International Journal of Microwave and Wireless Technologies

cambridge.org/mrf

Hand gesture recognition method using FMCW radar based on multidomain fusion

Tianhong Yang and Hanxu Wu

College of Electronic Information Engineering, Shenyang Aerospace University, Shenyang, China

Research Paper

Cite this article: Yang T, Wu H (2023). Hand gesture recognition method using FMCW radar based on multidomain fusion. *International Journal of Microwave and Wireless Technologies* 1–9. https://doi.org/10.1017/S1759078723000995

Received: 27 February 2023
Revised: 25 July 2023
Accepted: 27 July 2023

Keywords:
convolutional neural network; Frequency modulated continuous wave radar; hand gesture recognition; multidomain fusion

Corresponding author: Tianhong Yang;
Email: ytianhong@126.com

Abstract

Radar-based hand gesture recognition is a potential noncontact human–machine interaction technique. To enhance the recognition performance of hand gesture, a multidomain fusion-based recognition method using frequency-modulated continuous wave radar is proposed in this article. The received raw echo data of gestures is preprocessed to obtain the range–time matrix, Doppler–time matrix, and range–Doppler–frame tensor. The obtained three-domain radar data corresponding to each gesture are input into the three-channel convolutional neural network for feature extraction. In particular, the extracted features from three-domain data are fused with learnable weight matrices to obtain the final gesture classification results. The experimental results have shown that the classification accuracy of the proposed multidomain fusion network based on learning weight matrix-based fusion is 98.45%, which improves the classification performance compared with the classic average-based fusion and concatenation fusion.

Introduction

In recent years, hand gesture recognition technique has found wide applications in the field of human–machine interaction [1–3]. It has become a new research hotspot in the fields of sign language translation, vehicle infotainment systems, intelligent home, etc. Typically, the hand gesture recognition systems are implemented using optical cameras and wearable sensors. The optical camera systems are sensitive to the illumination condition and have low privacy. The wearable sensors can normally work only under the contact condition and bring uncomfortable use experience. Fortunately, the radar sensor can recognize different hand gestures in the case of noncontact operation, bad lighting condition, and non-line-of-sight [4–6]. Therefore, radar-based hand gesture recognition has received substantial attention in both academic world and industry world.

The radar sensors for hand gesture recognition can be broadly categorized as unmodulated continuous wave (CW) radar, pulse radar, and frequency-modulated continuous wave (FMCW) radar. In papers [7, 8], the micro-Doppler spectrograms corresponding to different gesture echoes are acquired by unmodulated single-tone CW radar. Then, the micro-Doppler spectrograms are classified by convolutional neural network (CNN) to realize gesture recognition. However, the single-tone CW radar cannot provide range information of gestures, which leads to the limited recognition performance of the system. Due to its ability to provide both range and Doppler information, FMCW radar system has been widely used for hand gesture recognition [9–11]. Some researchers adopt FMCW radar system with multiple-input multiple-output architecture to obtain the angle information of hand gesture [12, 13]. However, the angle information is only applicable to gestures with large-scale motion. In paper [14], a multistatic FMCW radar with one transmitter and four receivers is used to collect the gesture echo data. The spectrograms of four receivers are used as the input of multichannel two-dimensional CNN (2-D-CNN) to increase the recognition accuracy of gestures. Unfortunately, 2-D-CNN can only learn the spatial feature of the spectrograms for each gesture. Three-dimensional CNN (3-D-CNN) is proposed to extract the spatiotemporal characteristic information of dynamic gesture to increase the classification accuracy [15, 16]. In paper [15], only the range–Doppler–frame tensor is used as the input of 3-D-CNN to achieve the gesture classification, which makes the feature extraction incomplete. In paper [16], the range–Doppler map time sequence (RDMTS) and range–azimuth map time sequence (RAMTS) are combined to improve the gesture classification performance. However, the extract features of RAMTS and RDMTS are fused by the simple concatenation fusion method, which ignores the importance score of each domain data for the gesture recognition.

