


ORIGINAL RESEARCH

Adaptive framework towards radar-based diversity gesture recognition with range-Doppler signatures

Liyang Wang  | Zongyong Cui | Yiming Pi | Changjie Cao | Zongjie Cao

School of Information and Communication Engineering, University of Electronic Science and Technology of China, Chengdu, China

Correspondence

Zongjie Cao, School of Information and Communication Engineering, University of Electronic Science and Technology of China, Xiyuan Ave, West Hi-Tech Zone, Chengdu, Sichuan 611731, China.

Email: zjcao@uestc.edu.cn

Funding information

National Natural Science Foundation of China, Grant/Award Number: 61971101

Abstract

Radar-based hand gesture recognition (HGR) has attracted growing interest in human-computer interaction. A rich diversity in how people perform gestures causes a large intra-class variance, and the sample quality varies from person to person. It makes HGR more challenging to identify dynamic, complicated, and deforming hand gestures. It is urgent for the real world to explore a robust method that better identifies the gestures from non-specified users. To address the above issues, an adaptive framework is proposed for gesture recognition, and it has two main contributions. First of all, a trajectory range Doppler map (t-RDM) is obtained by non-coherent accumulating for inter-frame dependencies, and then t-RDM is enhanced to highlight the trajectory information. Taking into account different movement patterns of the gestures, a two-pathway convolutional neural network targeted for raw and enhanced t-RDMs is proposed, which independently mines discriminative information from the two t-RDMs with different salient features. Second, an adaptive individual cost (AIC) loss is proposed, aiming to establish a powerful feature representation by adaptively extracting the commonalities in variant gestures according to the sample quality. Based on a public dataset using soli radar, the proposed method is evaluated on two tasks: cross-person recognition and cross-scenario recognition. These two recognition modes require that the training set and the test set are mutually exclusive not only at the sample level but also at the source level. Extensive experiments demonstrate that the proposed method is superior to the existing approaches for alleviating the low recognition performance caused by gesture diversity.

KEYWORDS

gesture recognition, human-computer interaction, neural network, radar-based

1 | INTRODUCTION

In the interactive framework, the vigorous development of intelligent computing has driven the advancements in human-computer interaction (HCI), and an effective interaction interface is assuming importance in our daily lives [1]. Hand gesture recognition (HGR), as a non-verbal mode, is regarded as a natural way for HCI. It enables the users to control devices in a non-contact way without using a mouse, keyboard, and so on. [2].

Due to the user-friendly and expressive interactive interface, there is a large amount of research emerging on gesture

recognition. From the perspective of sensor types, these studies can be roughly divided into three categories, namely optical-based, inertial-based, and radar-based gesture recognition [3–8]. Among them, the technology of optical-based gesture recognition has achieved the most remarkable achievements [9], but the changeable ambient light and inevitable privacy leakage limit its ubiquitous application. Inertial elements capture the fine gesture movement accurately, however, the annoying wires and complicated equipment make it unfriendly to users. Radar, as an alternative option, compensates well for the inherent shortcomings of the first two sensors, and it is more robust to ambient light and is free from the

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2022 The Authors. *IET Radar, Sonar & Navigation* published by John Wiley & Sons Ltd on behalf of The Institution of Engineering and Technology.

constraints of wearable devices. Thanks to the boom in microwave engineering, radars have been widely used in the HGR field [10–13].

Although abundant research studies exist on radar-based HGR, the technology of device-free gesture recognition still has a long way to develop. Traditionally, radar systems are promoted by the applications such as detection, localisation, and identification of aircraft, ships, and other rigid targets [14]. These applications are quite different from identifying complex, dynamic, deforming hand shapes at close range [15]. In contrast to rigid targets, the hand has multiple degrees of freedom (DoF) and its shape may deform continuously during the motion. Consequently, the radar signals of multiple gestures are distinct, and the sample quality varies from person to person.

Intuitively, the individual sensitivity of gestures would lead to a large intra-class variance, and a large intra-class variance would degrade the performance of the recognition algorithm. Moreover, the gap within a gesture class would become more evident in the evaluation method that the training and test sets come from different data sources (people or scenarios). For the sake of description, these evaluation methods of the training set and test set from different people or scenarios are noted as cross-person or cross-scenario. The recognition using evaluation methods such as hold-out or cross-validation is referred to as traditional recognition. Traditional recognition requires the training set and test set to be mutually exclusive at the sample level, regardless of the sample sources.

From the results of the existing studies, it appears that the degraded recognition performance caused by the variant gestures from different sources has been demonstrated. Wang et al., the researchers from Google ATAP, found that the recognition accuracy is 87% in the traditional evaluation mode based on millimetre-wave radar. Whereas, this performance decreased to 85.75% in cross-scenario recognition and more severely to 79% in cross-person recognition [16]. Based on a frequency modulated continuous wave (FMCW) radar, Zhang et al. found that the recognition accuracy is 72.5% when using the fifth person samples as the test set. Compared with the accuracy of 96% in traditional recognition, the cross-person recognition performance decreases by nearly 24% [17]. Similarly, the issue of degraded recognition performance also occurred in the study by Berenguer et al. [18].

The above research results show that a recognition algorithm performs poorly in cross-person or cross-scenario recognition, even though it achieves a desirable recognition accuracy in traditional evaluation methods. Hold-out (e.g., training set: test set = 70%:30%) or cross-validation (e.g., fivefold) are commonly adopted as the traditional evaluation tools, and the pros or cons of the model on the test samples are considered to be convincing results due to the sample randomness [19]. Generally, this evaluation approach is considered fair to evaluate the model's performance when the random peculiarities of samples are disorder and irregular. Whereas, radar-based gesture signals are highly correlated with individual characteristics. The source characteristics of the gesture samples cannot be ignored. As shown in Figure 1, the

samples from different sources tend to show a noticeable variance in the data distribution. Specifically, the same type of gestures from different sources differ significantly, such as $G_{1,1}$, $G_{1,2}$, and $G_{1,3}$. Even worse, the variant samples are likely to be confused with other categories, such as $G_{1,1}$ and $G_{2,1}$. It is not sufficient for an evaluation method that does not consider the source characteristics to evaluate the algorithm performance in gesture recognition. Cross-person and cross-scenario evaluation modes undoubtedly amplify the impact of source characteristics on recognition performance. However, to evaluate the performance of a non-specified and unknown person or scenario is critical for gesture recognition in the real world. Unfortunately, certain existing studies only note this challenge in cross-person (cross-scenario) gesture recognition based on radar, but there are no efforts specifically dedicated to addressing this issue.

In this study, we are committed to alleviating the low recognition performance caused by the individual diversity of radar-based gesture signals. To this end, an adaptive gesture recognition framework is proposed considering the differences of gesture samples. The proposed method is based on range Doppler map (RDM), which is widely applied in radar-based gesture recognition. First of all, a trajectory RDM (t-RDM) is proposed to establish the interdependencies of the observed frames, which can effectively record the movement trajectory and improve the robustness of samples. Then, an enhanced t-RDM is proposed to highlight the trajectory in the raw t-RDMs. Subsequently, a two-pathway convolutional neural network (CNN) is proposed with dual-channel inputs especially for the t-RDM and enhanced t-RDM, which captures the fine-grained signatures and coarse-grained signatures of the gestures independently. Additionally, adaptive individual cost (AIC) loss is proposed considering that the gesture performing habits of different people are diverse. AIC loss allows the recognition model to adaptively mine the commonalities in the variant gestures according to the sample quality. The performance of the proposed method is demonstrated on a publicly available dataset using a Soli radar.

The main contributions of this work are summarised as follows:

- Aiming at the impact of individual variations and the characteristics of RDMs, a framework of adaptive gesture recognition is proposed, which can effectively improve the model's generalisation performance on the gestures from unknown data sources.
- Combined with the motion characteristics of fine-grained and coarse-grained gestures, a raw t-RDM and an enhanced t-RDM is proposed to improve the sample quality from the root of data. Furthermore, a two-pathway CNN is proposed to extract the features of the raw and enhanced t-RDMs independently.
- AIC loss is proposed. Considering the diversity of different source gestures, AIC loss allows the recognition model to adaptively capture the common signatures from various sources based on the sample quality, which facilitates the acquisition of clustered intra-class features.

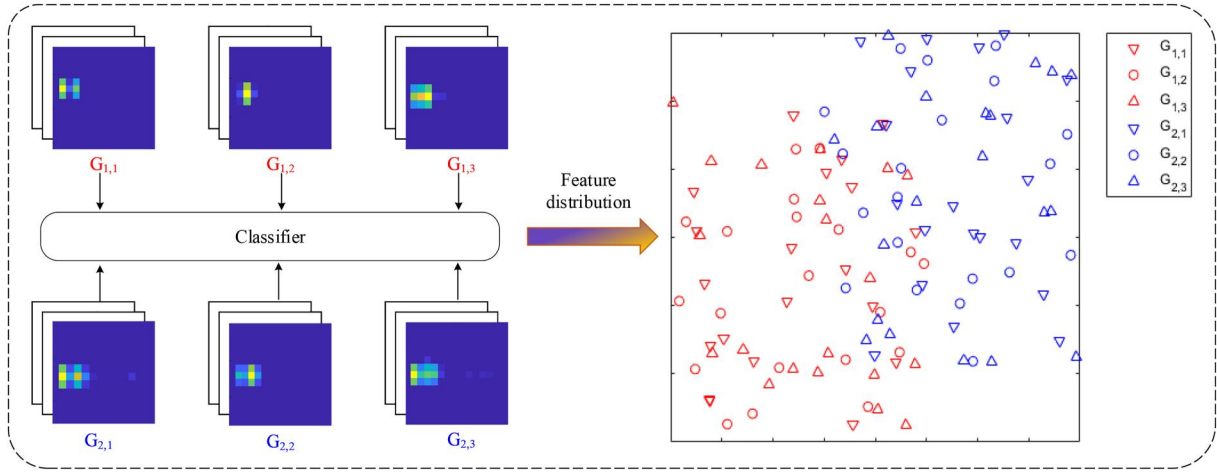


FIGURE 1 Schematic diagram of the impact of various sources on gesture recognition. Suppose that three people (scenarios) are involved in the recognition of two types of gestures. The sample features extracted by the classifier are represented as scatters, and the scatter shape indicates its data source, and the colour indicates the gesture category. G_{ij} is i th gesture signals from the j th source

2 | RELATED WORK

2.1 | Existing studies in radar-based HGR

A complete gesture recognition process is mainly divided into four steps based on radar, which are predefinition of gesture categories, data acquisition, signal processing, and recognition algorithm. First, some commonly used gestures are taken into consideration in the predefinition of gesture categories, such as sliding left or right [20]. To provide an interactive feeling, Google researchers designed some gestures with two fingers moving relative to each other, and the friction makes the user perceive the gesture movement more clearly [15, 16]. However, the complicated gestures pose a challenge to sensor sensing. Second, the FMCW radar is widely used in the data acquisition of gesture recognition. Compared with the pulsed radar, the FMCW radar has a more powerful sensing ability, and it can measure not only distance but also the speed of targets [21]. Subsequently, time-range [10], time-Doppler [22, 23] and range-Doppler (it is also noted RDM) [8] are typical signatures applied in radar signal processing to characterise the gestures uniquely. Finally, the vast majority of studies in the HGR field have focussed on recognition algorithms. These algorithms are mainly dedicated to improving the accuracy or reducing the response latency of gesture recognition etc [2, 24].

