artificial feature extraction and are mainly suitable for small-scale data learning, so the robustness and generalization ability of the model hinder its development. Deep learning breaks through the above limitations. The convolutional neural network (CNN) [11], [29], [30] is used to extract the gesture features from the spectrum such as range time maps, doppler time maps and angle time maps. In order to extract temporal and spatial features, both CNN and Long Short-Term Memory (LSTM) are designed to model the dynamic information of gesture [31]. In [32], the authors designed a 3D-CNN for short spatial-temporal modeling, LSTM for global temporal feature extraction, and a CTC layer for classifying hand gestures.

In recent years, with the performance improvement and cost reduction of high-speed RF switches, TMA has attracted the attention of researchers again. TMA is designed to generate ultra-low sidelobes, but periodic modulation causes the TMA to produce unwanted harmonics, which causes spectral interference. In order to solve the above problems, the researchers adopted differential evolution [33], genetic algorithm [34] and particle swarm optimization [35] to optimize the switching sequence of RF switches to achieve the suppression of sideband level. The convex optimization algorithm [36-38], FFT [39], [40] and artifcial neural network [41], [42] are employed to accomplish the synthesis of TMA. With the change of concept in recent years, harmonic components are no longer considered as adverse factors in new applications, such as harmonic beam scanning [43], multiuser communication [44] and radar-communication integration [45], [46]. In addition, TMA is widely used in direction finding [47-50]. In [47] and [48], the sparse signal recovery was proposed in TMA for the AOA estimation. In [50], the authors use *l*2-norm approximation method to covert 1bit TMA into conventional array, and uses the spatial spectrum estimation method to calculate the AOA.

We note that in order to obtain rich gesture information, previous studies have deployed multiple sensing nodes or used radar devices with high cost and complex hardware. Inspired by the research of existing gesture recognition systems and TMA-related applications, we propose a single-channel, low-complexity gesture recogniton system. To quickly demonstrate the feasibility of the technology, the USRP is configured as a CW radar. It is worth noting that commercial RF modules can replace USRP, making the system low-cost.

The main contributions of this paper are summarized as follows.

(1) We propose a TMA-based CW radar gesture sensing system. The conventional multi-channel array is restored by the harmonic components of the single-channel TMA, which simplifies the complexity of the multi-channel radar.

(2) We perform 2D-FFT on the recovered multi-channel data to acquire ADM to characterize gesture motion, and propose a lightweight neural network with an attention mechanism for spatiotemporal feature extraction.

(3) We built a prototype and conducted performance evaluation of the system. The experimental results show that the accuracy of cross-domain (i.e. new environments, new users, and new positions) recognition is 85.6% - 97.4% without model retraining.

The rest of this paper is organized as follows. Section II introduces the signal model of CW radar and the fundamental theory for TMA. In Section III, we present the design of our system. In Section IV introduces signal processing and gives the structure of neural network. Section V shows the experimental setup and performance evaluation. Section VI extends further discussions. Finally, Section VII draws conclusions.

II. Fundamental Theory

A. Signal Model

Consider an *N*-element uniform linear array and receive the echos of *L* targets from the far-field. The signal received by the *n*th antenna is:

$$x_n\left(t\right)=\sum_{i=1}^{L}\alpha_i e^{j2\pi f_c\left(t+\tau_{i,n}\left(t\right)\right)} \tag{1}$$

where f_c denotes the carrier frequency of the transmitted signal, α_i is the amplitude of the received signal related to the ith target's radar cross section, transmitting power and so on. $\tau_{i,n}(t)$ is the delay of the ith target received by the nth antenna and can be rewritten as:

$$\tau_{i,n}\left(t\right)=\frac{2\left(d_{0,i}+\int_0^t v_i\left(t\right)\cos\varphi_i dt\right)+d\left(n-1\right)\sin\theta_i}{c} \tag{2}$$

where $d_{0,i}$ is the radial distance of the ith target relative to the radar, $v_i(t)$ it the speed of the ith target， c is the speed of light, φ_i is the angle of the target's velocity vector relative to the radar line of sight, d is the distance between adjacent antennas, θ_i is the corresponding incident angle as shown in Fig. 1.

In the Doppler radar, the f_d is expressed as:

$$f_{di}\left(t\right)=f_c\frac{2v_i\left(t\right)\cos\varphi}{c} \tag{3}$$

Assuming that the speed of the target remains constant for a short time, then:

$$f_{di}t = \frac{\int_0^t f_c 2v_i(t)\cos\varphi dt}{c} \tag{4}$$

substituting (2) and (4) into (1) leads to:

$$x_n(t) = \sum_{i=1}^{L} \alpha_i e^{j2\pi(f_c+f_{di})t + j2\pi f_c \tau_{i,0} + j(n-1)\beta\sin\theta_i} \tag{5}$$

where *β=2π/λ* is the wavenumber with *λ* the carrier wavelength.

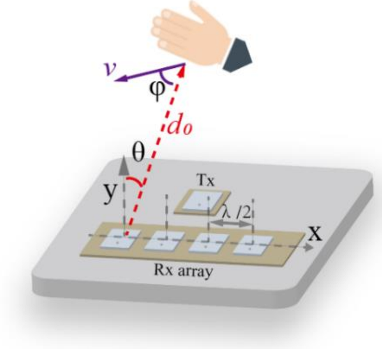
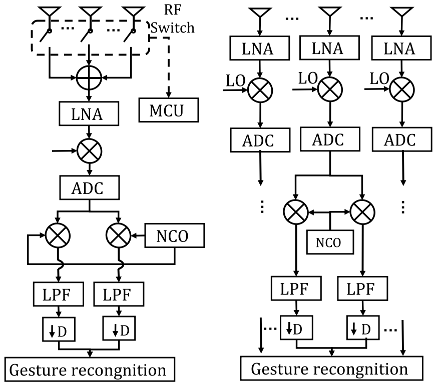


Fig. 1. Gesture recognition using multi-channel sensing system.



(a) (b)

Fig. 2. Structure of receiver. (a)Single channel TMA. (b) Conventional multichannel array.

B. Fundamental Principle of TMA

As shown in the Fig. 2, compared with the traditional multi-channel receiver, the TMA-based receiver uses only one RF link, which reduces the hardware complexity. In the proposed TMACW radar, the received signal is modulated by a periodic ON/OFF modulation function *U_n*(t). Then the received signal is:

$$y(t) = \sum_{n=1}^{N} U_n(t) x_n(t) \tag{6}$$

U_n(t) is a periodic pulse function with modulation period *T*. In each period, U_n(t) is expressed as:

$$U_n(t) = \sum_{q=1}^{Q} U_n^q C_n^q(t) \tag{7}$$

where *Q* is the length of the time-coding sequence, $C_n^q(t)$ is a pulse period function with modulation period *T* and is given by:

$$C_n^q(t) = \begin{cases} 1, & (q-1)\tau < t < q\tau \\ 0, & others \end{cases} \tag{8}$$

where $\tau = T_p/Q$ is the pulse width of $C_n^q(t)$, $U_n^q \in [0,1]$ is the amplitude of the *n*th element during the interval $(q\text{-}1)\,\tau \leqslant t \leqslant q\tau$. Specifically, when the switch is open $U_n^q = 1$, closed $U_n^q = 0$. Next, we decompose $C_n^q(t)$ into a Fourier series

$$C_n^q(t) = \sum_{k=-\infty}^{\infty} r_{n,k}^q e^{j2\pi k f_p t} \tag{9}$$

where $f_p=1/T_p$ is the modulation frequency. the Fourier coefficients $r_{n,k}^q$ are given by

$$r_{n,k}^q = \frac{1}{T_p}\int_0^{T_p} C_n^q(t) e^{-j2\pi k f_p t} \tag{10}$$

Thus, the Fourier series coefficients of the periodic function U_n(t) can can be expressed as

$$\begin{aligned} a_{n,k} &= \sum_{q=1}^{Q} U_n^q r_{n,k}^q \\ &= \sum_{q=1}^{Q} \frac{U_n^q}{T_p} \int_{(q-1)\tau}^{q\tau} e^{-j2\pi k f_p t} \\ &= \sum_{q=1}^{Q} \frac{U_n^q}{T_p} \text{sinc}\left(\frac{\pi k}{Q}\right) e^{\frac{-j\pi k(2q-1)}{Q}} \end{aligned} \tag{11}$$

The modulation function U_n(t) can be expanded by Fourier series

$$U_n = \sum_{k=-\infty}^{\infty} a_{n,k} e^{j2\pi k f_p t} \tag{12}$$

Inserting (11) and (12) into (6), the single-channel signal after time modulation can be written as:

$$\begin{aligned} y(t) &= \sum_{n=1}^{N} U_n(t) x_n(t) \\ &= \sum_{n=1}^{N}\left(\sum_{k=-\infty}^{\infty} a_{n,k} e^{j2\pi k f_p t}\right) x_n(t) \\ &= \sum_{k=-\infty}^{\infty} e^{j2\pi k f_p t}\left(\sum_{n=1}^{N} a_{n,k} x_n(t)\right) \end{aligned} \tag{13}$$

We can find that the signal received by a single channel is the sum of the fundamental component and each order harmonic component. In this study, we use a single-pole multi-throw (SPMT) RF switch to implement the modulation function. That is, only one link is in the on-state at a certain time. In order to simplify the control, the TMA units are switched on and off successively and the time modulation function is shown in Fig. 3.