To solve the above problem, a multidomain fusion network for hand gesture recognition based on FMCW radar is developed. First, the range–time matrix, the Doppler–time matrix, and the range–Doppler frame tensor are obtained by preprocessing the original gesture echo signal. Then, the range–time matrix and the Doppler–time matrix are used as inputs of

https://doi.org/10.1017/S1759078723000995 Published online by Cambridge University Press

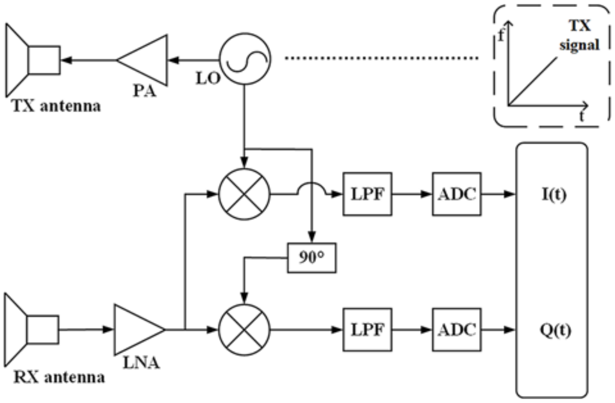


Figure 1. Block diagram of the used FMCW radar system.

2-D-CNNs, respectively. The range–Doppler–frame tensor is used as the input of 3-D-CNN. The trainable weight matrix is used to fuse the extracted feature information from different domain data so that the most comprehensive feature information is obtained in the fusion module. Finally, the additional dense layer and Softmax function are employed for multiclass classification of hand gestures. Experiments are conducted to prove the effectiveness of the multidomain radar data and the superiority of the learning weight matrix-based fusion method compared to the average-based fusion and concatenation fusion.

Reconstruction of multidomain data

Signal model

The simplified block diagram of the used FMCW radar is depicted in Fig. 1. The FMCW radar consists of a waveform generator, a transmitting antenna, a receiving antenna, a quadrature demodulator, two low-pass filters, and two analog-to-digital converters. The FMCW signal generated by the waveform generator is sent to the transmitting antenna. The transmitted signal reflected by the hand is received by the receiving antenna. The received signal is then amplified and mixed with transmitted signal to obtain the beat signal.

For an FMCW radar, the received signal can be expressed as

$$S_{rx}(t) = A_{rx}\cos(2\pi(f_0(t-\tau) + \frac{r}{2}(t-\tau)^2) + \varphi_{rx}) \quad (1)$$

where τ represents the round-trip propagation time delay, f_0 is the carrier frequency, r is the chirp rate, φ_{rx} is the initial phase of the received signal, and A_{rx} is the amplitude of the received signal and expressed as

$$A_{rx} = \frac{G\lambda\sqrt{P\sigma}}{(4\pi)^{1.5}R^2\sqrt{L}} \quad (2)$$

where G is the antenna gain; P and λ are the power and wavelength of transmitted signal, respectively; σ is the radar cross section of target; and L denotes other losses. After the received signal is demodulated by the I/Q demodulator, the beat signal is obtained as

$$S(t) = I(t) + jQ(t) = A\exp[j\psi(t)] \quad (3)$$

where $\psi(t)$ is the phase of the signal.