Thanks to good interpretability, handcrafted features have been prevalent in the field of gesture recognition, and they usually highlight the discriminative gesture signatures through mathematical formulas. Excellent recognition results reveal that handcrafted features are a good bridge to integrate the expert experience into the machine learning [25, 26]. However, considerable domain expertise and tedious feature screening limit the versatility of handcrafted features in specific scenarios. With the rise of neural networks, an effective and automatic feature extraction method has broken into the vision of researchers [27]. Numerous studies have shown that neural networks perform outstandingly in radar-based gesture

recognition. Kim et al. applied a deep 2-D CNN to recognise the Doppler features of 10 gestures, and its overall recognition rate reached 85.6% [28]. Hazra et al. adopted a depth 3-D CNN to recognise 6 designed gestures, and the overall accuracy achieved 94.5% [29]. Lei et al. employed the network of DS-3DCNN-LSTM, which is 3-D CNN combined with LSTM, to recognise 10 kinds of gestures based on their RDMs and Range Azimuth Maps, and it achieved 97.66% recognition accuracy [30].

2.2 | Existing studies in diversity target identification

As mentioned above, both the hand's flexibility and specific environmental signatures encoded in the wireless signals pose challenges to the robustness of gesture recognition algorithms. The device-free recognition of human activities, which have similar characteristics to the hand, also faces these challenges. As we know, human activities show a high degree of diversity, and even the activities from the same performer in the same environment may differ. In addition, wireless signals provide substantial target information that is specific to the environment where the activities occurred. These factors lead to the fact that an algorithm recognises the human activities from a specific person well in a specific scene but does not work well in another scene or from another new person.

These issues have been noticed by some researchers in human activity recognition via wireless sensing. Jiang et al. proposed a deep model targeted for extracting the robust discriminative signatures [31]. First, a CNN is adopted in this model as a feature extractor and mines the informative features. Then, the extracted features are, respectively, fed into an activity recogniser and a domain discriminator. The former is expected to maximise the predicted possibilities of the correct activity labels, and the latter module aims to obscure the domain characteristics of the targets, that is, minimise the

possibilities of the predicted domain labels. Finally, this method is demonstrated to capture effective domain-independent features and accomplish a robust activity recognition with the varying surroundings. To address the diversity activity recognition through WiFi sensing, Guo et al. proposed an LCED model to weaken the individual divergences in human activities, which is combined with long short time memory (LSTM) network and CNN [32]. First, an LSTM module is applied to capture the time-varying features, and the features contain individual signatures encoded in the WiFi signals. Then, a deep CNN further extracts the generalised features considering activity recognition rather than individual identification. LCED model is claimed to achieve a robust accuracy of 95% in 16 diverse activity recognition.

The common thing in the existing studies is that the recognition algorithms are dedicated to constructing a powerful and domain-independent feature representation. However, the above solutions with complicated recognition networks have to rely on high-precision equipment to collect a large amount of high-quality data. To some extent, it limits the widespread use of some complicated methods in diversity target recognition based on wireless sensing.

3 | THE ALGORITHM OF THE ADAPTIVE GESTURE RECOGNITION FRAMEWORK

3.1 | Two-pathway CNN targeted for raw and enhanced trajectory RDMs

RDMs characterise the targets uniquely with respect to the range and velocity, and it can be obtained by 2-D Fourier transform on a certain number of consecutive radar signals. A dynamic gesture usually contains multiple RDMs, and the RDMs encode the hand motion signatures at different moments. Considering the fact that the hand position is unique during its motion, a t-RDM is established by non-coherent accumulating on the previous RDMs. Specifically, t-RDM can be expressed as follows:

$$\begin{aligned} t\text{-RDM}_i &= \text{RDM}_i + t\text{-RDM}_{i-1}, \quad i \in [2, \dots, N], \\ t\text{-RDM}_1 &= \text{RDM}_1 \end{aligned} \quad (1)$$

where, N is the number of frames contained in a complete dynamic gesture. It is worth noting that the generation of t-RDM is obtained in one dynamic gesture sample. Therefore, the motion signatures among different gesture samples would not interfere with each other. Additionally, the generation of t-RDM does not change the number of frames contained in a gesture sample, that is, a complete gesture sample still corresponds to N frames of t-RDMs. Because the previous frames are observed and can be regarded as the prior knowledge, the operation of non-coherent accumulating on the previous frames does not increase the sample observation time, but it records a trajectory about range and velocity in t-RDM and constructs solid inter-frame dependencies. The comparison of the RDMs with different salient features of a push gesture is shown in Figure 2.

In the subfigures of Figure 2, the range-Doppler bins have different brightness, and the brightness value represents the energy intensity. The energy intensity is determined by the radar equation and is related to the target's range and radar cross section (RCS). In RDMs, the blue bins have the lowest energy and the yellow bins have the highest energy. It can be seen from Figure 2a that the bins with high energy are concentrated in low-speed and low-distance areas. In Figure 2b, the curve with lower energy is indicated as a trajectory of the push gesture according to the hand's movements, which starts and ends with a velocity of zero and moves a certain distance approaching radar. However, the bins with the highest energy in the raw t-RDM are still concentrated in the low-speed and low-distance areas. In other words, the micro signatures of gestures are prominent in the raw t-RDM.

In the coarse-grained gestures with fast speed or large motion amplitude, the brightest bins in the t-RDMs are predominantly from the human joints or some static objects. These interfering signals are enhanced in the low-speed and low-distance areas by frame-by-frame incoherent accumulation due to the tiny motion characteristics. Intuitively, the movement trajectory is conducive to characterising the coarse-grained gestures, which usually have evident movement trajectories. Therefore, an enhanced t-RDM is proposed to highlight the movement trajectory by subtracting non-zero area from the matrix composed of the maximum value of the raw t-RDM, and it can be expressed as follows:

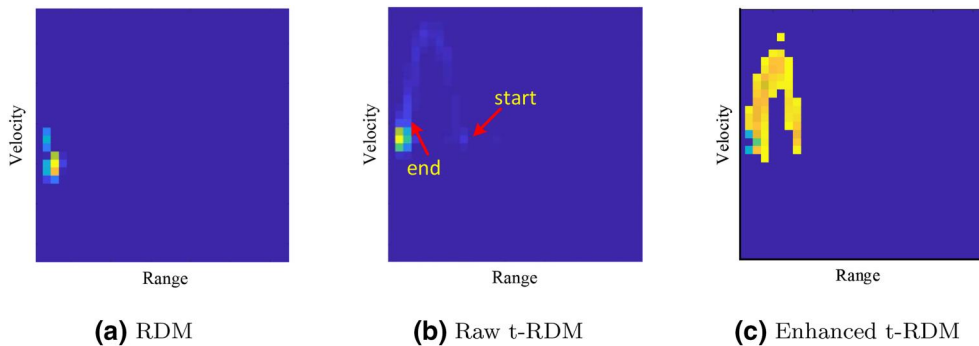


FIGURE 2 Comparison of the range Doppler map (RDM), raw trajectory-RDM and the corresponding enhanced t-RDM

$$\begin{aligned}
\text{enhanced } t\text{-RDM} &= A_{\text{mask}} \odot (B_{\text{max}} - t\text{-RDM}) \\
A_{\text{mask}} &= \text{zeros}(m, n) \oplus t\text{-RDM} \\
B_{\text{max}} &= \max(\max(t\text{-RDM})) \cdot \text{Ones}(m, n)
\end{aligned} \quad (2)$$

where, \odot is Hadamard product, \oplus is exclusive or sign, A_{mask} is a mask matrix used to extract non-zero area in t-RDM, and B_{max} is a matrix composed of the maximum value of the t-RDM, and it has the same dimension of $m \times n$ as t-RDM. The corresponding enhanced t-RDM is obtained by Equation (2), and it is given in Figure 2c. It can be seen from Figure 2c that the characteristics of the movement trajectory are highlighted in the enhanced t-RDM.

However, not all kinds of gestures have noticeable movement trajectories. Fine-grained gestures with low-speed and small-amplitude movements are widely used in close HCI, for example, the interaction between people and smartphones etc. Limited by the movement characteristics of the hand and the resolution capabilities of the radar, the trajectories in the fine-grained gestures are not evident and are usually concentrated in the low-speed and low-distance areas. Moreover, the motion signatures of fine-grained gestures are easily coupled with other micro-motion interfering signals. It is challenging for the enhanced t-RDM to accurately characterise the movement trajectories of fine-grained gestures, but the raw t-RDM has advantages in highlighting the micro-motion signatures of gestures as mentioned above.

To fully exploit the valid features of the two t-RDMs, a two-pathway CNN targeted for raw and enhanced t-RDMs is proposed, and the illustration of the network is presented in Figure 3. Hierarchical convolutional layers have powerful feature extraction capability, and a module composed of convolutional layers in each pathway can be regarded as a feature extractor [33]. The structure of each feature extractor consists of two 2-D convolution layers, each of which is followed by a max-pooling layer. First, 32 convolution kernels with a size of 5

$\times 5$ are applied in the first convolution layer. Then, 64 convolution kernels with a size of 3×3 are adopted in the second convolution layer. The same padding is utilised in each convolution layer as shown in the feature extractor module in Figure 3.