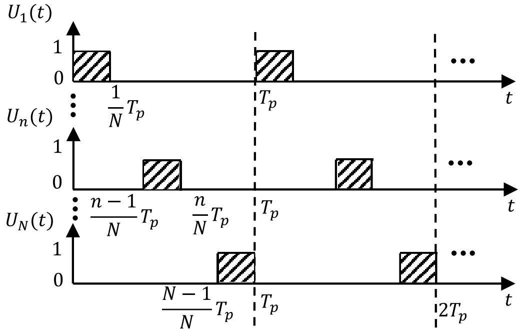


Fig. 3. Periodical modulation function U_n(t), n = 1,2, …, *N*.

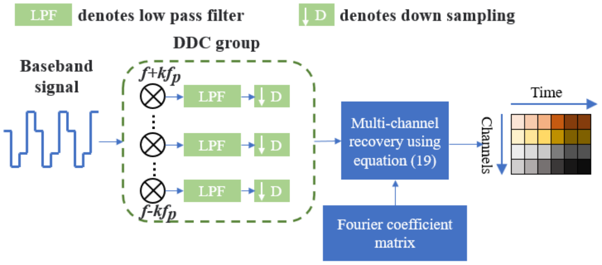

LPF denotes low pass filter
↓D denotes down sampling
DDC group
Baseband signal
f+kfp
LPF
↓D
f-kfp
Multi-channel recovery using equation (19)
Fourier coefficient matrix
Time
Channels

Fig. 4. Signal processing steps for multi-channel array recovery.

C. Multichannel Array Recovery

The signal processing diagram of TMA recovering conventional multi-channel is shown in the Fig. 4.

We designed a series of digital down conversion (DDC) to obtain harmonic components. For the *k*th DDC, we mix the received signal with the *k*th harmonic to move the *k*th harmonic to the baseband. Then, the digital baseband signal of the *k*th harmonic is obtained through low-pass filter (LPF). Finally, downsampling is used to reduce the complexity and resource consumption of storage and processing. The baseband signal of the *k*th harmonic component at f_0+kf_p in the digital domain can be expressed as

$$y_k(m) = \sum_{n=1}^{N} a_{n,k} x'_n(m) \tag{14}$$

where $x'_n(m)$ is the baseband signal received by the *n*th antenna. Add noise to the model and rewrite (14) with vector notation:

$$\begin{aligned}\mathbf{Y}(m) &= \boldsymbol{\Gamma}\left(\mathbf{X}(m)+\mathbf{N}(m)\right) \\ &= \boldsymbol{\Gamma}\mathbf{X}(m)+\mathbf{N}'(m)\end{aligned} \tag{15}$$

where $\boldsymbol{Y}(m)=[y_{-K}(m), y_{-K+1}(m)\dots, y_K(m)]^T$ is the harmonic vector generated by the received signal modulated by the RF switch, $\mathbf{X}(m)=[x'_1(m), x'_2(m)\dots, x'_N(m)]^T$ is baseband signals received by conventional multichannel arrays, where $x'_n(m)$ can be obtained by downconverting (5).

$$x'_n(m) = \sum_{i=1}^{L} \alpha_i e^{j2\pi f_{di} m + j2\pi f_c \tau_{i,0} + j(n-1)\beta \sin\theta_i} \tag{16}$$

$\mathbf{N}'(m)=[n_{-K}(m), n_{-K+1}(m), \dots, n_K(m)]^T$ is the noise vector, $[\cdot]^T$ denotes the transpose operator and $\boldsymbol{\Gamma} \in \mathbb{C}^{(2K+1)\times N}$ is the Fourier coefficient matrix corresponding to the modulation function and is given by:

$$\boldsymbol{\Gamma} = \begin{bmatrix} a_{-K,1} & a_{-K,2} & \cdots & a_{-K,N} \\ a_{-K+1,1} & a_{-K+1,2} & \cdots & a_{-K+1,N} \\ \vdots & \vdots & \ddots & \vdots \\ a_{K,1} & a_{K,2} & \cdots & a_{K,N} \end{bmatrix} \tag{17}$$

It can be found that the harmonic component can be obtained by linear transformation of the received signal of the conventional linear array, and the transformation matrix is $\boldsymbol{\Gamma}$. If $\boldsymbol{\Gamma}$ is full rank, $\mathbf{X}$(m) can be restored by $\mathbf{Y}$(m).

$$\min_{\mathbf{X}(m)} \left\| \boldsymbol{\Gamma}\mathbf{X}(m) - \mathbf{Y}(m) \right\| \tag{18}$$

where $\|\cdot\|$ is a norm on $\mathbb{C}^{2K+1}$. The best-fit solution can be obtained applying complex least square：

$$\mathbf{X}(m) = \left(\boldsymbol{\Gamma}^H \boldsymbol{\Gamma}\right)^{-1} \boldsymbol{\Gamma}^H \mathbf{Y}(m) \tag{19}$$

where $[\cdot]^H$ indicates the conjugate transpose operator, $\mathbf{X}(m)$ is the recovered signal of conventional linear array, which contains Doppler and AOA information of the targets.

III. System Design

A. Antenna Array Design

In this paper, a standard probe-fed microstrip patch antenna is used. The beam width of the transmitting antenna and each receiving antenna needs to be wide enough to meet the application scenario of gesture recognition. The antenna is designed for FR4 substrate with a thickness of 0.508 mm. The maximum realized gain is approximately 5.6 dB and the front-to-back ratio is approximately 10 dB. The Rx array is a uniform linear array composed of four microstrip elements, and the distance between adjacent antennas is λ/2 to avoid the grating lobe issue. The Rx array and Tx antenna of our system are shown in Fig. 5:

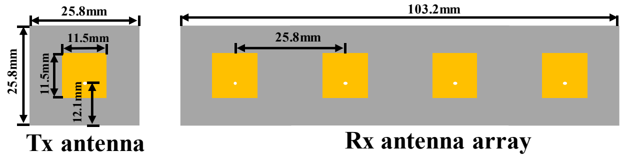


Fig. 5. Layout of Tx antenna and Rx antenna array.

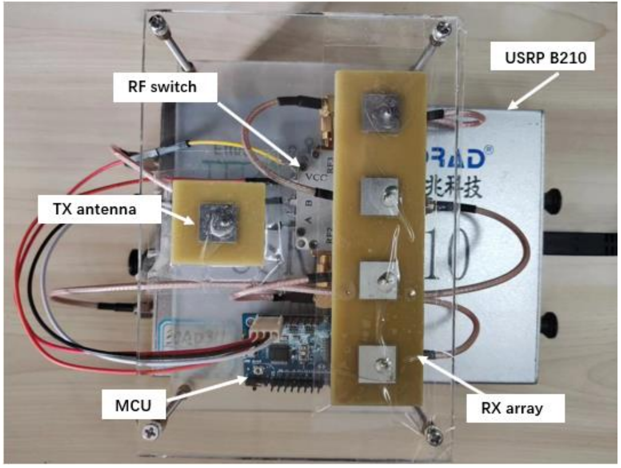


Fig. 6. Photograph of the TMACW radar system.

B. TMACW Radar Design

In the paper, we build a TMACW prototype with commercial off-the-shelf (COTS) hardware. The hardware used can be easily replaced with other general COTS substitutes of similar functionality.

We use the open source GNURadio framework to build the system flowgraph. The USRP B210 sends the collected data to the host through the USB cable. The TMA consists of a single pole four throw (SP4T) RF switch, a receiving antenna array (cost~$1) and a STM32F193C8T6 (STM32) microcontroller. The SP4T switch used is HMC7992 (cost~$9) which provides fast switching speed, i.e. 150 ns, low insertion loss, i.e. 1dB and high isolation at 5GHz, i.e. 30db. STM32 (cost~$0.9) is used as the external MCU, and its GPIO controls the HMC7992 to turn on and off the element in turn according to the order shown in Fig. 3. The prototype of the system is shown as Fig. 6:

IV. Signal processing

A. Harmonic components acquisition

After receiving the signal, we mix the received signal with different harmonic components and then pass through a low-pass filter to obtain the baseband signal. However, the modulation frequency generated by the MCU deviates from the ideal modulation frequency, which affects the recovery of the traditional array. It assumes that the modulation frequency offset of the *k*th harmonic is $f_{res,k}$, the harmonic component is expressed as:

$$y_k(m)=e^{j2\pi f_{res,k}m}\sum_{n=1}^{N}a_{n,k}x'(m) \tag{20}$$

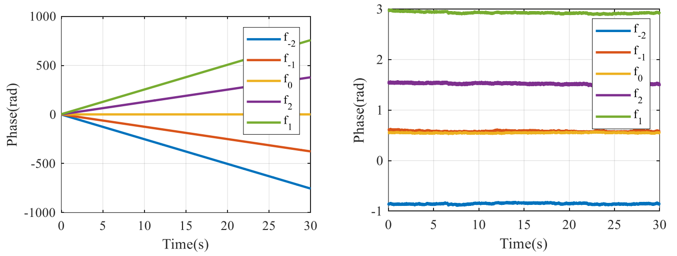
Rewrite (16) with vector notation:

$$\mathbf{Y}(m)=\mathbf{B}(\mathrm{m})\mathbf{\Gamma}\mathbf{X}(m)+\mathbf{N}'(m) \tag{21}$$

where $\mathbf{B}$(m) = diag[$e^{j2\pi f_{res,-K}m}, e^{j2\pi f_{res,-K+1}m}, \dots, e^{j2\pi f_{res,K}m}$]^T. Therefore, the recovered traditional array is expressed as:

$$\mathbf{X}(\mathrm{m})=\left(\mathbf{\Gamma}^H\mathbf{\Gamma}\right)^{-1}\mathbf{\Gamma}^H\mathrm{B}(m)^{-1}\mathbf{Y}(m) \tag{22}$$

In order to ensure accurate and effective harmonic recovery, it is necessary to estimate and eliminate the frequency offset. First, we collect 30 s of data in a static environment, then we mix the received signal with ideal harmonic components and finally obtain a rough harmonic baseband signal through a low-pass filter. We use the Filter Designer app of MATLAB to design a Butterworth low-pass filter with a passband of 0-200Hz. It can be seen from (16) that the $f_{res,k}$ causes the phase of the harmonic to change linearly with time. We calculate the slope ω_k of the phase of each harmonic component by linear regression, then $f_{res,k}=\omega_k/2\pi t$. The stable harmonic components can be obtained by compensating the $f_{res,k}$ in the mixing stage. The harmonic phase comparison before and after carrier offset compensation is shown in Fig. 7.