Data preprocessing

The range–time matrix, Doppler–time matrix and range–Doppler–frame tensor corresponding to each gesture can be obtained by preprocessing the beat signal collected by FMCW radar. The flowchart of raw echo data preprocessing is shown in Fig. 2. First, the original echo data are reshaped into a two-dimensional raw data matrix, where the horizontal axis represents the slow time dimension and the vertical axis represents the fast time dimension. The raw data matrix is transformed into the range–time matrix by performing the fast Fourier transform (FFT) along the fast time dimension. Then the fourth-order Butterworth high-pass filter with the cut-off frequency of 0.0075 Hz is used as moving target indicator to suppress the background static clutter in the range–time matrix. Considering the distance between the hand and the radar for different gestures, the range bins starting from 0.75 m to 2.25 m are chosen for further short time Fourier transform (STFT) processing. The STFT with length of 0.2 s and overlap coefficient of 95% is performed for each range bin along the slow time dimension. Accordingly, the Doppler–time matrix is obtained by coherently summing the STFT results of each range bin. A frame is established by stacking arrays of the sampled beat signals for a certain number of frequency modulation periods. A series of range–Doppler matrix can be obtained by performing FFT along the slow time for each frame. Finally, the range–Doppler–frame tensor is obtained by stacking the range–Doppler matrix of each frame. In this article, the range–Doppler–time tensor is composed of 20 frames, with the duration of 100 ms per frame.

The range–time matrix, Doppler–time matrix, and range–Doppler–time tensor are combined to implement the classification of hand gestures. Since each domain data has its unique and valuable feature information, how to effectively integrate the features of three-domain data is critical to enhance the classification performance of hand gestures.

Proposed multidomain fusion network

In order to make full use of feature information of each domain, each domain data is input into different CNN models according to its own property. The entire architecture of the proposed multidomain fusion network is depicted in Fig. 3. The proposed network is composed of three parts: feature extraction module, feature fusion module, and classification module.

Feature extraction

In order to extract the feature information from the three-domain data, different network models are designed according to the different characteristics of three-domain data. The range–time matrix contains the time-varying range information between the radar and the hand. The range–Doppler matrix reflects the time-varying Doppler frequency information of various scattering points of the hand. Each range–Doppler frame characterizes the changes in range–Doppler signature over time. In the proposed network, two 2-D-CNNs with the same structure are used to extract the spatial features of range–time matrix and Doppler–time matrix, respectively. A 3-D-CNN model is built to extract the spatiotemporal features of range–Doppler–frame tensor. As shown in Fig. 4(a), the built 2-D-CNN model consists of three 2-D convolution layers, each of which is followed by a rectified linear unit (ReLU) activation function and a max pooling layer. The last max pooling layer is

https://doi.org/10.1017/S1759078723000995 Published online by Cambridge University Press

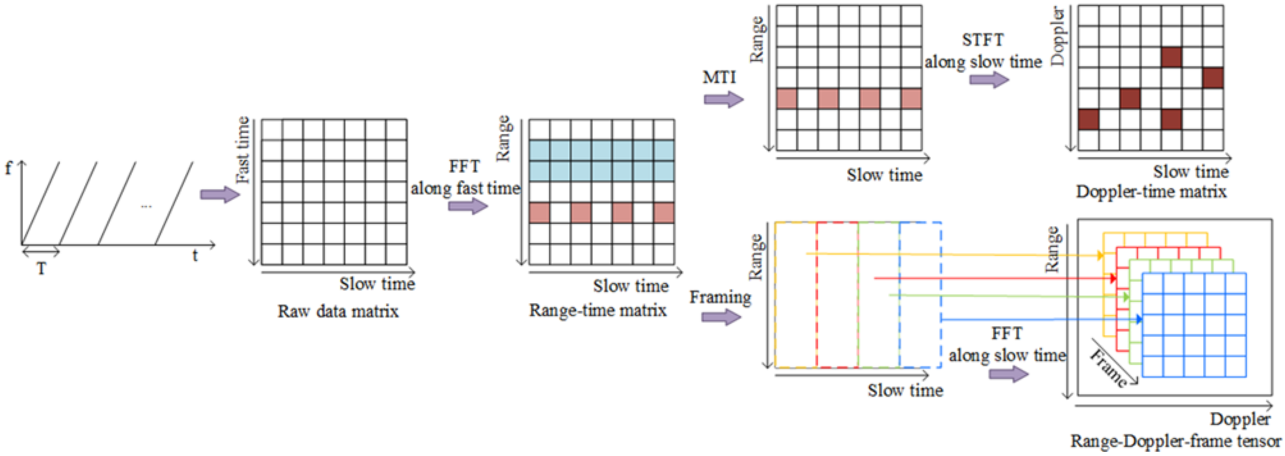


Figure 2. Flowchart of raw echo data preprocessing.