Raw t-RDMs and the corresponding enhanced t-RDMs have the same dimension, and they are, respectively, fed into each pathway feature extractor of the model. The output of each feature extractor has 64 independent feature maps, and they are flattened into a vector, respectively. To ensure that the features from the raw t-RDMs and enhanced t-RDMs are not mutually destroyed, the two flattened feature vectors are straightly concatenated into one feature vector. Subsequently, the concatenated feature vector is fed into the first fully connected layer with 128 neurons to obtain discriminative features. Finally, the last fully connected layer with n classes of neurons is adopted as a classification layer and outputs the predicted probabilities. In this study, 11 neurons are employed in the last fully connected layer.

3.2 | AIC loss

The gesture quality varies from its source to source. To extract the discriminate features in a targeted manner, AIC loss is proposed in this part. It allows the two-pathway CNN model to adaptively capture the commonalities in irregular personalised samples according to the contributions.

The premise assumption of AIC loss is that although the individual performing characteristics make the intra-class gestures divergent, some similarities still exist among these gestures. It is the similarities within a class that provide the possibility for correct identification. Therefore, the core idea of AIC loss is to minimise the intra-class variance and make the same type of gestures more clustered no matter from which

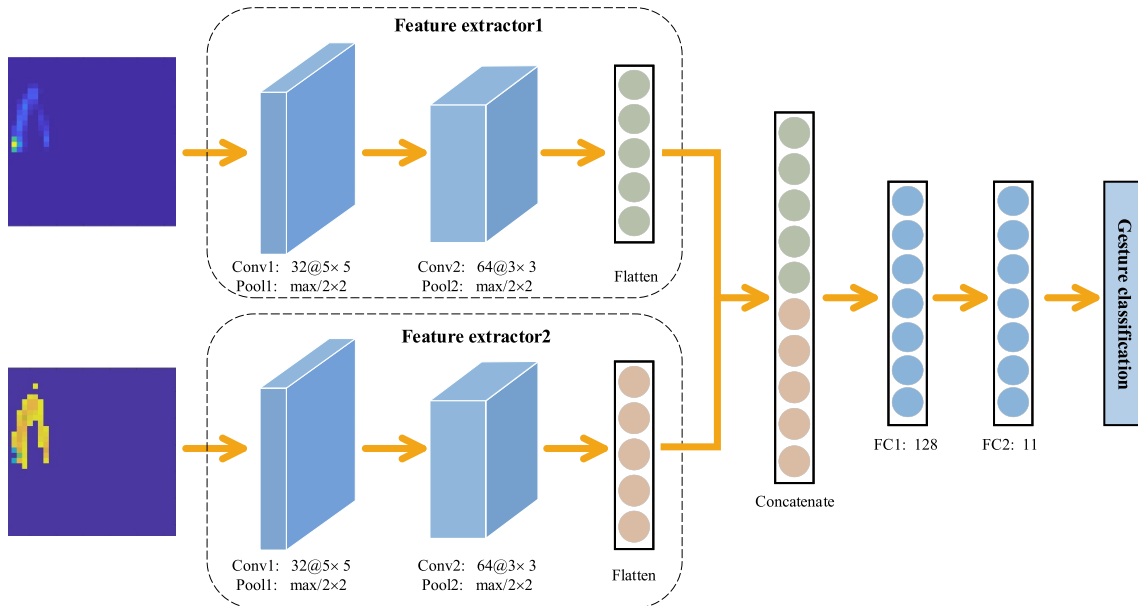


FIGURE 3 Illustration of the two-pathway CNN targeted for raw and enhanced t-RDMs

sources. Specifically, a novel data structure is first constructed, which prepares for the calculation of AIC loss. In this novel dataset, a gesture frame corresponds to two labels, which are the category label y_c and the source label y_s . The category label indicates the gesture category and the source label represents the gesture source. Noted that the source label is specific to the person ID in cross-person gesture recognition and the scenario ID in cross-scenario gesture recognition. The source labels of the gestures can be recorded in advance of the experiment without much extra effort.

Then, to accomplish a reasonable feature representation, an intuitive idea is to make a small intra-class variance and a large inter-class variance. Whereas, the various sources with individual characteristics or environment interferences create a gap in the intra-class data distribution as mentioned above. To this end, a loss of \mathcal{L}_{inter} is proposed and it restricts the intra-class features from inter-sources to be more clustered by minimising the Euclidean distances of the features. However, the contributions of each source to the common signatures are different. It is rigid to assign the same criteria to the gestures from varied sources. Therefore, adaptive weights are assigned in AIC loss for the different source gestures. Ideally, the loss of \mathcal{L}_{inter} should be applied to the entire training set. But the resulting huge burden of storage space and computational resources makes this scheme impractical. To address this issue, the clustered intra-class features constraining in a batch becomes an alternative choice. The loss functions of \mathcal{L}_{inter} can be expressed as follows:

$$\mathcal{L}_{inter} = \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^m w_{c_i, s_j} \left\| (1 - \delta(s_j = s_0)) (\bar{f}_{c_i, s_j} - \bar{f}_{c_i, s_0}) \right\|_2^2 \quad (3)$$

where, \bar{f}_{c_i, s_j} is the average feature vector of the c_i -th class from the s_j -th source label. Similarly, \bar{f}_{c_i, s_0} is the average feature vector of the c_i -th class from the initial source label of s_0 . To be fair, this initial source label s_0 is randomly reselected in each batch. The function of $\delta(\cdot)$ is the Dirac delta function, and the value of $1 - \delta(s_j = s_0)$ is 0 when s_j is equal to s_0 , otherwise it is 1. The item of n is the number of gesture categories in a batch, and m is the number of sources in a batch. The parameter of w_{c_i, s_j} is the adaptive weight corresponding to \bar{f}_{c_i, s_j} .

As stated in Ref. [34], cross-entropy loss contributes to making inter-class samples dispersion and provides a basis for category identification. In AIC loss, a cross-entropy (CE) loss is applied to the joint supervised for a large inter-class variance. Its formula is shown as Equation (4), where $p(x)$ is the expected output of the category labels, and $q(x)$ is the actual output of the category labels. Finally, AIC loss can be expressed as Equation (5). The detailed algorithm process of the adaptive gesture recognition framework is presented in Algorithm 1. Specifically, an adaptive weight vector W contains the weight elements of all categories from all sources in the training set. First, the corresponding weight elements are selected in each batch through a mask operation of A_{mask} . Then, the partial weights are updated with the training in each

iteration. Finally, the updating of all elements in the weight vector is accomplished through continuous iterations.

$$\mathcal{L}_{CE} = - \sum_x p(x) \log q(x) \quad (4)$$

$$\mathcal{L} = \mathcal{L}_{CE} + \mathcal{L}_{inter} \quad (5)$$

Algorithm 1 Iterative Optimisation for the adaptive gesture recognition framework

Input: $\{X_{raw}, X_{enhanced}, Y_c, Y_s\}$: training set;
 θ_c : initial parameters of the adaptive gesture recognition framework; n : the number of the gesture categories; m : the number of the data sources in the training set; n_b : the sample number of a batch; k : the total number of the training samples; max_{ep} : the max epoch index; l_r : learning rate;

Output: The parameters of θ_c .

```

1: initial  $W \leftarrow \text{ones}(mn, 1)$ ;
2: while not converge do
3:   for  $i = 0; i < max_{ep}; i++$  do
4:     for  $t = 0; t < \lceil \frac{k}{n_b} \rceil; t++$  do
5:        $A_{mask}^t \leftarrow \text{onehot}(mY_c^t + Y_s^t)$ 
6:        $W^t \leftarrow A_{mask}^t \times \text{sigmoid}(W)$ 
7:        $Out^t, F^t \leftarrow \text{forward-pass}(X^t, X_{enhanced}^t, \theta_c)$ 
8:        $s_0 \leftarrow \text{select randomly from } Y_s^t \text{ in each gesture category}$ 
9:       According to Equation (3) and Equation (4), compute the joint loss
        $\mathcal{L}^t \leftarrow \text{pass}(Out^t, F^t, Y_c^t, Y_s^t, W^t, s_0)$ 
10:       $Grad \leftarrow \text{backward-pass}(\theta_c, W^t, \mathcal{L}^t)$ 
11:       $\theta_c^{t+1}, W^{t+1} \leftarrow \text{Adam}(Grad, l_r, \theta_c, W^t)$ 
12:       $\theta_c^t, W^t \leftarrow \theta_c^{t+1}, W^{t+1}$ 
13:     end for
14:   end for
15: end while

```

4 | EXPERIMENTS

4.1 | Database

A public gesture dataset 1 collected by the Soli radar is adopted in this study for evaluation [16]. Soli is a millimetre-wave radar chip, and it is designed and developed by the Google company, especially for wearable and portable electronic equipment. As one of the pioneers, the great success brought by the Soli radar chip has promoted the development of high-resolution radar in the HGR field [15]. This radar chip has a centre frequency of 60 GHz and 2 transmitting and four receiving antennas. With

the help of extremely fast frame rates, Soli has a high temporal resolution to identify the complex fine-grained motions.

The public dataset is favoured by some researchers due to its reasonable experimental configurations [16, 18]. It includes 11 kinds of gestures and covers most of the common gestures oriented to tiny electronic devices, such as pinch index, finger rub, and so on. The specific illustrations of the gestures are presented in Figure 4, and it shows the corresponding gesture categories and snapshots. To be more suitable for actual application scenarios, the gesture acquisitions are conducted, respectively, in two experimental conditions. First, 10 volunteers are asked to perform 11 kinds of gestures under the same experimental scenario, and a total of 2750 gesture samples are collected. This gesture subset is utilised to verify cross-person gesture recognition in subsequent experiments. Second, a single volunteer is asked to perform 11 gestures in six different experimental scenarios. A total of 2750 gesture samples are collected, and this gesture subset is employed in subsequent cross-scenario gesture recognition.

To illustrate the diversity of the gesture samples, the t-RDMs from the cross-person subset and the cross-scenario subset are given in Figure 5, which includes both fine-grained gestures and coarse-grained gestures. It can be seen from Figure 5 that there are evident differences among the gesture samples from different data sources. In addition, it is difficult to observe the movement trajectory from fine-grained gesture samples, but it is prominent in coarse-grained gesture samples.