(a) (b)

Fig. 7 (a) Phase without removing frequency offset. (b) phase after compensation.

After the above processing, we can obtain a stable phase and accurate frequency of each harmonic component as shown in Fig. 7 (b). Then, the calibrated harmonic components are substituted into (15) to restore the conventional array.

B. TMA Synchronization

In the research of TMA application, many simulation studies assume synchronous sampling between switch and ADC. In a simple sampling system, the control complexity of synchronous sampling is high and difficult to achieve. If the starting point of time modulation is not correctly located, different phases are introduced into different harmonics [42]. Given the delay Δ*t* in (8), the *k*th harmonic coefficient should be rewrited as:

$$a_{n,k}=e^{j2\pi kf_p\Delta t}a_{n,k} \tag{23}$$

Then the harmonic coefficient matrix should be rewritten as:

$$\mathbf{\Gamma}=\mathbf{D\Gamma} \tag{24}$$

where $\mathbf{D}$ = diag[$e^{-j2\pi Kf_p\Delta t}, e^{j2\pi(1-K)f_p\Delta t}, \dots, e^{j2\pi Kf_p\Delta t}$]^T is the delay matrix. If Δ*t* is not considered and $\mathbf{\Gamma}$ is still used to restore the array, it will cause errors and affect the AOA estimation. Finding the starting position of a modulation period is also the key to recover multi-channel. We use sliding window and cosine similarity to locate the switching time as shown in the Fig. 8. First, the RF switch is not activated for a short period of time at the beginning of sampling. Then, we calculate the similarity using two sliding windows of length T_p/N, and y_{w2} is T_p/N samples ahead of y_{w1}. The synchronous indicator can be calculated as:

$$v(i)=1-\left|\frac{\sum_{i=1}^{T_p/N}\mathrm{y}_{w1}(i)\times\mathrm{y}_{w2}(i)}{\sqrt{\sum_{i=1}^{T_p/N}\mathrm{y}_{w1}(i)}\times\sqrt{\sum_{i=1}^{T_p/N}\mathrm{y}_{w2}(i)}}\right| \tag{25}$$

At the beginning of sampling, *v* is close to 0. When the RF switch starts to work, *v* will increase. In our experiment, the synchronization threshold is set as 0.1. Assuming that the value of the synchronization indicator at *n* is greater than the threshold, the initial sampling point of time modulation is *n*+ T_p/N.

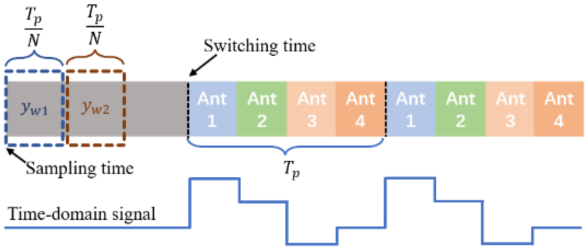


Fig. 8. The signal of the single-channel signal in time-domain.

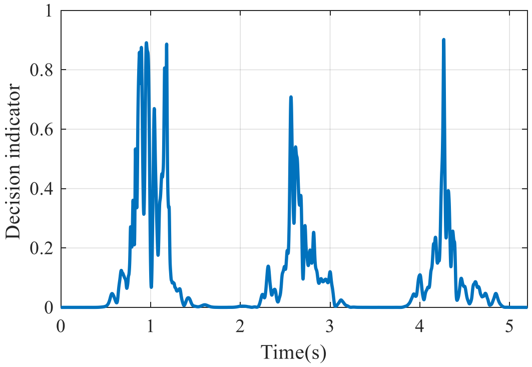


Fig. 9. Signal processing flowchart for constructing ADM.

C. Gesture Segmentation

In order to segment continuous gestures, previous studies mostly used the variance of sliding window data to determine whether there is a moving gesture in the current time. However, the range of variance data is affected by the received signal power, and it is difficult to determine the appropriate threshold to distinguish the beginning and end of the gesture. At the same time, when a hand gesture has round-trip motion, the decision indicator may have deep fading, resulting in error in hand gesture segmentation. To avoid the above problems, we also use two sliding windows and calculate their cosine similarity. Compared with the method based on variance, cosine similarity can clarify the range of values and help to select the appropriate threshold. When the motion occurs, the similarity of the data in the two sliding windows begins to decrease. When the motion ends, the similarity value is close to 1. The decision indicator can be calculated as

$$\eta(m)=1-\frac{1}{N}\sum_{1}^{N}\frac{\sum_{m}^{m+w}x_n(m)x_n(m+W)}{\sqrt{\sum_{m}^{m+w}x_n(m)^2}\sqrt{\sum_{m}^{m+w}x_n(m+W)^2}} \qquad (26)$$

where N is the number of antennas and W is the length of the sliding window. In this study, we set the length of the sliding window to 0.05s to ensure correct gesture segmentation. The decision indicator is as shown in the Fig.9 when the subject executes three different gestures in succession. In our experiment, the threshold is set as 0.01. Assuming that the decision indicator at m is greater than the threshold, the starting position of gesture is $m+W$.

D. Construct ADMs

In this part, we will introduce the method of construct ADM in detail. The signal processing flow is shown in the Fig.10.

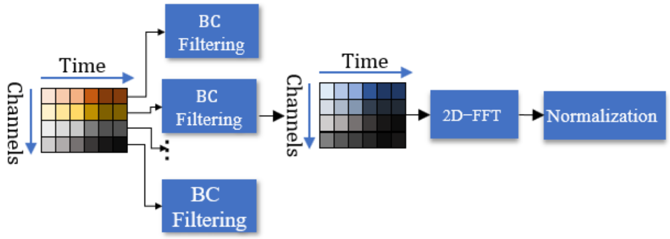


Fig. 10. Signal processing flowchart for constructing ADM.

1) *Background Cancelation*

In the experiment, the CW radar is subject to the stationary clutter reflections and the leakage between transmitter and receiver. Compared with static clutter, the signal reflected by the hand is very weak which will degrade the signal-to-noise ratio of the target's signal. Therefore, it is important to separate the target signal from the received signal. In this study, we suppress static clutter using the background subtraction method, which is based on the exponential average. Each channel needs to perform background subtraction processing to obtain accurate AOA information. The background mean at time m is calculated from the previous background mean $S_n(m\text{-}1)$ and the signal $x_n^{'}(m)$

$$S_n(m)=\rho x_n^{'}(m)+(1-\rho)S_n(m-1) \qquad (27)$$

where ρ is the exponential weighting factor. ρ is set to 0.95 in this paper. The signal after eliminating static interference is expressed as

$$x_n(m)=x_n^{'}(m)-S_n(m-1) \qquad (28)$$

2) *2D-FFT Construct ADM*

Assume that the incident angle of the gesture reflection signal is θ, the phase difference of adjacent antennas is $\omega=d\beta sin\theta$. Thus, the phase difference can be used to estimate the AOA of the target $\theta = arcsin\frac{\omega}{d\beta}$. We increase the dimension of the receive antenna by zero-padding from N to P, to avoid the fence effect of FFT. The FFT is performed along the antenna dimension to obtain the AOA. The angular resolution φ is given by

$$\varphi=\frac{1}{\cos\psi}\frac{50.8\lambda}{Nd}(^\circ) \qquad (29)$$

where ψ is the steering angle, N denotes the element number of the TMA, d is the distance of two adjacent antennas.

It is assumed that the speed of hand motion is constant within the sampling time T_s, the phase difference of echo should be $\omega=4\pi vT_s/\lambda$. When $\omega<\pi$ can make sure that the measured doppler is unambiguous. Therefore, the maximum unambiguous speed is

$$v_{\max}=\frac{\lambda}{4T_s} \qquad (30)$$

the resolution of Doppler-FFT determines the ability to distinguish the phase difference in $\omega 1$ and $\omega 2$, $|\omega 1\text{-}\omega 2|=4\pi\Delta vT_s/\lambda\geq 2\pi/N$, where N is the number of sampling points. Thus, the velocity resolution is

$$\vartriangle v=\frac{\lambda}{2NT_s} \qquad (31)$$

The ADM of the gesture is calculate using 2D-FFT

$$H(a,b)=\sum_{n=0}^{P}\left(\sum_{m=0}^{M}W_h\otimes x_n(m)e^{-j2\pi a\frac{m}{M}}\right)e^{-j2\pi b\frac{n}{P}} \qquad (32)$$

where W_h is the Hamming window function, ⊗ is an elementwise multiplication. To obtain continuous ADM, STFT with 0.2 s Hamming window and 80% overlapping between successive 2D-FFTs is performed on the signal. We draw a circle clockwise as an example and select 4 positions to demonstrate the change of ADMs with the movement of gestures as shown in Fig. 11. The blue arrows indicate the motion trajectory of the gesture. The abscissa and ordinate corresponding to the highlighted part of ADM indicate the speed and direction of the gesture, respectively. For example, when the hand moves to position 2, both the AOA and doppler of the gesture reach the maximum value in the opposite direction as shown in Fig. 11(c).