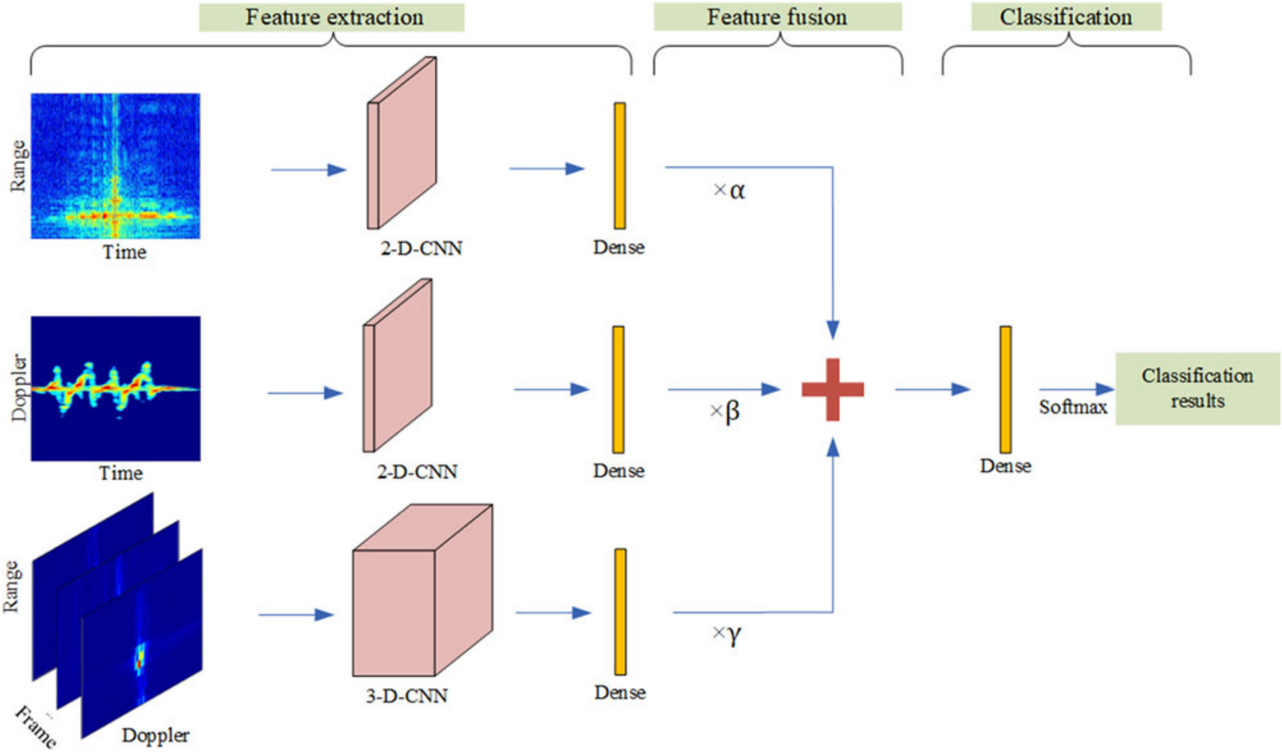


Figure 3. The proposed network architecture for hand gesture recognition.

followed by a dense layer, which outputs the 10 × 1 feature vector. In order to reduce the parameters of network, the built 3-D-CNN model only contains two 3-D convolution layers, each of which is followed by a ReLU activation function and a 3-D max pooling layer. Finally, the feature vector with 10 × 1 dimension is output through the dense layer. The structure of 3-D-CNN model is shown in Fig. 4(b).