4.2 | Implementation details

To facilitate the learning process of the network, the gestures are first resized into a fixed-length by downsampling or inserting. Considering the length of most gesture samples, 50 frames are selected as the uniform length of the resized gestures. The four closely spaced receiving antennas lead to similar RDMs in each channel. To reduce the computational burden and improve the signal-to-noise ratio, similar RDMs from

different channels are averaged into a single frame. First, frame-by-frame non-coherent accumulation is applied to a complete gesture sample, and a sequence of t-RDMs is obtained. Each frame of t-RDM records the trajectory of its previous frames. Then, the enhanced t-RDM is acquired by subtracting all valid features from the matrix composed of the maximum energy value in the raw t-RDM. Therefore, the original trajectory information with lower energy is highlighted in the enhanced t-RDM. Finally, the raw t-RDMs and enhanced t-RDMs are fed into the two-pathway CNN, respectively.

Parametric rectified linear unit (PReLU) is used as an activation function [35], and a dropout of 0.8 is applied after the penultimate fully connected layer. Adam is chosen as the method of stochastic optimisation, and the learning rate is 3×10^{-4} . The size of a mini-batch is set as 512, and 10 epochs are selected empirically as the maximum number of epochs. The output of the penultimate fully connected layer is used as the feature vector, and it participates in the calculation of AIC loss together with the output of the last fully connected layer.

4.3 | Evaluation results

To explore the effectiveness of the proposed method in diversity gesture identification, it is assessed on the novel evaluation approaches, respectively, in the cross-person and cross-scenario gesture recognition.

For a reliable result, the corresponding subset from the Soli dataset is chosen in each group of evaluation experiments. Specifically, a subset of the gestures from 10 performers in the same scenario is adopted in cross-person recognition. As the experimental configuration in Ref. [16], a leave-one-person-out mode is used to verify the impact of unknown people on recognition performance in turn. There are 10 volunteers included in the subset, therefore, a total of 10 validation experiments are conducted for convincing results. In each validation experiment, the gesture samples from one of the volunteers are chosen to be left for testing, and the remaining

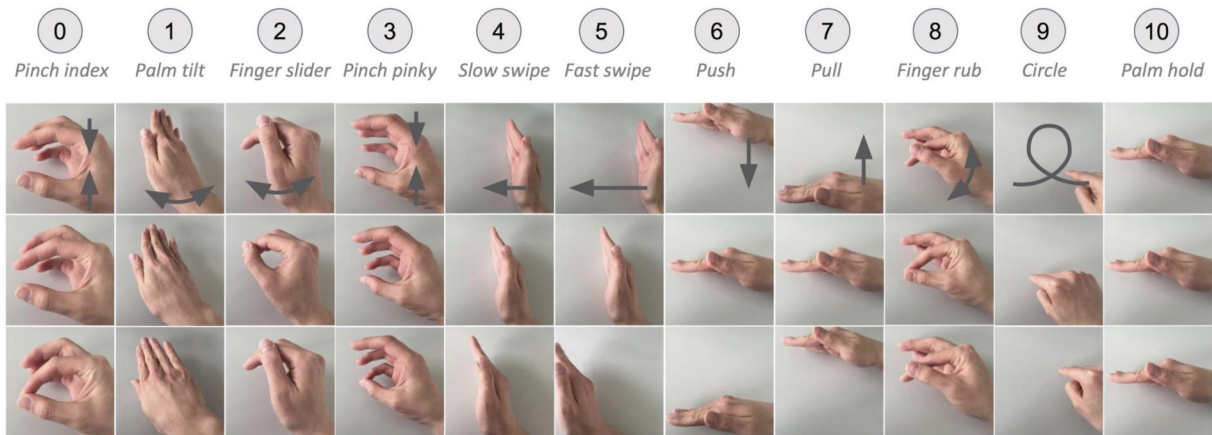


FIGURE 4 Illustration of 11 gestures in the soli dataset

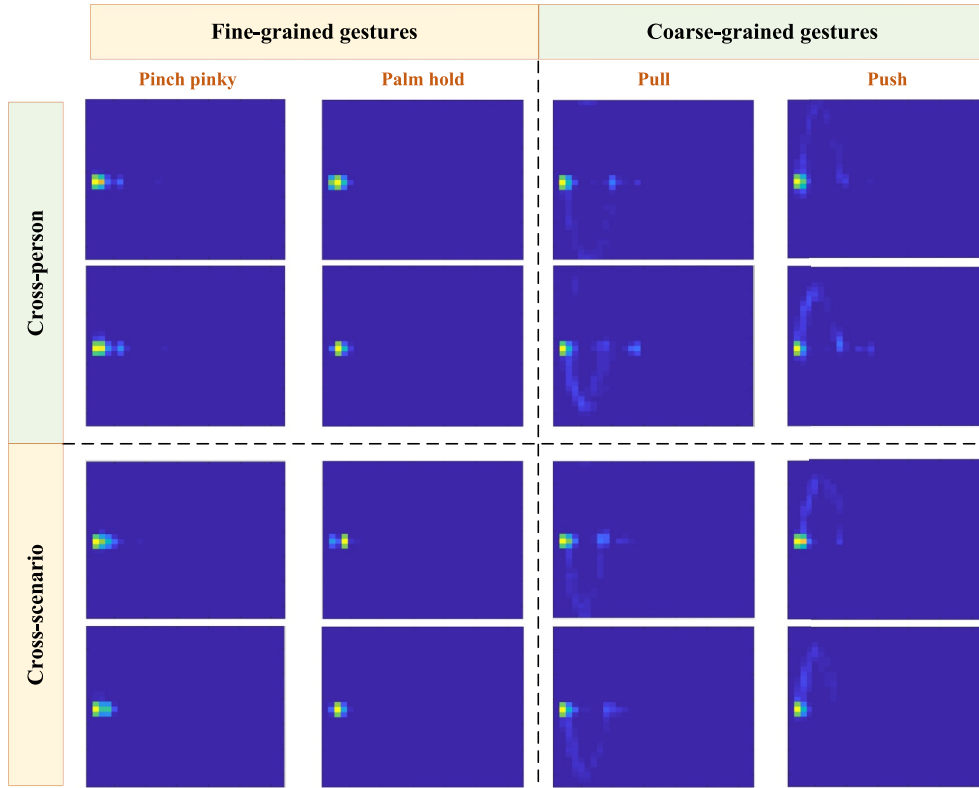


FIGURE 5 Illustration of the diversity of the raw trajectory range Doppler maps

volunteers' samples are used for training. Similarly, a gesture subset collected in multiple scenarios is employed in cross-scenario recognition. A leave-one-scenario-out mode is applied to examine the classifier's generalisation performance on the unknown scenario. Finally, a total of 6 validation experiments are conducted in the cross-scenario recognition, and the gestures from each scenario are chosen as the test set in turn. The average results of the cross-person recognition and cross-scenario recognition are shown in Table 1.

Table 1 shows that the average accuracy of the adaptive gesture recognition framework in cross-person recognition is 81.45%, and the accuracy in cross-scenario recognition is 92.02%. To facilitate a clear view of the classification rate among the gestures, the confusion matrices of two recognition results are shown in Figure 6. It can be seen from Figure 6 that the proposed method has a prominent distinguishing ability on most gestures in both evaluation modes. In the cross-person mode, the accuracy rate of less than 80% is mainly in the three gestures of G0, G2, and G3. The gestures with tiny movements pose a more challenging issue in cross-person gesture recognition. Some joint occlusion may exist during the hand movement, and the energy of their range-Doppler bins shows a more evident gap across different people. In cross-scenario recognition, the accuracies of the majority of hand gestures have been improved significantly, including the tiny gestures of G0, G2, and G3. Such improvements are mainly caused by two aspects, one of which is that the performing differences for the same person become more narrowed, and another one

TABLE 1 The average accuracies of cross-person gesture recognition and cross-scenario gesture recognition

	Cross-person	Cross-scenario
Accuracy	81.45%	92.02%

is that there is less interference of the specific environment at a long distance compared with the gesture operation at a range close to the radar.

In the proposed method, the adaptive gesture recognition framework is constrained by AIC loss. To obtain more generalised features, individual cost weights are adopted on the features from different sources. The cost weights are learnt and optimised by the network according to the recognition task, and the final weights of the cross-person recognition and cross-scenario recognition are shown in Figure 7. The colour of the scattering points represents the gesture category, and each scattering point corresponds to the weight of a source. The value at the x-coordinate is the cost weight value, and the average weight value of each category is shown as a curve marked with circles.

Theoretically, the distribution of individual cost weights should follow the two criteria: (1) Just as no two leaves are exactly the same in the world, the weights of different sources tend to be different. (2) For the same gesture category, the sources with more deviating samples should be given heavier cost weights to extract the common characteristics. The experimental results show that the cost weights of various sources are indeed different as shown in Figure 7. By

comparing the scatter plots of Figure 7a,b, it can be noticed that the cost weight distributions show a similar trend among the gesture categories, even though they are from different source types.

On the whole, the cost weights of gestures G1, G5, G6, and G7 are relatively high, and these gestures have a larger RCS and an obvious movement action. Interestingly, the gestures of G4 and G5 are similar, and they are both swipe actions but have different speeds. However, the average cost weight of G4 is relatively lower but that of G5 is relatively higher. Additionally, the gesture G10 seems to have a comparable RCS with G6 and G7, but the cost weights of G10 are far below the weights of G6 and G7, which is a palm hold action with only a slight hand shivering. On the other hand, the gestures with small RCS are given relatively lower cost weights, such as G2 and G3. The above results demonstrate that the penalty factors are related to the targets' RCS and the movement characteristics. The gestures with larger RCS and evident movement action are more likely to be assigned heavier cost weights adaptively by the network.

5 | ANALYSIS

5.1 | Effect of the two-pathway CNN targeted for raw and enhanced t-RDMs

In this part, the contributions of the two-pathway CNN targeted for raw and enhanced t-RDMs are verified. Compared with the existing radar-based gesture recognition methods, the proposed module has two main contributions: (1) the raw RDMs and the enhanced t-RDMs, which establish a strong inter-frame dependency and form a movement trajectory; (2) the feature-oriented two-pathway CNN network, which mines the discriminative features from two pathway inputs independently. In the subsequent experiments, the two contributions are proven one by one.