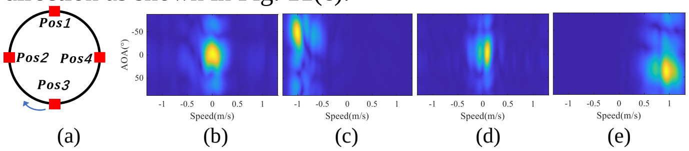


Fig. 11. ADMs of drawing a circle counterclockwise.

3) *ADM normalization*

In the actual measurement, there are many factors that affect the quality of ADM. For example, the reflected signal strength of different subjects and the adjustment of transmit and receive power will affect the value of ADM. In order to eliminate the influence of signal strength, we normalize each frame of ADMs. In addition, the speed of gesture execution by different users is different, which makes the number of ADM frames of gesture different. We fixed the number of frames N_f of ADMs so that they can contain complete gestures as much as possible. If the number of frames is greater than N_f, the extra frames will be discarded. Otherwise, zero padding is performed. After the process, the ADM becomes related to gestures only, and is input to the deep learning model.

E. Spatial Feature Extraction

The input to the network model is a sequence of ADMs, which is similar to the video streams. Each frame of ADMs represents the speed and direction of the gesture at the current moment. The continuous ADMs characterize changes in gestures over time. It is very necessary to mine the spatial and temporal features of ADM, which determines the robustness of the gesture recognition system.

CNN is widely used in the field of computer vision and has achieved great success, with its powerful spatial feature extraction ability. We use a shallow neural network consisting of two convolutional layers to extract the local features of each frame of ADM. After each convolutional layer, the Maxpooling layer and activation function are connected. Finally, the output of the last pooling layer is flattened into a 1D vector and used as the input of the temporal feature extraction.

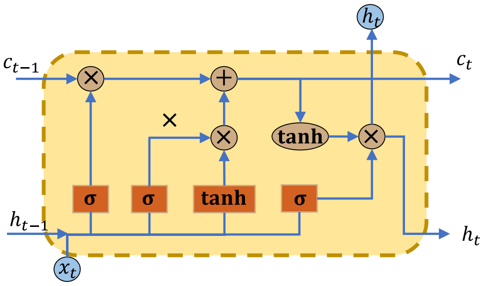


Fig. 12 Typical structure of LSTM.

In time series modeling, LSTM is widely used to deal with complex time dynamic sequences [51]. An LSTM cell consists of forgetting gate ($\mathbf{f}_t$), input gate ($\mathbf{i}_t$) and output gate ($\mathbf{o}_t$)as show in Fig. 12. $\mathbf{f}_t$ determines how much information of the cell state at time t-1 needs to be reserved to the present time. $\mathbf{i}_t$ controls how much information of the input at time t will be retained. $\mathbf{o}_t$ determines the output of the LSTM cell. The LSTM cell can be expressed as follow

$$\begin{aligned}
\mathbf{f}_t &= \sigma\left(\mathbf{W}_f\left[\mathbf{h}_{t-1}, \boldsymbol{x}_t\right]+\mathbf{b}_f\right) \\
\mathbf{i}_t &= \sigma\left(\mathbf{W}_i\left[\mathbf{h}_{t-1}, \boldsymbol{x}_t\right]+\mathbf{b}_i\right) \\
\tilde{\mathbf{c}}_t &= \tanh\left(\mathbf{W}_c\left[\mathbf{h}_{t-1}, \boldsymbol{x}_t\right]+\mathbf{b}_c\right) \\
\mathbf{c}_t &= \mathbf{f}_t \otimes \mathbf{c}_{t\text{-}1}+\mathbf{i}_t \otimes \tilde{\mathbf{c}}_t \\
\mathbf{o}_t &= \sigma\left(\mathbf{W}_o\left[\mathbf{h}_{t-1}, \boldsymbol{x}_t\right]+\mathbf{b}_o\right) \\
\mathbf{h}_t &= \mathrm{o}_t \otimes \tanh\left(\mathbf{c}_t\right)
\end{aligned} \tag{33}$$

where $[\mathbf{h}_{t\text{-}1}, \boldsymbol{x}_t]$ is a concatenation vector of the previously hidden state $\mathbf{h}_{t\text{-}1}$ and the current input. $\tilde{\mathbf{c}}_t$ and $\mathbf{c}_t$ are cell candidate and cell state, respectively. $\{\mathbf{W}_f, \mathbf{W}_i, \mathbf{W}_c, \mathbf{W}_o, \mathbf{b}_f, \mathbf{b}_i, \mathbf{b}_c, \mathbf{b}_o\}$ are weights and biases. The function of $\sigma(\cdot)$ and $tanh(\cdot)$ are sigmoid and tanh activation functions, respectively.

Compared with the traditional LSTM, Bi-LSTM can perform forward and backward processing on continuous ADM, while considering the past and future information of the data to extract richer features. Bi-LSTM contains two independent LSTMs as shown in the Fig. 13. Bi-LSTM's forward hidden state $\overrightarrow{\mathbf{h}}_t \in \mathbb{R}^{\mathrm{M}}$ and the backward hidden state $\overleftarrow{\mathbf{h}}_t \in \mathbb{R}^{\mathrm{M}}$ concatenate together to get output $\mathbf{h}_t = [\overrightarrow{\mathbf{h}}_t, \overleftarrow{\mathbf{h}}_t] \in \mathbb{R}^{2\mathrm{M}}$.

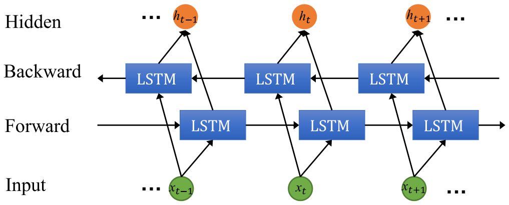


Fig. 13. structure of Bi-direction LSTM.

For continuous ADM, different features and time steps have different contributions to the final gesture recognition. Therefore, a hidden representation is constructed by integrating these scores to obtain better classification performance or improve the robustness of the model. In this study, the attention mechanism assigns appropriate weights to the extracted space-time features. First, average pooling is performed on the hidden layer state of the output of Bi-LSTM, which is equivalent to compressing the features of each time step into a scalar u_t. Then, the u_t is input to the activation function tanh to obtain the weight of each time step

$$u_t = \mathrm{avgpooling}\left(\mathbf{h}_t\right) \tag{34}$$

$$s_t = \sigma\left(u_t\right) \tag{35}$$

Thus, the final output hidden state $\mathbf{h}_{wt}$ is calculated as a weighted sum of all the hidden states $\mathbf{h}_t$

$$\mathbf{h}_{wt} = \sum_{t=1}^{N} s_t \mathbf{h}_t \tag{36}$$

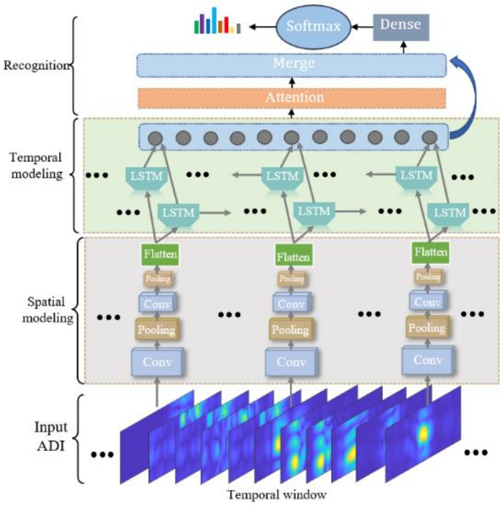


Fig. 14. Network architecture for spatial-temporal feature extraction.

The proposed network model is shown in the Fig. 14. Specifically, the input ADM series is a tensor with dimension as $T\times M\times P$, where T is the number of ADM snapshots. For the tth sampling ADM$_t$, the ADM$_t$ is fed into the CNN with two convolutional layers. Each layer of Conv2D contains 4 filters with a kernel size of 2×2 to extract the Doppler and AOA information of gestures. Each Cov2D is followed by the MaxPooling2D layer with a pooling size of 2×2, which halves the size of the feature map output by the previous layer of convolution filters and prevents overfitting. At the same time, the ReLU activation function behind the MaxPooling layer increases the nonlinear relationship between the layers of the neural network. Inputs output from CNN are fed into Bi-LSTM, and 16 neurons are used for each of the two LSTM layers. The attention mechanism generates a $1\times T$ weight vector to assign a weight to each time step. After the above feature extraction, the feature vector with a length of $32T$ is output and input to the output layer with 10 neurons to obtain the probability of each gesture. The details of the proposed network are shown in Table I.

In the training stage, the dropout mechanism is used to avoid our model overfitting. We used cross-entropy as the loss function. The Adam optimizer with a learning rate of 0.001 is chosen as the method of stochastic optimization. The size of a mini-batch is set as 32, and 20 epochs are selected empirically as the maximum number of epochs.