Feature fusion

The feature fusion is carried out on the extracted 1-D feature vectors from three-domain data. Since the three feature vectors contain different feature information of gestures, they have different importance scores the final recognition results. In order to obtain the complementary feature, the trainable weight matrices are used to merge three feature vectors. The fused feature vector can be expressed as

$$\mathbf{F}_{\text{fuse}} = \alpha \times \mathbf{F}_1 + \beta \times \mathbf{F}_2 + \gamma \times \mathbf{F}_3 \tag{4}$$

where $\mathbf{F}_1$, $\mathbf{F}_2$ and $\mathbf{F}_3$ are extracted feature vectors from range–time matrix, Doppler–time matrix, and Doppler–range–frame tensor, respectively. α, β, and γ are the trainable weight matrices of the three feature vectors, respectively. Because the extracted features from different domain data have their own importance for hand gestures classification, the learnable weight matrices can help the model find the most appropriate fusion way to represent different significance of three-domain input data. If the contributions of features from three-domain data are consider to be equal, the

https://doi.org/10.1017/S1759078723000995 Published online by Cambridge University Press

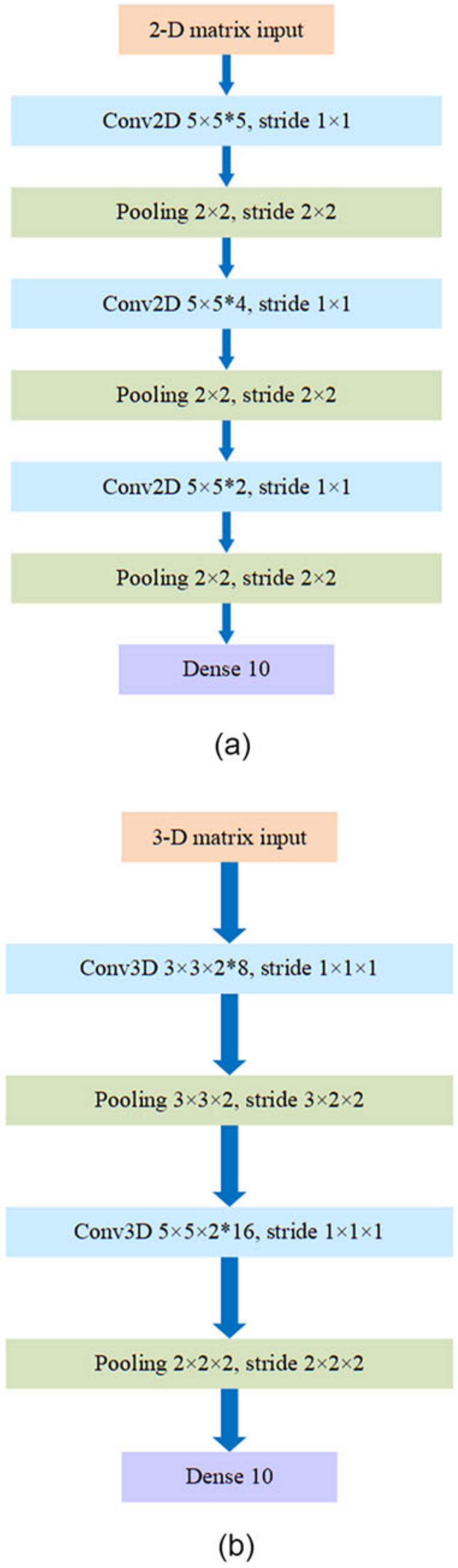


Figure 4. 2-D-CNN/3-D-CNN structure: (a) 2-D-CNN structure and. (b) 3-D-CNN structure.

recognition accuracy of gesture will be reduced, which will be proved this in the following experiment part. Since the fused vector obtained by the weight matrices contains importance proportion of each domain data, it can provide the most discriminative feature representation of three-domain data, which effectively improves the classification ability of the network model.