First, a baseline model without any proposed modules is designed, and it has two convolution layers; each layer is followed by a max-pooling. In the first convolution layer, 32 filters with the size of 5×5 , and 64 filters with the size of 3×3 are applied in the second convolution layer. Subsequently, two

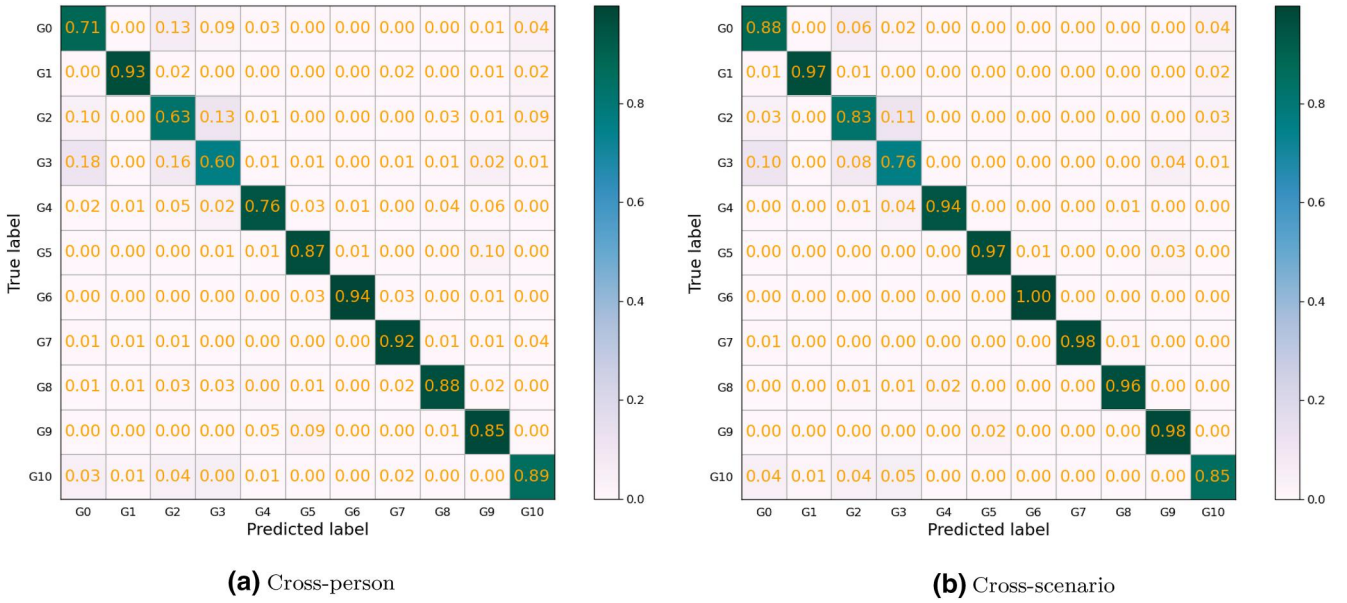


FIGURE 6 Confusion matrices of the cross-person gesture recognition and cross-scenario gesture recognition

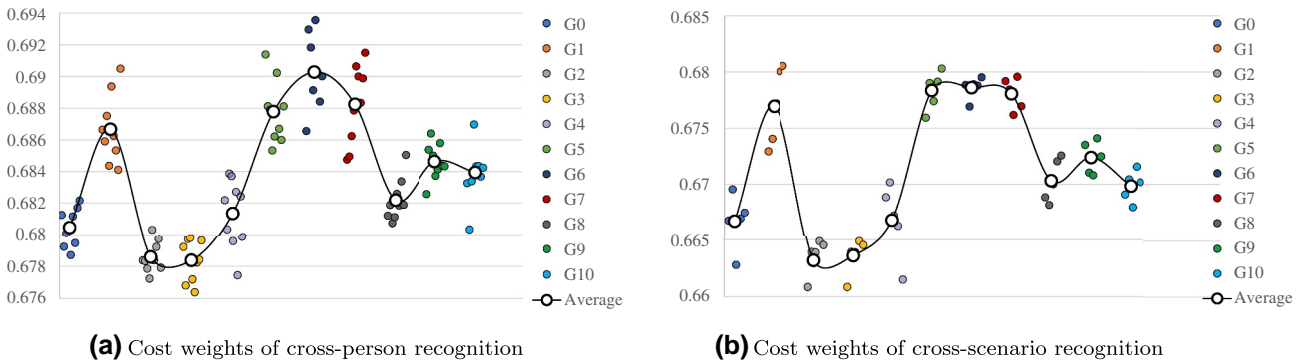


FIGURE 7 The individual cost weights of the cross-person gesture recognition and cross-scenario gesture recognition

fully connected layers are followed with 128 and 11 neurons, respectively. In brief, the baseline has the same structural configuration as the single path model in the two-pathway CNN. However, the input gesture samples are replaced with only the original RDMs without any processing, and the classic cross-entropy loss (CE) is adopted as the loss function. The specific structure of the baseline model is given in Figure 8.

Second, to evaluate the effect of data enhancement, the single-pathway CNN incorporating the raw and enhanced t-RDMs is designed in the baseline + concatenation experiment. In this configuration, the raw t-RDM and the corresponding enhanced t-RDM are concatenated into a new input sample, and then the new inputs are fed into the baseline model. The dimensions of raw t-RDM and enhanced t-RDM are 32×32 ; therefore, the concatenated RDM's dimension becomes 64×32 . This experimental configuration maintains the same number of features as the proposed two-pathway CNN model. Finally, to testify the superiority of the two-pathway model structure, an experiment with the two-

pathway CNN structure is conducted, but CE loss replaces the AIC loss as the model's loss function.

For convincing results, this set of experiments is evaluated in cross-person recognition and cross-scenario recognition, respectively. The specific recognition rates under different methods are compared through the histograms in Figures 9 and 10, which correspond to cross-person recognition and cross-scenario recognition, respectively. The results of the proposed overall framework, that is the adaptive gesture recognition framework, are compared in two figures together for comparing conveniently.

In cross-person recognition, the average accuracies of baseline, baseline + concatenation, and two-pathway CNN + CE are 54.9%, 78.1%, and 78.5%, respectively. These experiment results in cross-session recognition are 67.4%, 88.9%, and 89.4%, respectively. It is well accepted that the CNN model has a more powerful ability in capturing the local signatures of the inputs rather than the sequential dependencies. A signal frame of a complete gesture as the input

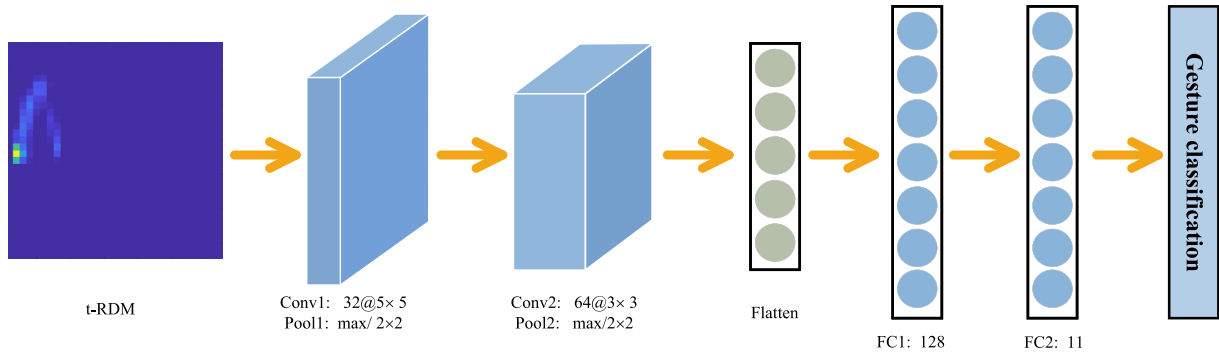


FIGURE 8 Illustration of the baseline model

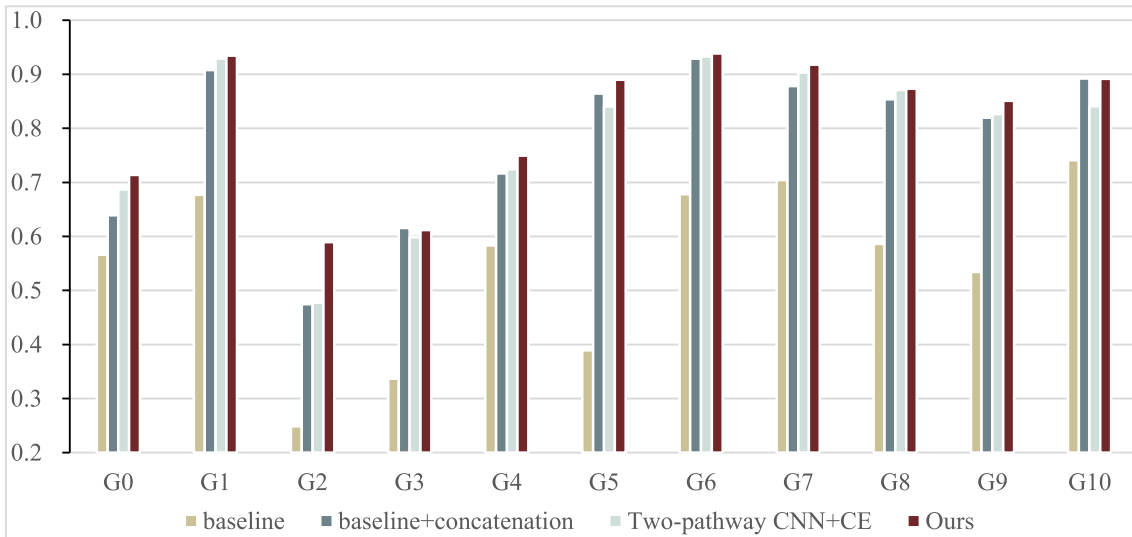


FIGURE 9 A set of experiments to evaluate the effectiveness of two-pathway convolutional neural network targeted for raw and enhanced trajectory range Doppler map in cross-person gesture recognition

exacerbates the difficulty of diversity gesture recognition. Consistent with the theoretical expectations, the performance of baseline shows the worst recognition rate in identifying the diverse gestures for both cross-source recognition. Whereas, the recognition accuracy is improved by more than 20% in both recognition modes when the original data are enhanced to establish a strong inter-frame dependency. As shown in Figures 9 and 10, the gesture accuracy improves significantly in baseline + concatenation compared with that of in baseline. The results demonstrated that the data enhancement method is beneficial for CNN to mine the discriminative features.