Table I
DETAILS OF EACH LAYER IN THE PROPOSED NETWORK, "CONV2D" REFERS TO THE CONVOLUTION LAYER, "K" DENOTES KERNEL SIZE, "C" DENOTES THE NUMBER OF OUTPUT CHANNELS, AND "N" DENOTES THE NUMBER OF NEURONS

CNN	Conv2D(k=2×2, C=4)
	Maxpooling(2)
	ReLU
	Conv2D(k=2×2, C=4)
	Maxpooling(2)
	ReLU

Bi-LSTM	LSTM(N=16) LSTM(N=16)
Attention	Global Average Pooling
	Sigmoid
Output layer	FC layer(N=10)

Table II
SYSTEM PARAMETERS

Parameters	Values
Carrier frequency F_c	5.8 GHz
Sampling frequency F_s	200 KHz
The passband of the LPF in DDC	0~200 Hz
Modulation Period	0.5 ms
USRP Tx gain	60 dB
USRP Rx gain	60 dB
Number of transmit antenna	4
Number of receive antenna	1
Antenna space	$\lambda/2$
Frame time	30 ms

V. EXPERIMENT AND ANALYSIS

The proposed sensing system consists of TMA, transceiver, MCU and computer. In order to quickly verify the feasibility of the system, we use USRP as the transceiver. It is worth noting that the expensive USRP can be replaced by low-cost wireless transceiver modules to achieve low-cost deployment. The TMA works in receiving mode, so it will not radiate infinite harmonics to space and will not interfere with other communication systems. We list detailed TMACW radar configuration parameters in Table II. After collecting data, we use MATLAB to perform the proposed signal processing on the received signal to obtain the gesture ADMs dataset. The system is implemented using Pytorch1.5 framework on a computer with NVIDIA GeForce GTX 1060 GPU and 32 GB RAM.

A. Dataset

In this paper, we collected 8 gesture data from 13 volunteers (9 male and 4 female) in 4 environments. Fig. 15 shows different indoor environments. Specifically, the dataset includes gestures commonly used in human-computer interaction, such as Pull (PL), Push (PS), Left Swipe (LS), Right Swipe (RS), closing of the fist (CF), Push-Pull (PP), Left-Right (LR), Zigzag (Z) and opening of the fist (OF) as shown in Fig. 17. The above gestures are made by full hand and the experimental scene is shown in the Fig. 16. We randomly selected the gesture data of 8 volunteers in the D region to form the in-domain dataset. The in-domain dataset contains 4000 gesture samples (8 users × 10 gestures × 50 times). We divide the in-domain dataset into training set and test set according to the ratio of 7:3. It is worth noting that the proposed network model is only trained by in-domain dataset. The gesture data of the remaining 5 volunteers constitute cross-people dataset which contains samples 1000 samples (5 users × 10 gestures ×20 times). In addition, we collected gesture data in areas A, B and C to form a cross-environment dataset, which includes 900 samples (3 Users ×3 environment ×10 gestures ×10 times). In order to reduce the impact of the torso on gestures, the radar transmits and receives signals in a vertical direction. In the above datasets, the distance between the volunteer's hand and the antenna array is about 20 to 30cm. The AOA of the hand relative to the radar is about 0° and allows a slight deviation. We collected 1500 samples (2 users×25 times×10 gestures×3) at a distance of 30 cm, 40 cm, and 50 cm and 1500 samples (2 users×25 times×10 gestures×3) at AOA of 15°, 30°, and 45° to construct the cross-position dataset as shown in Fig. 23. Each dataset is independent of each other. We only use the samples from the in-domain dataset to train the proposed neural network model. The trained neural network model is used to test multiple cross-domain data sets to verify the domain-independent sensing ability of our system.

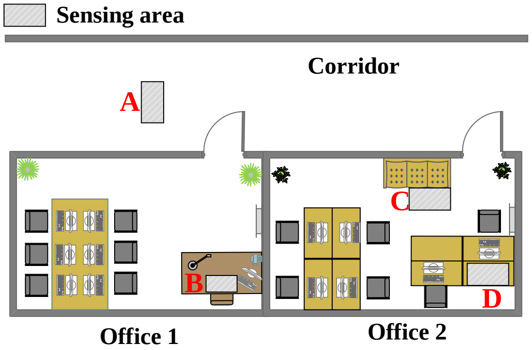


Fig. 15. The layout of four different environments in our gesture dataset.

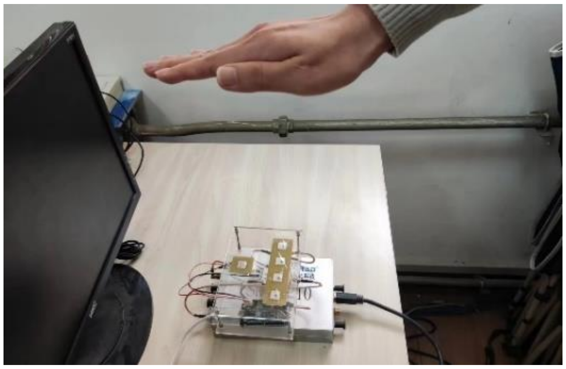


Fig. 16. Experimental scene

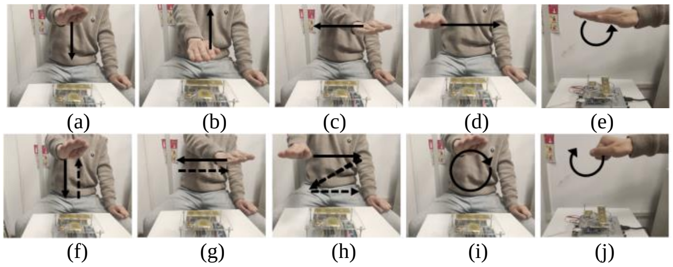


Fig. 17. All kinds of gestures. (a) Up (U). (b) Down (D). (c) Left swipe (L). (d) Right swipe (R). (e) closing of the fist (CF), (f) Up-Down (UD). (g) Left-Right (LR). (h) Zigzag (Z). (i) Circle clockwise (Cir). (j) opening of the fist (OF).

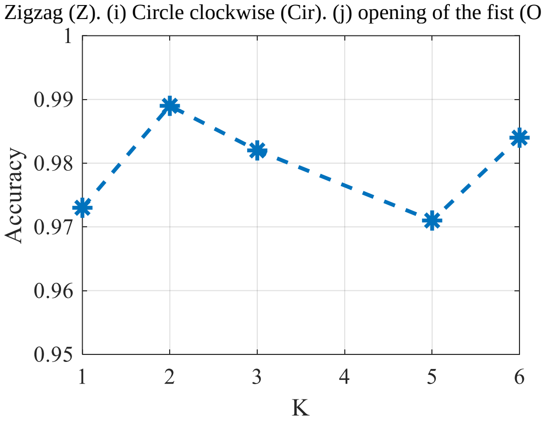


Fig. 18. The Impact of harmonic number on accuracy.

B. Impact of the harmonic number on gesture recognition

In section II B, it is known that the number of harmonic components and fundamental waves needs to be greater than the number of array elements to ensure accurate restoration of the conventional array. Most of the energy of the received signal is concentrated in the first few harmonic components. It is very necessary to explore the influence of harmonic number on gesture recognition accuracy. It can be seen from (14) that the Fourier coefficient of the harmonic component of integer multiples of 4 generated by the modulation method adopted in this study is 0, which means that these harmonics do not exist. Therefore, we consider and evaluate the recognition accuracy when K=1, 2, 3 and 5. Fig. 18 shows the influence of the number of harmonic components on the accuracy of gesture recognition. It can be found that the accuracy does not change significantly with the increase of the number of harmonics. The reason is that the energy of higher-order harmonics is relatively small and does not affect the harmonic recovery processing. In addition, recovering conventional array with more harmonics means that more digital down-conversion is required, which will consume a lot of CPU time and computing resources. In summary, we use the ±2nd harmonic, ± 1st harmonic and the fundamental wave to recover the conventional array.

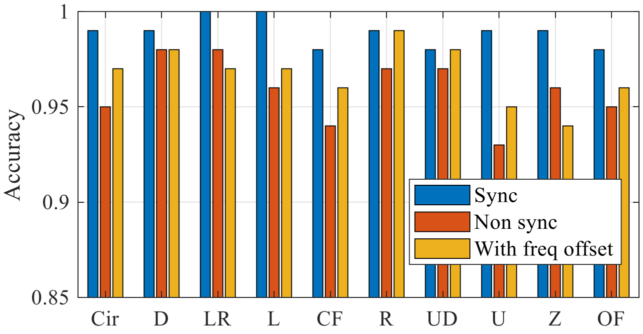


Fig. 19. The influence of non-synchronization and frequency offset on recognition accuracy.

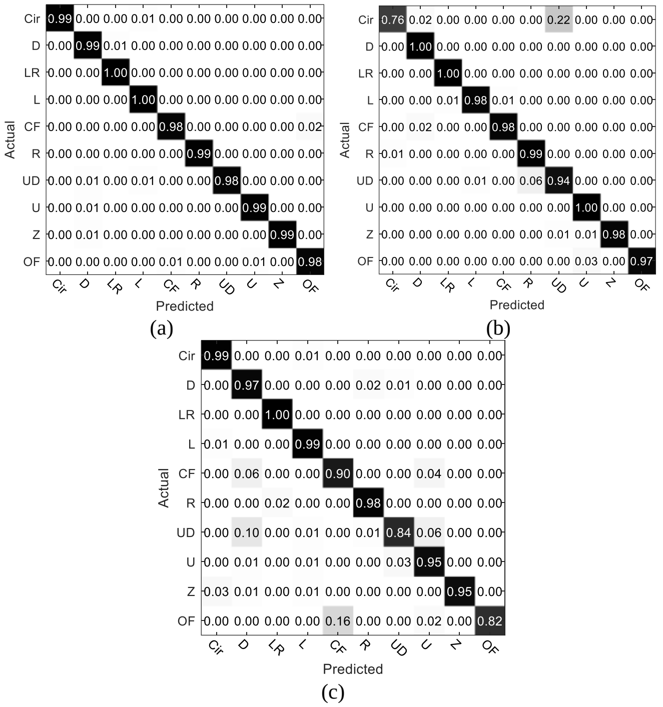


Fig. 20. Confusion matrices of different input. (a) Confusion matrices of ADM. (b) Confusion matrices of Doppler. (c) Confusion matrices of AOA.