Classification

After fusing the features of the three-domain data, the final fused feature vector is obtained. Then the dense layer is used to map the fused feature to the label space. Finally, the Softmax function is deployed to convert the real values of the vector into probability values between 0 and 1. The output probability of Softmax function is given by

$$y_i = \frac{\exp(z_i)}{\sum_{j=1}^{k}\exp(z_j)} \quad i = 1, 2, \cdots, 8 \tag{5}$$

where y_i represents the prediction probability that the data belongs to the i th category, k denotes the total number of gesture classes and is equal to 8, andz_i is the output value of the ith neuron of the dense layer.

Given a set of labeled samples, the cross-entropy loss function is adopted to train the proposed network model. The loss function is given as

$$L = -\frac{1}{N}\sum_{n=1}^{N}\sum_{i=1}^{k} y'_{n,i}\log\left(y_{n,i}\right) \tag{6}$$

where N denotes the size of a minibatch, $y'_{n,i}$ is the actual prediction probability of the i th category corresponding to the nth sample, and $y_{n,i}$ represents the expected prediction probability of the i th category corresponding to the nth sample.

Experimental results

Data collection

The commercial K-band FMCW radar platform SDR-KIT-2400AD developed by Ancortek incorporation is used to collect data samples of different hand gestures. The radar transmits FMCW waveform with a carrier frequency of 24 GHz and a bandwidth of 2 GHz. The FMCW sweep time is set to 1 ms and the data sampling frequency is 128 kHz. The transmitted power of the system is 18 dBm, and the two horn antennas with 10 dBi gain are used for transmitting and receiving signals. The experiment is performed in an indoor laboratory with one subject sitting in front of the radar. The hand of the subject has a distance of approximately 30 cm from the antennas. The experimental scene is demonstrated in Fig. 5. The gesture data are recorded by five volunteers, which are composed of four men and one woman. Eight types of gestures used for evaluation are listed as follows: (a) raise, (b) swipe left to right, (c) swipe back to front, (d) circle, (e) push, (f) pinch, (g) flick, and (h) snap. The illustrative images of the different gestures are shown in Fig. 6. The recording time for each gesture is set as 2 s. Each subject is required to repeat each gesture 40 times and thus the total number of each gesture is 200 (5 subjects × 40 repetitions).

The collected raw radar data are preprocessed by MATLAB software to obtain range–time matrix, Doppler–time matrix, and range–Doppler–frame tensor for each gesture. The proposed network is implemented using Tensorflow and Keras library. All the

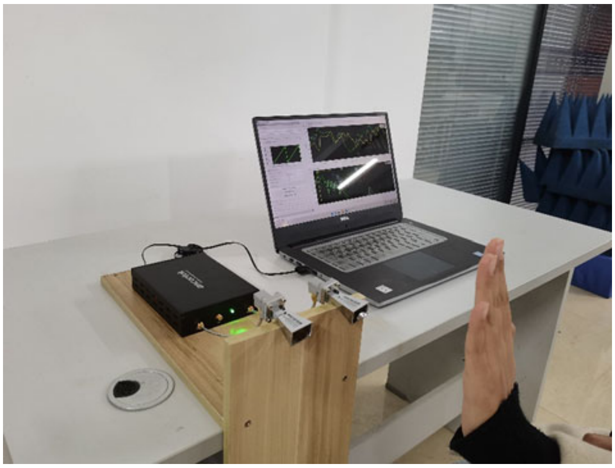


Figure 5. Experimental scene.