The last group of experimental results demonstrated that the two-pathway CNN model is superior to the signal CNN model by comparing the average recognition accuracies of baseline + concatenation and two-pathway CNN + CE. Although the overall accuracy improvement is slight, the recognition results of each type of gesture illustrate that the two-pathway CNN is beneficial to improving the recognition performance for most of the gestures as shown in Figures 9 and 10. More importantly, the two-pathway CNN extracts the features from both input data independently [36], which is convenient to observe the extraction of coarse-grained and fine-grained motion signatures of gestures and enhances the interpretability of the classification model. To illustrate the two-pathway CNN's ability to extract full-grained features, the feature maps of its first two convolutional layers for coarse-grained and fine-grained gestures are, respectively, given in Figure 11.

First, the feature maps of the coarse-grained gestures are presented in Figure 11a. It can be seen clearly from Figure 11a that the classifier tends to focus on the bins with higher energy in the input sample. Therefore, the one pathway feature extractor based on t-RDM tends to focus on the micro-movement signatures of gestures, and the other pathway feature extractor based on enhanced t-RDM tends to focus on

the movement trajectory of gestures. Second, the feature maps of the fine-grained gesture are presented in Figure 11b. Due to the tiny movements, it is difficult to discover the above rules based on the feature maps of the fine-grained gestures. The complex motion and occlusion among fingers make it difficult for enhanced t-RDM to accurately characterise the motion trajectory of gestures, as shown in Figure 11b. As stated in reference [16], the fine-grained gestures pose a more challenging task for recognition algorithms. However, the experimental results illustrated that the two-pathway input samples with different salient features are beneficial to comprehensively describe the motion signatures of the fine-grained gestures by comparing the results of baseline and two-pathway CNN + CE.

5.2 | Effect of the AIC loss for personalised gestures

In this part, the contribution of the proposed AIC loss to the diversity gesture recognition is demonstrated. Similar to the above experimental configuration, the single pathway CNN with CE loss is still applied as the baseline model in this part. Whereas, the CE loss of the baseline model is replaced by the AIC loss in the comparison experiment. As usual, the validation trials of AIC loss are conducted in cross-person and cross-scenario recognition, respectively.

In cross-person recognition, the recognition rates of baseline and baseline + AIC loss are 54.9% and 57.2%, respectively. In addition, their average accuracy rates are separately 67.4% and 71.3% in cross-scenario recognition. As shown in Figures 12 and 13, the accuracies of each gesture type are compared in different recognition algorithms. The experimental results show that the algorithm combined with AIC loss contributes to the diversity gesture recognition. It indicates

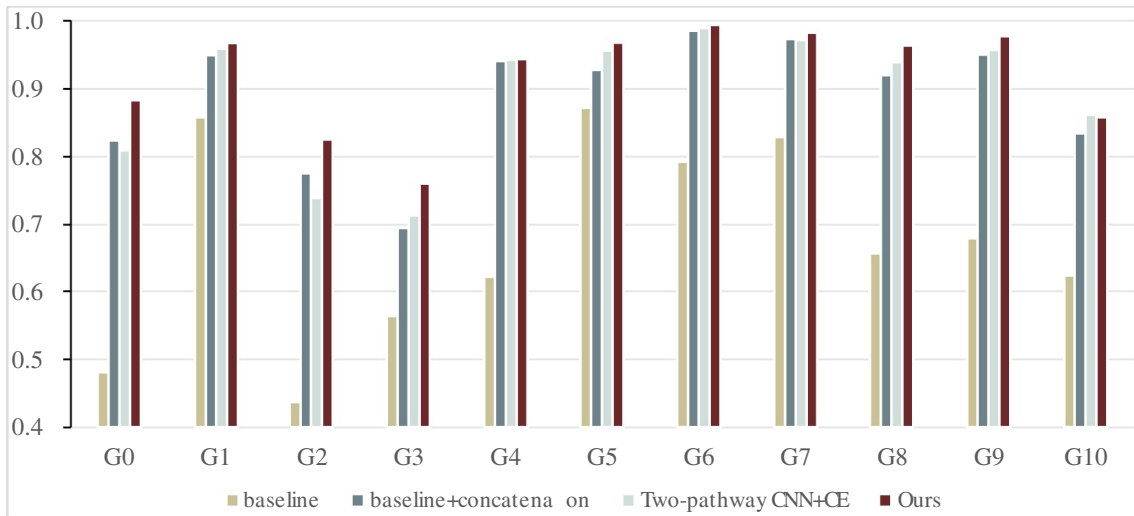
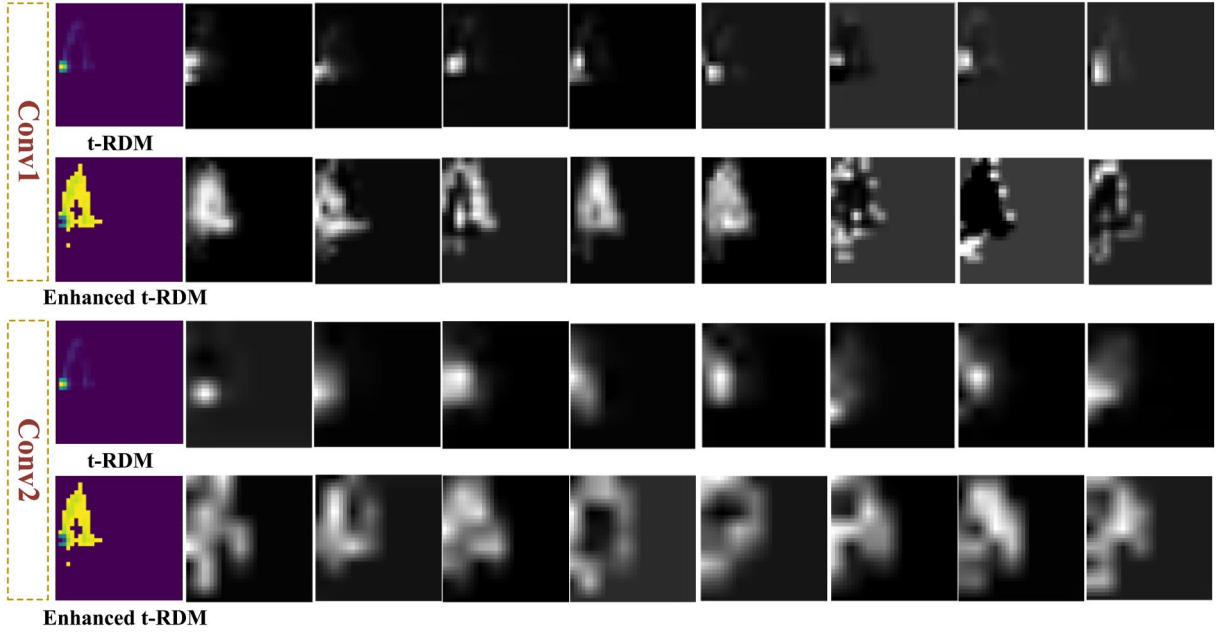
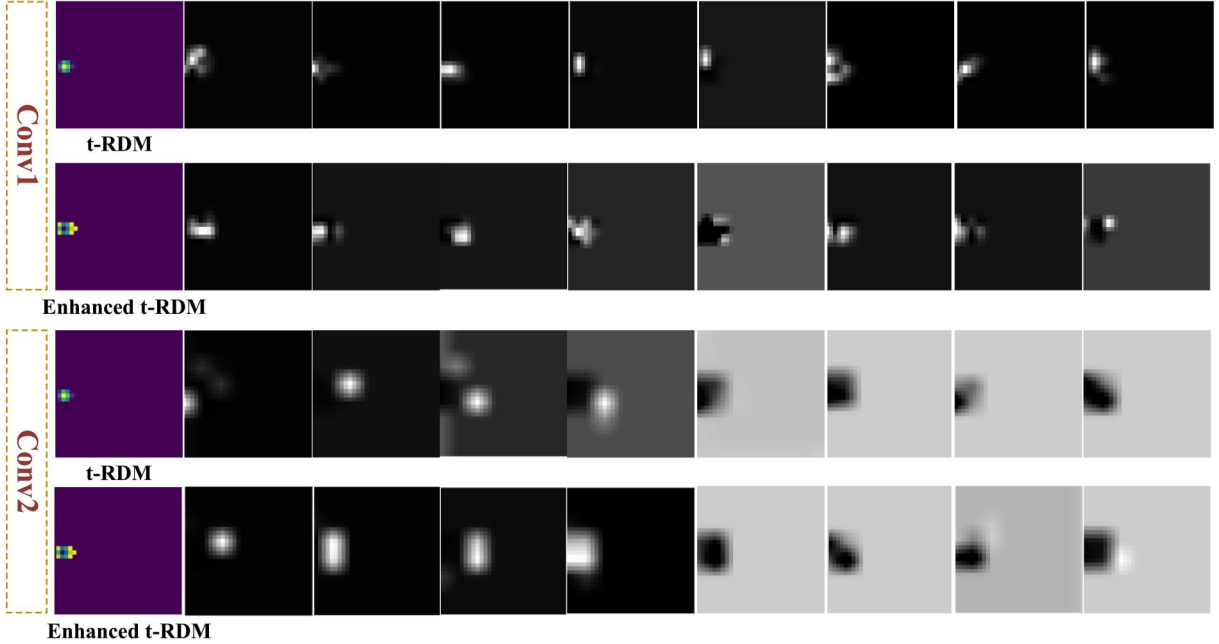


FIGURE 10 A set of experiments to evaluate the effectiveness of two-pathway convolutional neural network targeted for raw and enhanced trajectory range Doppler map in cross-scenario gesture recognition



(a) The feature maps of coarse-grained gestures



(b) The feature maps of fine-grained gestures

FIGURE 11 Comparison of feature maps of the coarse-grained gestures and the fine-grained gestures

that AIC loss is effective to extract the robust features in the diverse gestures, even in the original samples without any enhancements. However, the improvement of AIC loss is less significant than that of the data enhancement method. The network is driven by data, and it is reasonable to infer that the data quality has a primary and critical impact on the network performance. Therefore, AIC loss can alleviate the challenges in diversity gesture recognition but enhance the sample quality that has an outstanding improvement as shown in the above analysis and experiments.

5.3 | Compared with existing state-of-the-art methods

To the best of our knowledge, there are only a few studies to explore the classifier's performance on unknown gesture sources, among which Ref. [16, 18] are the most representative ones. Both studies investigated this issue based on the publicly available Soli dataset, which is consistent with the dataset we used, and therefore the comparison is reasonable. In this part, the experiment configuration is consistent with that in Ref. [16,

[18], and the average results of 10 leave-one-person-out experiments and 6 leave-one-scenario-out experiments are summarised in Table 2.