C. Impact of non-synchronization and frequency offset

As described in Section III A and B, we need to complete system synchronization and frequency offset elimination to ensure the normal operation of the radar. We evaluated the impact of the above factors on the proposed system. In fact, it is very easy to use a spectrograph to obtain the modulation frequency, but there is a deviation of several hertz. The existence of frequency offset is equivalent to the harmonic vector multiplied by a time-varying diagonal array, which affects the recovery of traditional array. The recognition accuracy will be reduced by 2.2% when the system has no frequency calibration as shown in Fig. 19. Besides, the time delay caused by non-synchronization will lead to phase shift of each element in $\mathbf{\Gamma}$ which also affects the recovery of conventional arrays. We compare the recognition accuracy under the two conditions of synchronous and random determination of initial sampling time as shown in the Fig. 19.

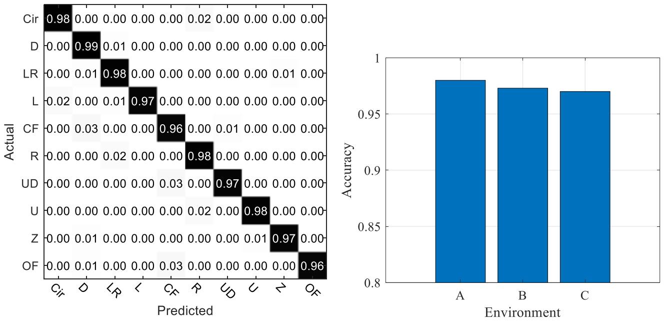
The AOA with errors leads to the distortion of the motion track of the highlighted part in the ADM. In the case of non-synchronization, the average recognition accuracy of gesture decreases by 4%. In summary, it is very important to calibrate the harmonic carrier frequency and synchronization for the hand gesture recognition system based on TMA.

D. Comparison of ADM, Doppler and AOA

In order to compare the impact of input type on gesture recognition accuracy, we extracted the doppler and AOA information of TMACW radar respectively. Specifically, the doppler spectrum and DOA spectrum are acquired by short-time Fourier transform (STFT) and multiple signal classification (MUSIC) algorithms, respectively. The sliding window length for calculating doppler spectrum and AOA spectrum is 0.1 s with 80% overlap. We feed the AOA spectrum and doppler spectrum to Resnet18 for training and testing, respectively. We use confusion matrix, which each column represents the instances in a predicted class and each row represents the instances in an actual class, to evaluate the performance of our system. The average gesture recognition accuracy using ADM, doppler and AOA is 98.9%, 96% and 93.9%, respectively. The confusion matrix for gesture recognition using only doppler information is shown in the Fig. 20(b). It is hard to obtain information about the horizontal motion of the hand from the doppler spectrum, which makes it difficult to distinguish between Cir and UD gestures. In addition, it is difficult to distinguish the motion in the vertical direction of the gesture using only AOA information, which makes it easy for the model to mistake UD for D as shown in Fig. 20(c). The ADM combined with AOA and doppler information can make up for the limitations of single domain features and effectively solve the above problem that similar gestures are difficult to distinguish as show in Fig. 20(a).

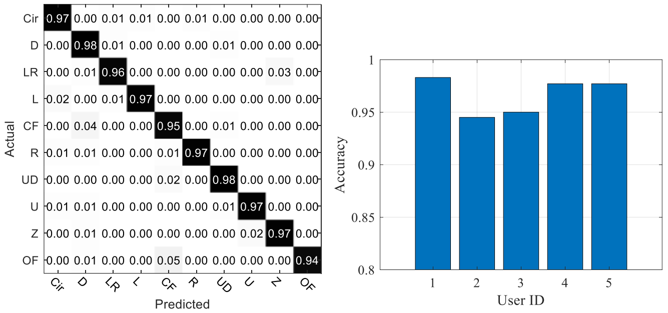
E. Impact of environment

Gesture recognition based on RF sensing is easily affected by the environment. Walls and furniture in the indoor environment cause rich multipaths during the propagation of wireless signals. In this experiment, we verified the robustness of the system in three environments, such as corridor, laboratory and meeting room. As shown in Fig. 21(b), our system can achieve 98%, 97.3%, and 97% average recognition accuracy in the corridor, laboratory and meeting room, respectively. The recognition accuracy of each gesture is not less than 97% as shown in Fig. 21(a). In general, the proposed system achieves high accuracy for different environments.



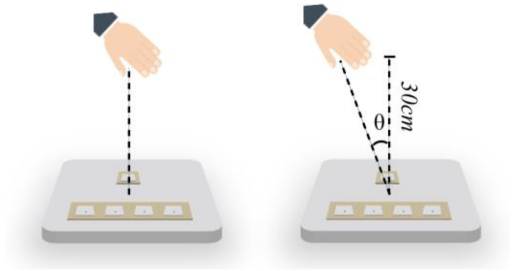
(a) (b)

Fig. 21. Gesture recognition results cross environments. (a) Confusion matrix (b) average recognition accuracy in different environments.



(a) (b)

Fig. 22. Gesture recognition results cross users. (a) Confusion matrix (b) gesture recognition accuracy on new users.



(a) (b)

Fig. 23. Experimental scene. (a) Different distances. (b) Different incident angles.

F. The impact of new users

Different people perform gestures in different ways, such as speed and range, which leads to large intra-class differences. Sample quality varies from person to person. To evaluate the performance of the proposed system on new users, we will test the trained model on a cross-person dataset containing 5 new users. It can be seen from Fig. 22(b) that the average recognition accuracy is 96.6%, and the recognition accuracy of the 5 people remained above 94.5% as shown in Fig. 22(b). The experimental results show that the proposed system has the ability of cross-user gesture recognition.

G. Impact of Position

In order to evaluate the effect of position on gesture recognition, we conducted a large number of experiments at different distances and different incident angles. First, the gestures with different distances are measured when the incident angle is fixed to zero. The distances between the hand and the radar are 30, 40, and 50 cm, respectively. The experimental scene is shown in Fig. 23 (a). Table III shows the classification accuracies for different distances. As the distance increases, the power of the hand echo decreases and the signal-to-noise ratio (SNR) is low, resulting in a decrease in the accuracy of gesture recognition. When the distance is 50cm, the recognition accuracy drops to 90.3% which can be still acceptable. In addition, the gestures with different incident angles are measured when the distance between the target hand and the antenna plane is fixed at 30 cm. The angles of the hand to the antenna plane are 15°, 30°, and 45°, respectively. The experimental scene is shown in Fig. 23 (b). Table IV shows the classification accuracies for different incident angle. It can be found that with the increase of incidence angles, the recognition accuracy decreases significantly. This is because the Doppler and AOA of the samples under this experimental setting are quite different from those of the samples in the in-domain dataset. When the incident angle of the gesture is greater than 45 °, there will be significant errors, which cannot be applied to HMI in this case.

Table III
THE RECOGNITION ACCURACY AT DIFFERENT DISTANCES

Distance	30 cm	40 cm	50 cm
Accuracy	98%	93.5%	90.3%

Table IV
THE RECOGNITION ACCURACY AT DIFFERENT ANGLES

Angle	15°	30°	45°
Accuracy	91%	80.2%	63.4%

Table V
COMPARISON OF DIFFERENT MODELS

Model	Model size (MB)	In domain	Cross people	Cross env.	Dis. within 50 cm	AOA within 30 °
Widar 3.0	1.001	96.2%	92.4%	93.6%	84.5%	80.5%
Model 1	0.126	97.3%	96.5%	95.4%	87.4%	82.6%
Model 2 (Proposed)	0.129	98.9%	96.7%	97.4%	93.9%	85.6%

env. means environment.
dis. means distance.

Table VI
COMPARISON OF DIFFERENT METHODS

	[12]	[23]	[13]	[16]	[29]	Ours
Tx signal	CW	OFDM	FMCW	FMCW	FMCW	CW
Channel	1T/4R	1T/6R	1T4R	3T4R	3T4R	**1T1R**
T/R antanna	1/4	1/6	1/4	3/4	3/4	1/4
Working Frequncy	10 GHz	5.8 GHz	24-26 GHz	60-64 GHz	60-64 Ghz	5.8 GHz
Range res.	NA	NA	7 cm	3 cm	3cm	NA
Max range	0.5 m	NA	0.6 m	1 m	NA	0.5 m
Cross-domain	NA	Yes	NA	NA	NA	**Yes**
Number of gestures	8	6	8	20	10	10
Acc. (%)	96.6	92.7	98.75	NA	95	**98.9**

res. means resolution.

H. Comparison with different network models

In order to demonstrate the superior domain-independent ability of the proposed system, we compared the proposed neural network with Widar 3.0. as shown in Table V. Model 2 represents the proposed neural network. Model 1 represents proposed neural network without the attention mechanism. The input of Widar3.0 is the body-coordinate velocity profile (BVP) sequence which is a series of two-dimensional matrices similar to our ADMs. The backbone of the Widar 3.0 is 2DCNN and Gated Recurrent Unit (GRU). We only modified the output layer of Widar 3.0 to fit our dataset. Widar3.0 is shallower than the proposed model, which makes the output layer output large number of features. Therefore, the last layer contains a large number of neurons, resulting in a larger model size than our models. All three models can achieve high-precision gesture recognition in In-domain dataset. In the cross-position scenario, we compared the recognition accuracy of all models under two conditions: distance within 50 cm and incident angle within 30 °. In the cross-domain sensing, it can be observed that the recognition accuracy of the proposed network model is higher than that of Widar 3.0. To verify the effectiveness of the proposed attention mechanism, we conduct ablation experiments. It can be found that the attention mechanism can focus on more important features and improve the generalization ability of the system.