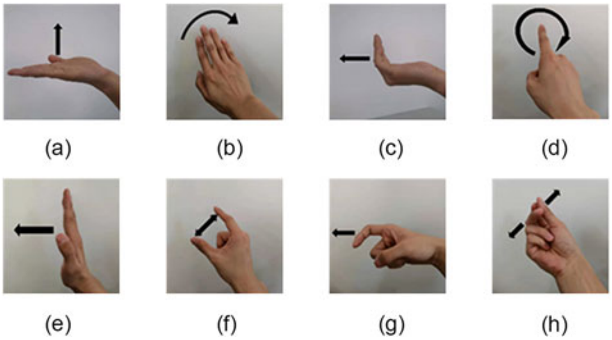


Figure 6. Illustrative images of the different gestures: (a) raise, (b) swipe left to right, (c) swipe back to front, (d) circle, (e) push, (f) pinch, (g) flick, and (h) snap.

experiments are performed on a workstation with an NVIDIA Quadro P5000 GPU and an Intel(R) Xeon(R) Gold 6132 processor.

Network training

The three-domain data are input into the proposed network model for training. The network is trained from scratch using a learning rate of 0.001 and the adaptive moment estimation (Adam) optimizer. The datasets are split into 60% for the training, 20% for the validation, and 20% for the testing. The training dataset and validation dataset are shuffled every epoch and the validation dataset is not included in the test dataset. The test dataset is used after the complete training to inspect the performance of the network. The network model is trained for 250 epochs. The minibatch size is set to 16 and each epoch consists of 60 batches. The training and validation accuracies are calculated at the end of each epoch. Figure 7 shows the accuracy of training and validation process for the proposed multidomain fusion network. It is observed from Fig. 7 that there is no further increase of accuracy when the number of training epochs is more than 180.

In the process of network training, the cross-entropy loss function which measures the difference between the predicted value and the actual value is exploited to evaluate the fitting degree of trainable parameters. Figure 8 shows the value of loss function for the training and validation process. As can be seen from Fig. 8, the loss value gradually decreases as the number of training epochs increases. When the training reaches the 180th epoch, the loss does not decrease further, which means that the network has attained maturity.

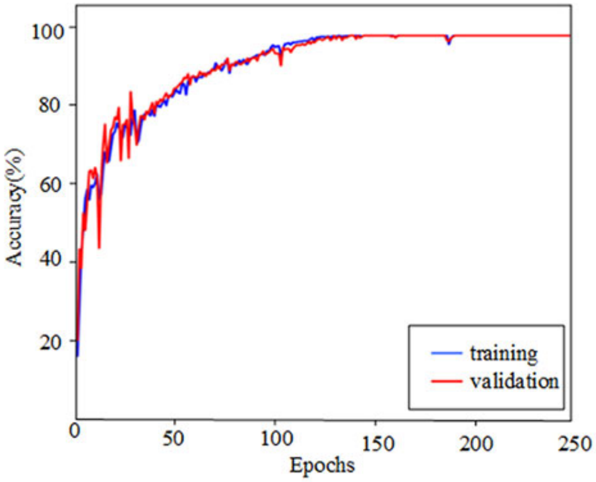


Figure 7. Training and validation accuracy at each epoch.

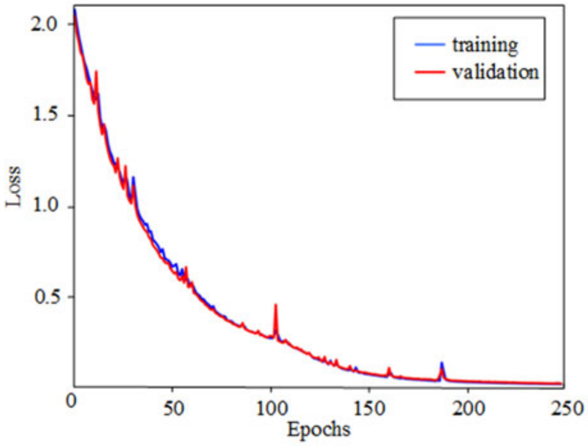


Figure 8. Training and validation loss at each epoch.