The results of the proposed method are, respectively, 81.45% and 92.02% in cross-person and cross-scenario gesture recognition. The proposed method shows a more outstanding recognition performance than the other state-of-the-art methods, especially for some gestures such as G0 and G5 etc. By comparing the accuracies of each type of gestures, it can be found that the results of G0, G2, and G3 are relatively lower than the other gestures in any of the identification approaches. These gestures have tiny movement actions, and fingers in

motion may cover each other in the line of sight of radar. Therefore, it is difficult for radar to perceive these fine-grained gestures with high precision. Moreover, similar shapes in the gestures of G0, G2, and G3 with only small differences in movement lead to the signatures of these gestures being easier to be confused. As reported in Ref. [18], it is still a challenge for a classifier to correctly identify the type of fine-grained gestures based on radar.

Another phenomenon is that the cross-scenario recognition shows consistently superior performance than the cross-person recognition among these approaches mentioned above. Theoretically, it may be due to the fact that the

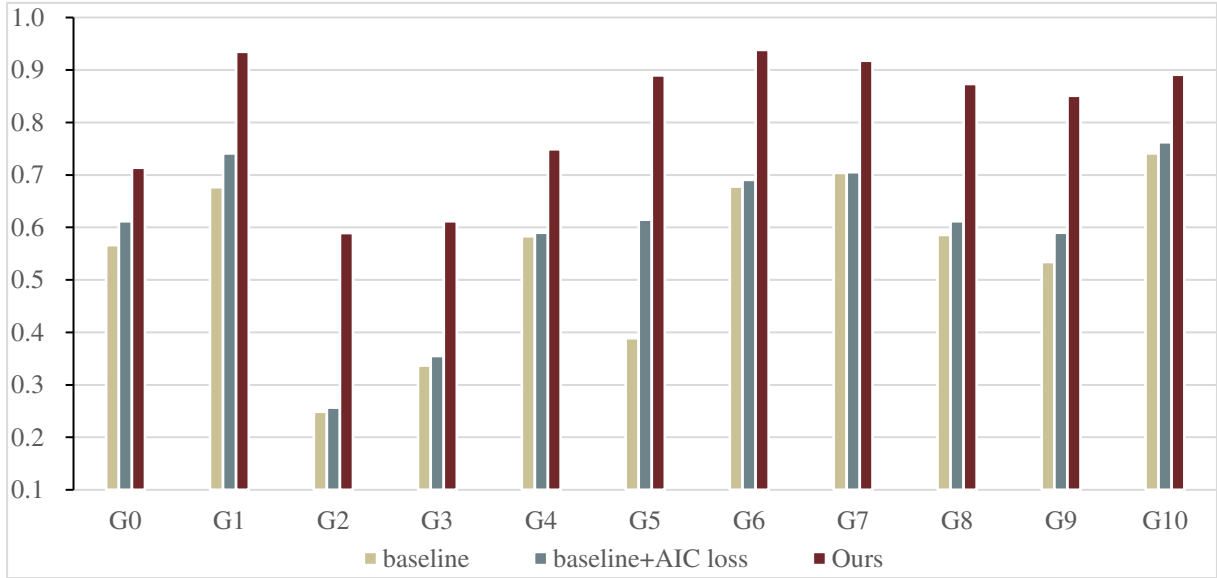


FIGURE 12 A set of experiments to evaluate the effectiveness of adaptive individual cost loss for cross-person gesture recognition

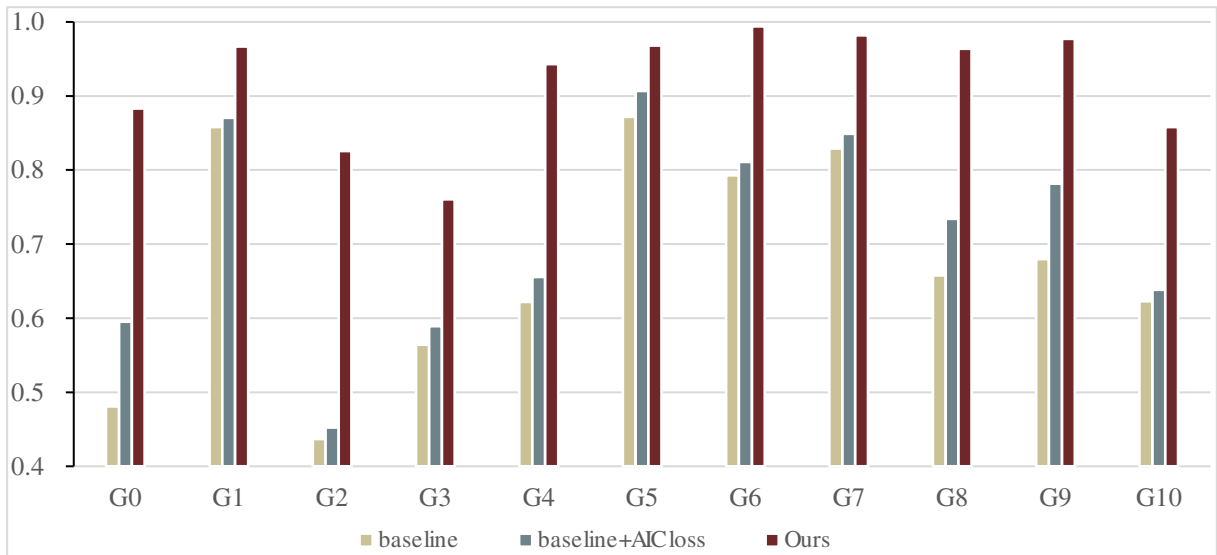


FIGURE 13 A set of experiments to evaluate the effectiveness of adaptive individual cost loss for cross-scenario gesture recognition

distribution gap of the cross-scenario dataset is smaller than that of the cross-person dataset. To verify this hypothesis, the original data is directly applied for visualisation by t-SNE with no further extraction by any recognition algorithms. The method of t-SNE is a powerful technique for visualising the data distribution, and it does not need excessive tuning parameters [37]. The data distribution visualisation results from cross persons and cross scenarios are shown in Figure 14.

Both subgraphs of Figure 14 are uniformly constrained within the same x-axis and y-axis ranges, and the annotation shows the radius of the clustered region. It can be seen from Figure 14 that the original data distribution from cross persons is indeed more dispersed than that from cross scenarios as the theoretical analysis. There are two main reasons for this difference in the distribution gap between the cross-person and cross-scenario datasets. First, the gestures from a single person generally have a more consistent performing habit than those from multiple people. Second, different people are performing gestures close to the radar in the cross-person dataset, while in the cross-scenario dataset, radar signals record the signatures from multiple scenes at a faraway distance from the radar. According to the radar equation, the scenario-specific signals at a farther distance have less interference for a radar chip with

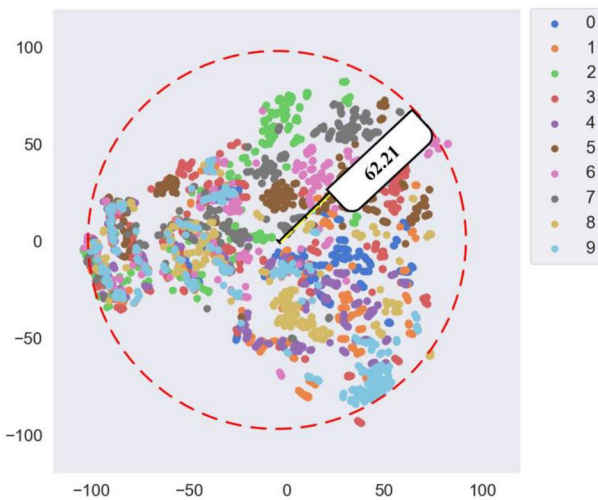
the narrowed space. The results of this visualisation experiment further demonstrate that the data quality has a significant impact on the algorithm performance.

6 | CONCLUSION

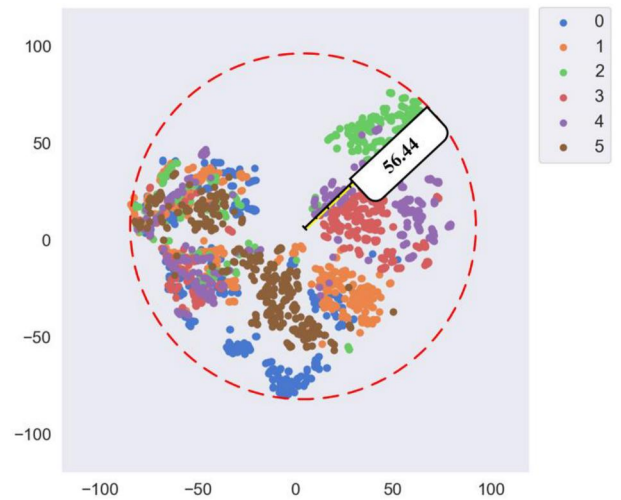
The hand's flexibility and the signals from the specific environment result in the diversity in gesture samples. In practical applications, a good HGR system is required to make a reliable response, no matter what person or scenario it comes from. In this paper, we presented an adaptive gesture recognition framework for the gestures with individual variations. It is mainly based on (1) a two-pathway CNN targeted for raw and enhanced t-RDM. The data enhancement first establishes strong inter-frame dependencies, and different movement characteristics are highlighted for the physical states of the gesture. Then, a two-pathway CNN is proposed to independently extract the discriminative features from the two-pathway inputs. (2) AIC loss is proposed to minimise the intra-class variance by constraining the features from different sources to be more clustered. More importantly, it adaptively adjusts the penalty weights for the common signatures in diverse

TABLE 2 The average results of cross-person recognition and cross-scenario recognition using different methods

	Methods	Avg.Acc. (%)	G0	G1	G2	G3	G4	G5	G6	G7	G8	G9	G10
Cross-person	UFR-CNN-LSTM [18]	78	65	93.37	62.5	62.78	76.95	60.18	95.98	90.46	90.51	85.9	74.41
	CNN-LSTM [16]	79.06	58.71	95.16	64.8	67.62	72.31	72.91	93.4	89.99	91.82	82.8	80.24
	Ours	81.45	71.35	93.46	58.94	61.18	74.92	89.01	93.83	91.76	87.34	85.08	89.13
Cross-scenario	UFR-CNN-LSTM [18]	84.01	59.58	96.63	73.51	63.52	86.74	84.55	99.06	96.12	93.66	91	79.84
	CNN-LSTM [16]	85.75	56.69	95.33	76.43	61.98	92.73	81.38	98.42	97.79	96.83	96.92	89.1
	Ours	92.02	88.31	96.69	82.56	76.06	94.33	96.81	99.40	98.20	96.36	97.71	85.80



(a) The intra-class data distribution from cross persons



(b) The intra-class data distribution from cross scenarios

FIGURE 14 Comparison of intra-class data distribution in cross-person recognition and cross-scenario recognition

samples. The gestures from different sources are assigned appropriate weights so that they can be fully exploited. Comprehensive experiments on a publicly available dataset demonstrated the effectiveness of the proposed method on cross-person and cross-scenario gesture recognition. Subsequently, the contributions of the proposed modules to the proposed method have also been proven separately. Finally, comparison results with the existing state-of-the-art methods show that the proposed method has superior performance on the diverse personalised gestures.