I. Comparison with different time modulation functions

The Fourier coefficient matrix of the arbitrary time modulation function is derived. It should be noted that if TMA is restored to a traditional array, Fourier coefficient matrix Γ needs to satisfy the full rank. In order to explore the influence of different time modulation functions on the recognition accuracy, we generated three kinds of time modulation functions as shown in Fig. 23. $\boldsymbol{\Gamma}_a$, $\boldsymbol{\Gamma}_b$, and $\boldsymbol{\Gamma}_c$ are the Fourier coefficient matrices of U_a, U_b and U_c, respectively. $\boldsymbol{\Gamma}_a$ and $\boldsymbol{\Gamma}_b$ are full rank matrices, while the rank of $\boldsymbol{\Gamma}_c$ is 2. We collected 20 samples for each gesture under each time modulation function. The vertical distance between the hand and the radar is approximately 30cm. We use the Model 2 neural network to test the collected samples. The average recognition accuracy of gestures with different time modulation functions is shown in the Table VII. It can be found that $\boldsymbol{\Gamma}_a$ and $\boldsymbol{\Gamma}_b$ meet the conditions for array recovery, so the proposed method can accurately recognize gestures. However, $\boldsymbol{\Gamma}_c$ cannot accurately recover the array. The estimated AOA of the gesture is incorrect, further leading to a significant decrease in recognition accuracy.

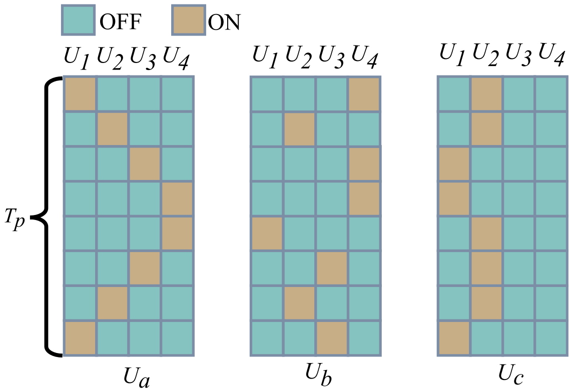


Fig. 23. Three kinds of time modulation functions

Table VII
COMPARISON OF DIFFERENT TIME MODULATION FUNCTIONS

Modulation function	U_a	U_b	U_c
Accuracy	97.5%	97%	66.5%

VI. DISCUSSION

A comparison with other similar works is listed in Table VI. Compared to multi-channel radar, the proposed system achieves the same performance using only a single channel. Compared with [12], [13], [16], and [29], our system has the ability of cross-domain sensing and high recognition accuracy, which benefits from the proposed domain independent feature and the designed neural network. In addition, we use TMA as a receiving array that does not radiate harmonics into space and does not affect communication signals in the same frequency band. Compared to radar in [13], [16], and [29] with wideband, the proposed system has lower spectrum occupancy.

There are two limitations to this study. First, the Tx signal in the proposed system is CW, which leads to a lack of ranging capability. The reason is that the original signal is repeated on the spectral axis with interval f_p. To avoid spectral aliasing between adjacent harmonic components, the bandwidth of the original signal must be less than f_p. Thus, we use CW as the Tx signal. Second, the efficiency of the proposed system is relatively low. The overall time-modulation efficiency is defined as $\eta=\eta_{\text{TMA}}\times\eta_{\text{s}}$ [52], where the efficiency $\eta_{\text{TMA}}=P_U/P_R$ represents the harmonic efficiency defined by the ratio of the power in useful harmonics to that in all harmonics. P_R is the total mean power of the TMA and is given by [53]

$$P_R = 4\pi \sum_{n=1}^{N} \left(t_{off,n} - t_{on,n} \right) T_p \tag{37}$$

where $t_{on,n}$ and $t_{off,n}$ represent the turn-on time and turn-off time of the nth channel of the RF switch, respectively. $P_U = \sum_{k=-K}^{K} P_k$ is the useful mean power. $P_k = 4\pi \sum_{n=1}^{N} \left| a_{n,k} \right|^2$ is the mean power at the kth harmonic [54]. Then, the η_{TMA} can be written as

$$\eta_{\text{TMA}} = \frac{\sum_{-K}^{K} \sum_{n=1}^{N} \left| a_{n,k} \right|^2}{\sum_{n=1}^{N} \left(t_{off,n} - t_{on,n} \right) T_p} \tag{38}$$

We set K=2 in this study. η_s is the switched feeding network efficiency and is given by

$$\eta_{\text{s}} = \frac{1}{N} \sum_{n=1}^{N} D_n \tag{39}$$

where $D_n = 1/4$ is the duty cycle of the excitation of the nth channel of the RF switch in this paper. Therefore, D_n=1/N=1/4 and $\eta=\eta_{\text{TMA}}\times\eta_{\text{s}}\approx 0.215$ in this study. The efficiency of the constructed TMA is relatively low, which may be the reason for the shorter detection distance compared to other multi-channel radar schemes. In future work, the SP4T RF switch will be replaced by a power splitter equipped with phase shifter to improve the efficiency of TMA [54].

VII. Conclusion

In this paper, a novel single-channel TMACW radar is proposed to enable cross-domain gesture recognition. The harmonic generated by time modulation is used to restore the TMA to the traditional array, which reduces the hardware complexity and cost of multi-channel radar. The 2D-FFT, normalization, and clutter suppression are used to obtain robust the ADM for gestures. Then we develop a neural network with attention mechanism to fully exploit spatial-temporal characteristics of ADM for gesture recognition. The experimental results show that our system can achieve high accuracy gesture recognition across different domains (i.e. environments, users, and positions). In the future, we believe that the proposed low-cost TMA sensing system can not only be applied to gesture recognition, but also promote other sensing tasks.

References

[1] P. Molchanov et al., “Online detection and classification of dynamic hand gestures with recurrent 3d convolutional neural networks,”, *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, pp. 4207-4215, 2016.

[2] P. Narayana, J. R. Beveridge, and B. A. Draper, “Gesture recognition: Focus on the hands,” in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5235–5244.

[3] M. Z. Islam, L. Yu, H. Abuella, J. F. O'Hara, C. Crick and S. Ekin, “Hand Gesture Recognition through Reflected Infrared Light Wave Signals,” *in 2023 10th International Conference on Electrical and Electronics Engineering (ICEEE)*, Istanbul, Turkiye, 2023, pp. 1-5.

[4] J. Gong, Y. Zhang, X. Zhou and X.-D. Yang, “Pyro: Thumb-tip gesture recognition using pyroelectric infrared sensing”, *Proc. 30th ACM Annu. ACM Symp. User Interface Softw. Technol*., pp. 553-563, 2017.

[5] B. Fang et al., “Dynamic gesture recognition using inertial sensors-based data gloves,” *in Proc. IEEE 4th Int. Conf. Adv. Robot. Mechatronics* (ICARM), 2019, pp. 390–395.

[6] Liu Y T, Zhang Y A, Zeng M. Y.-T. Liu, Y.-A. Zhang and M. Zeng, "Novel algorithm for hand gesture recognition utilizing a wrist-worn inertial sensor", *IEEE Sensors J.*, vol. 18, no. 24, pp. 10085-10095, Dec. 2018.

[7] C. Xu, P. H. Pathak, and P. Mohapatra, “Finger-writing with smartwatch: A case for finger and hand gesture recognition using smartwatch,” in *Proc. ACM Int. Workshop Mobile Comput. Syst*. Appl., 2015, pp. 9–14.

[8] X. Zeng, C. Wu and W.-B. Ye, "User-definable dynamic hand gesture recognition based on Doppler radar and few-shot learning", *IEEE Sensors J.*, vol. 21, no. 20, pp. 23224-23233, Oct. 2021.

[9] Y. Kim and B. Toomajian, “Hand gesture recognition using microDoppler signatures with convolutional neural network,” *IEEE Access*, vol. 4, pp. 7125–7130, 2016.

[10] P. Qifan, S. Gupta, S. Gollakota, and S. Pate, “Whole-home gesture recognition using wireless signals,” Comput. Commun. Rev., vol. 43, no. 4, pp. 485–486, 2013.

[11] S. Skaria, A. Al-Hourani, M. Lech and R. J. Evans, "Hand-gesture recognition using two-antenna Doppler radar with deep convolutional neural networks", *IEEE Sensors J.*, vol. 19, no. 8, pp. 3041-3048, Apr. 2019.

[12] A. A. Pramudita, “Contactless hand gesture sensor based on array of CW radar for human to machine interface,” *IEEE Sensors J.*, vol. 21, no. 13, pp. 15196–15208, Jul. 2021.

[13] L. Qu, H. Wu, T. Yang, L. Zhang, and Y. Sun, “Dynamic hand gesture classification based on multichannel radar using multistream fusion1-D convolutional neural network,” *IEEE Sensors J.*, vol. 22, no. 24, pp. 24083–24093, Dec. 2022.

[14] 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 Sensors J.*, vol. 20, no. 18, pp. 10706–10716, May 2020.

[15] X. Dong, Z. Zhao, Y. Wang, T. Zeng, J. Wang and Y. Sui, "FMCW radar-based hand gesture recognition using spatiotemporal deformable and context-aware convolutional 5-D feature representation", *IEEE Trans. Geosci. Remote Sens*., vol. 60, pp. 1-11, 2022.

[16] Y. Li, C. Gu and J.-F. Mao, "4-D gesture sensing using reconfigurable virtual array based on a 60 GHz FMCW MIMO radar sensor," *IEEE Trans. Microw. Theory Techn.*, vol. 70, no. 7, pp. 3652-3665, Jul. 2022.

[17] H. Abdelnasser, M. Youssef and K. A. Harras, "WiGest: A ubiquitous WiFi-based gesture recognition system", *Proc. IEEE INFOCOM*, 2015.

[18] M. A. A. Haseeb and R. Parasuraman, “Wisture: Touch-less hand gesture classification in unmodified smartphones using Wi-Fi signals,” *IEEE Sensors J.*, vol. 19, no. 1, pp. 257-267, Jan. 2019.

[19] H. Abdelnasser, K. Harras and M. Youssef, “A Ubiquitous WiFi-Based Fine-Grained Gesture Recognition System,” *IEEE Trans. Mobile Comput.*, vol. 18, no. 11, pp. 2474-2487, 1 Nov. 2019.

[20] Wenfeng He, Kaishun Wu, Yongpan Zou, and Zhong Ming. “Wig: Wifi-based gesture recognition system,” In Computer Communicationand Networks (ICCCN), 2015 24th International Conference on.

[21] H. Zou, Y. Zhou, J. Yang, H. Jiang, L. Xie and C. J. Spanos, “WiFi-enabled device-free gesture recognition for smart home automation,” *Proc. IEEE 14th Int. Conf. Control Autom*. (ICCA), pp. 476-481, 2018.

[22] N. Yu, W. Wang, A. X. Liu and L. Kong, “QGesture: Quantifying gesture distance and direction with WiFi signals,” *Proc. ACM Interact. Mobile Wearable Ubiquitous Technol*., vol. 2, no. 1, pp. 1-23, 2018.

[23] Y. Zhang et al., “Widar3.0: Zero-effort cross-domain gesture recognition with Wi-Fi,” *IEEE Trans. Pattern. Anal. Mach. Intell*., vol. 44, no. 11, pp. 8671-8688, Nov. 2022.

[24] K. Niu, F. Zhang, X. Wang, Q. Lv, H. Luo and D. Zhang, "Understanding WiFi signal frequency features for position-independent gesture sensing, " *IEEE Trans. Mobile Comput.*, Mar. 2021.

[25] Z. Zhou, Z. Cao, and Y. Pi, ''Dynamic gesture recognition with a terahertz radar based on range profile sequences and Doppler signatures,'' *Sensors*, vol. 18, no. 1, p. 10, Dec. 2017.

[26] S. Zhang, G. Li, M. Ritchie, F. Fioranelli and H. Griffiths, "Dynamic hand gesture classification based on radar micro-Doppler signatures," *Proc. CIE Int. Conf. Radar* (RADAR), pp. 1-4, Oct. 2016.

[27] K. A. Smith, C. Csech, D. Murdoch and G. Shaker, "Gesture recognition using mm-wave sensor for human-car interface," *IEEE Sens. Lett*., vol. 2, no. 2, pp. 1-4, Jun. 2018.

[28] Wang Z, Li G, Yang L. "Dynamic hand gesture recognition based on micro-doppler radar signatures using hidden Gauss–Markov models," *IEEE Geoscience and Remote Sensing Letters*,vol. 18,no. 2, pp. 291-295, 2020.

[29] S. Ahmed, W. Kim, J. Park, et al., "Radar based air-writing gesture recognition using a novel multi-stream CNN approach," *IEEE Internet of Things Journal*, 2022.

[30] 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, Jun. 2021.

[31] S. Wang, J. Song, J. Lien, I. Poupyrev, and O. Hilliges, "Interacting with soli: Exploring fine-grained dynamic gesture recognition in the radiofrequency spectrum," in *Proc. 29th Annu. Symp. User Interface Softw. Technol.* (UIST). New York, NY, USA: ACM, 2016, pp. 851–860.

[32] Z. Zhang, Z. Tian and M. Zhou, "Latern: Dynamic continuous hand gesture recognition using FMCW radar sensor," *IEEE Sensors J.*, vol. 18, no. 8, pp. 3278-3289, Apr. 2018.

[33] S. Yang, Y. B. Gan, and A. Qing, "Sideband suppression in time modulated linear arrays by the differential evolution algorithm," *IEEE Antennas Wireless Propag. Lett.*, vol. 1, no. 1, pp. 173–175, 2002.

[34] G. R. Hardel, N. T. Yallaparagada, D. Mandal, and A. K. Bhattacharjee, "Introducing deeper nulls for time modulated linear symmetric antenna array using real coded genetic algorithm," in 2011 *IEEE Symposium on computers & informatics*, 2011, pp. 249–254, IEEE.

[35] L. Poli, P. Rocca, L. Manica and A. Massa, "Handling sideband radiations in time-modulated arrays through particle swarm optimization," *IEEE Trans. Antennas Propag.*, vol. 58, no. 4, pp. 1408-1411, Apr. 2010.

[36] F. Yang, S. Yang, Y. Chen, S. Qu, and J. Hu, "Effcient pencil beam synthesis in 4-D antenna arrays using an iterative convex optimization algorithm," *IEEE Trans. Antennas Propag.*, vol. 67, no. 11, pp. 6847–6858, Nov. 2019.

[37] C. Z. Feng, W. T. Li, C. Cui, Y. Q. Hei, J. C. Mou and X. W. Shi, "An efficient and universal static and dynamic convex optimization for array synthesis," *IEEE Antennas Wireless Propag. Lett.*, vol. 21, no. 10, pp. 2060-2064, Oct. 2022.

[38] Y. -X. Zhang, Y. -C. Jiao and L. Zhang, "Efficient directivity maximization of time-modulated arrays with two-stage convex optimization," *IEEE Antennas Wireless Propag. Lett.*, vol. 19, no. 11, pp. 1847-1851, Nov. 2020

[39] F. Yang, S. Yang, M. Huang, and L. Wang, "Synthesis of large nonuniform spaced 4D linear arrays using an iterative FFT method," *in Proc. Int. Conf. Comput. Electromagn* (ICCEM), 2018, pp. 1–2.

[40] Y. Liu, J. Bai, J. Zheng, H. Liao, Y. Ren and Y. J. Guo, "Efficient shaped pattern synthesis for time modulated antenna arrays including mutual coupling by differential evolution integrated with FFT via least-square active element pattern expansion," *IEEE Trans. Antennas Propag.*, vol. 69, no. 7, pp. 4223-4228, Jul. 2021.

[41] C. Zhao, Y. Chen, Y. Feng and S. Yang, "Efficient Synthesis of Large-Scale Time-Modulated Antenna Arrays Using Artificial Neural Network and Inverse FFT," early Access in *IEEE Trans. Antennas Propag.*, doi: 10.1109/TAP.2023.3340349.

[42] Y. Q. Hei, L. Y. Ma, W. T. Li, J. C. Mou and X. W. Shi, "Effective Artificial Neural Network Framework for Time-Modulated Arrays Synthesis," *IEEE Trans. Antennas Propag.*, vol. 71, no. 10, pp. 7728-7740, Oct. 2023.

[43] H. Li, Y. Chen and S. Yang, "Harmonic beamforming in antenna array with time-modulated amplitude-phase weighting technique", *IEEE Trans. Antennas Propag.*, vol. 67, no. 10, pp. 6461-6472, Oct. 2019.

[44] Y. Liu, C. He, X. Liang, *et al.*, "Multiuser Communication by Electromagnetic Vortex Based on Time-Modulated Array," *IEEE Antennas Wirel. Propag. Lett*., vol. 19, no. 2, pp. 282-286, Feb. 2020.

[45] C. Shan, J. Shi, Y. Ma, *et al.* "Power loss suppression for time-modulated arrays in radar-communication integration," *IEEE Journal of Selected Topics in Signal Processing*, vol. 15, no. 6, pp. 1365-1377, 2021.

[46] Huang G, Ding Y, Ouyang S, *et al.*,"Target Localization Using Time-Modulated Directional Modulated Transmitters," *IEEE Sensors J*., vol. 22, no. 13, pp. 13508-13518. 2022.

[47] W. T. Li, Y. J. Lei, and X. W. Shi, "DOA estimation of time-modulated linear array based on sparse signal recovery," *IEEE Antennas Wirel. Propag. Lett.*, vol. 16, pp. 2336–2340, 2017.

[48] F. Yang, S. Yang, L. Sun, et al., "DOA estimation via sparse signal recovery in 4-D linear antenna arrays with optimized time sequences," *IEEE Trans. Veh. Technol.*, vol. 69, no. 1, pp. 771–783, 2020.

[49] G. Ni, C. He, Y. Liu, J. Chen and R. Jin, "Direction-finding based on time-modulated array without sampling synchronization," *IEEE Antennas Wirel. Propag. Lett*, vol. 19, no. 12, pp. 2149-2153, Dec. 2020.

[50] G. Ni, C. He, J. Chen, L. Bai and R. Jin, "Direction finding and performance analysis with 1 bit time modulated array," *IEEE Trans. Antennas Propag.*, vol. 69, no. 10, pp. 6881-6893, Oct. 2021.,

[51] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[52] R. Maneiro-Catoira, J. C. Brégains, J. A. García-Naya, L. Castedo, P. Rocca and L. Poli, "Performance analysis of time-modulated a*rrays for the angle diversity reception of digital linear modulated signals," IEEE J. Sel. Topics Signal Process*., vol. 11, no. 2, pp. 247-258, Mar. 2017.

[53]Gassab O, Azrar A, Dahimene A, et al. "Efficient electronic beam steering method in time modulated linear arrays," *IET Microwaves Antennas Propag.*, vol. 14, no.5, pp. 402-408. Feb. 2020.

[54] G. Ni, C. He, Y. Gao, J. Chen, and R. Jin, "High-efficiency modulationand harmonic beam scanning in time-modulated array,"*IEEE Trans.Antennas Propag*., vol. 71, no. 1, pp. 368–380, Jan. 2023.