ACKNOWLEDGEMENTS

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 61971101.

ORCID

Liyang Wang  <https://orcid.org/0000-0002-8964-0797>

REFERENCES

- Mitra, S., Acharya, T.: Gesture recognition: a survey. *IEEE Trans. Syst. Man Cybernet Part C*. 37(3), 311–324 (2007). <https://doi.org/10.1109/TSMCC.2007.893280>
- Fan, T., et al.: Wireless hand gesture recognition based on continuous-wave Doppler radar sensors. *IEEE Trans. Microw. Theor. Tech.* 64(11), 4012–4020 (2016). <https://doi.org/10.1109/TMTT.2016.2610427>
- Rautaray, S.S., Agrawal, A.: Vision based hand gesture recognition for human computer interaction: a survey. *Artif. Intell. Rev.* 43(1), 1–54 (2015). <https://doi.org/10.1007/s10462-012-9356-9>
- Li, Y., et al.: Deep attention network for joint hand gesture localization and recognition using static RGB-D images. *Inf. Sci.* 441, 66–78 (2018). <https://doi.org/10.1016/j.ins.2018.02.024>
- Kim, K.-W., et al.: Recognition of sign language with an inertial sensor-based data glove. *Technol. Health Care*. 24(s1), S223–S230 (2016). <https://doi.org/10.3233/THC-151078>
- Galka, J., et al.: Inertial motion sensing glove for sign language gesture acquisition and recognition. *IEEE Sensor. J.* 16(16), 6310–6316 (2016). <https://doi.org/10.1109/JSEN.2016.2583542>
- Amin, M.G., Zeng, Z., Shan, T.: Hand gesture recognition based on radar micro-Doppler signature envelopes. In: 2009 IEEE Radar Conference, pp. 1–6. (2019). <https://doi.org/10.1109/RADAR.2019.8835661>
- Molchanov, P., et al.: Short-range FMCW monopulse radar for hand-gesture sensing. In: 2015 IEEE Radar Conference, pp. 1491–1496. IEEE, Arlington (2015). <https://doi.org/10.1109/RADAR.2015.7131232>
- Plouffe, G., Cretu, A.-M.: Static and dynamic hand gesture recognition in depth data using dynamic time warping. *IEEE Trans. Instrum. Meas.* 65(2), 305–316 (2016). <https://doi.org/10.1109/TIM.2015.2498560>
- Wang, L., et al.: Negative latency recognition method for fine-grained gestures based on terahertz radar. *IEEE Trans. Geosci. Rem. Sens.* 58, 7955–7968 (2020). <https://doi.org/10.1109/TGRS.2020.2985421>
- Choi, J.-W., Ryu, S.-J., Kim, J.-H.: Short-Range Radar Based Real-Time Hand Gesture Recognition Using LSTM Encoder. *IEEE Access*. 7, 1–33618 (2019). <https://doi.org/10.1109/ACCESS.2019.2903586>
- Khan, F., Leem, S.K., Cho, S.H.: Hand-based gesture recognition for vehicular applications using IR-UWB radar. *Sensors*. 17(4), 833 (2017). <https://doi.org/10.3390/s17040833>
- Gravina, R., et al.: Multi-sensor fusion in body sensor networks: state-of-the-art and research challenges. *Inf. Fusion*. 35, 68–80 (2017). <https://doi.org/10.1016/j.inffus.2016.09.005>
- Brown, L., March, R.H.: A radar history of world war II: technical and military imperatives. *Phys. Today*. 53(10), 82–84 (2000). <https://doi.org/10.1063/1.1325205>
- Lien, J., et al.: Ubiquitous gesture sensing with millimeter wave radar. *ACM Trans. Graph.* 35(4), 1–19 (2016). <https://doi.org/10.1145/2897824.2925953>
- Wang, S., et al.: Interacting with Soli: exploring fine-grained dynamic gesture recognition in the radio-frequency spectrum. In: Proceedings of the 29th Annual Symposium on User Interface Software and Technology - UIST '16, pp. 851–860. ACM Press, Tokyo (2016). <https://doi.org/10.1145/2984511.2984565>
- Zhang, Z., Tian, Z., Zhou, M.: Latern: dynamic continuous hand gesture recognition using FMCW radar sensor. *IEEE Sensor. J.* 18(8), 3278–3289 (2018). <https://doi.org/10.1109/JSEN.2018.2808688>
- Berenguer, A.D., et al.: GestureVLAD: combining unsupervised features representation and spatio-temporal aggregation for Doppler-radar gesture recognition. *IEEE Access*. 7, 137122–137135 (2019). <https://doi.org/10.1109/ACCESS.2019.2942305>
- Browne, M.W.: Cross-validation methods. *J. Math. Psychol.* 44(1), 108–132 (2000). <https://doi.org/10.1006/jmps.1999.1279>
- Zhou, Z., Cao, Z., Pi, Y.: Dynamic gesture recognition with a terahertz radar based on range profile sequences and Doppler signatures. *Sensors*. 18. <https://doi.org/10.3390/s18010010>
- Ahmed, S., et al.: Hand gestures recognition using radar sensors for human-computer-interaction: a review. *Rem. Sens.* 13(3), 527 (2021). <https://doi.org/10.3390/rs13030527>
- Gurbuz, S.Z., et al.: Micro-Doppler-based in-home aided and unaided walking recognition with multiple radar and sonar systems, *IET Radar. Sonar & Navigation*. 11(1), 107–115 (2017). <https://doi.org/10.1049/iet-rsn.2016.0055>
- Wang, Z., Li, G., Yang, L.: Dynamic hand gesture recognition based on micro-Doppler radar signatures using hidden Gauss–Markov models. *Geosci. Rem. Sens. Lett. IEEE*. 18(2), 291–295 (2021). <https://doi.org/10.1109/LGRS.2020.2974821>
- Min, R., et al.: Early gesture recognition with reliable accuracy based on high resolution IoT radar sensors. *IEEE Internet Things J.* 8, 15396–15406 (2021). <https://doi.org/10.1109/JIOT.2021.3072169>
- Dardas, N.H., Georganas, N.D.: Real-time hand gesture detection and recognition using bag-of-features and support vector machine techniques. *IEEE Trans. Instrum. Meas.* 60(11), 3592–3607 (2011). <https://doi.org/10.1109/TIM.2011.2161140>
- Li, G., et al.: Sparsity-driven micro-Doppler feature extraction for dynamic hand gesture recognition. *IEEE Trans. Aero. Electron. Syst.* 54(2), 655–665 (2018). <https://doi.org/10.1109/TAES.2017.2761229>
- Liu, W., et al.: A survey of deep neural network architectures and their applications. *Neurocomputing*. 234, 11–26 (2017). <https://doi.org/10.1016/j.neucom.2016.12.038>
- Kim, Y., Toomajian, B.: Hand gesture recognition using micro-Doppler signatures with convolutional neural network. *IEEE Access*. 4, 7125–7130 (2016). <https://doi.org/10.1109/ACCESS.2016.2617282>
- Hazra, S., Santra, A.: Short-range radar-based gesture recognition system using 3D CNN with triplet loss. *IEEE Access*. 7, 125623–125633 (2019). <https://doi.org/10.1109/ACCESS.2019.2938725>
- Lei, W., et al.: Continuous gesture recognition based on time sequence fusion using MIMO radar sensor and deep learning. *Electronics*. 9(5), 869 (2020). <https://doi.org/10.3390/electronics9050869>
- Jiang, W., et al.: Towards environment independent device free human activity recognition. In: Proceedings of the 24th Annual International Conference on Mobile Computing and Networking, MobiCom '18, Association for Computing Machinery, pp. 289–304. New York (2018). <https://doi.org/10.1145/3241539.3241548>
- Guo, L., et al.: Towards CSI-based diversity activity recognition via LSTM-CNN encoder-decoder neural network. *Neurocomputing*. 444, 260–273 (2021). <https://doi.org/10.1016/j.neucom.2020.02.137>
- Liu, L., et al.: From BoW to CNN: two decades of texture representation for texture classification. *Int. J. Comput. Vis.* 127(1), 74–109 (2019). <https://doi.org/10.1007/s11263-018-1125-z>
- Wen, Y., et al.: A discriminative feature learning approach for deep face recognition. In: Leibe, B., et al. (Eds.), vol. 9911, pp. 499–515. Springer International Publishing, Cham (2016). https://doi.org/10.1007/978-3-319-46478-7_31
- He, K., et al.: Delving deep into rectifiers: surpassing human-level performance on ImageNet classification. In: 2015 IEEE International

- Conference on Computer Vision (ICCV), pp. 1026–1034. IEEE, Santiago (2015). <https://doi.org/10.1109/ICCV.2015.123>
36. Bai, X., et al.: Radar-based human Gait recognition using dual-channel deep convolutional neural network. *IEEE Trans. Geosci. Rem. Sens.* 57(12), 9767–9778 (2019). <https://doi.org/10.1109/TGRS.2019.2929096>
37. Gisbrecht, A., Schulz, A., Hammer, B.: Parametric nonlinear dimensionality reduction using kernel t-SNE. *Neurocomputing.* 147, 71–82 (2015). <https://doi.org/10.1016/j.neucom.2013.11.045>

How to cite this article: Wang, L., et al.: Adaptive framework towards radar-based diversity gesture recognition with range-Doppler signatures. *IET Radar Sonar Navig.* 1–16 (2022). <https://doi.org/10.1049/rsn2.12280>