Performance evaluation

In order to compare the classification performance of multidomain data and single-domain data, four sets of experiments are performed to input the range–time matrix, the Doppler–time matrix [8], the range–Doppler–frame tensor [14] and three-domain data into the corresponding single-channel or three-channel network. In order to make the results more general, we conduct a fivefold cross-validation experiment for each case. The whole dataset is divided into five parts, where four parts are used for training and one part is kept for testing. The average confusion matrices of four cases are shown in Fig. 9. It can be observed from Fig. 9 that the proposed three-channel network with multidomain data input has the lower misclassification rates compared to other single-channel networks with single-domain data input. In addition, when only single-domain data are used as the input of the network, the misclassification rates is relatively high for the gestures of swipe left to right, circle, pinch, flick, and snap.

https://doi.org/10.1017/S1759078723000995 Published online by Cambridge University Press

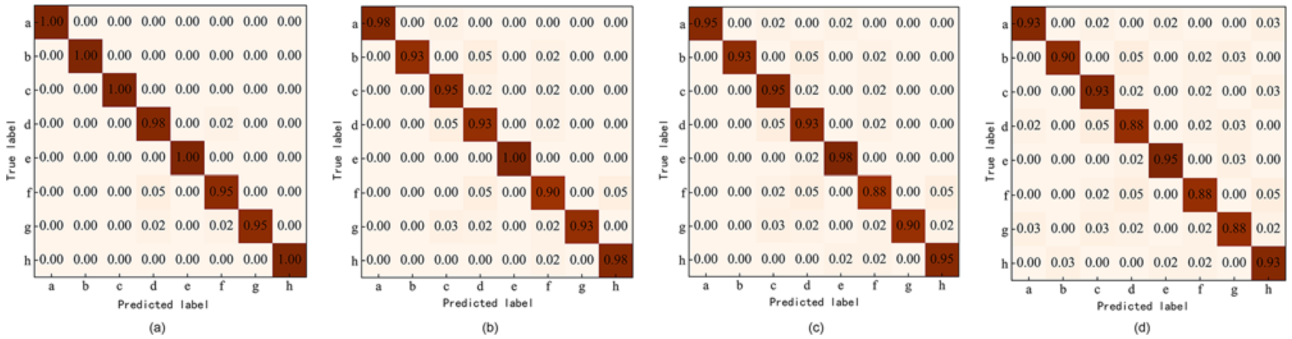


Figure 9. Confusion matrices: (a) three-domain data, (b) range–time matrix, (c) Doppler–time matrix [8], and (d) range–Doppler–frame tensor [14].

Table 1. Classification accuracies comparison of different input data

Input data	Average accuracy (%)
Three-domain data	98.45
Range–time matrix	90.63
Doppler–time matrix [8]	93.13
Range–Doppler–frame tensor [14]	94.69

Table 2. Classification accuracies of different fusion methods

Fusion method	Accuracy (%)
Learning weight matrix-based fusion	98.45
Average-based fusion	89.38
Concatenation fusion	96.56

This phenomenon is due to the fact that the hand movement of these gestures is relatively small and the single-channel networks which has incomplete feature extraction cannot distinguish these gestures well. Table 1 provides the comparison result of classification accuracies for different input data. As shown in Table 1, the classification performance using the range–Doppler–frame tensor data is the best for single-domain data input. However, the gesture feature information contained in the single-domain data is relatively incomplete, and the classification accuracies based on single-domain data are not satisfactory. When the three-domain data are used as the network input, more feature information of gesture actions can be extracted. The classification accuracy based on three-domain data can reach to 98.45%. Therefore, the recognition accuracy of multidomain data is higher than that of single-domain data.

In order to compare the effect of different fusion methods on the result of gesture recognition, the classification performances of three fusion modes are compared in this article. (1) Learning weight matrix-based fusion—It assigns a trainable weight matrix to the extracted feature vector from each domain data and adds the three feature vectors after multiplying the weight matrix. (2) Average-based fusion—It directly adds up the extracted feature vectors of three-domain radar data without considering the importance degree of each domain radar for hand gesture recognition. (3) Concatenation fusion—The extracted feature vectors from three-domain radar data are concatenated to obtain the new feature vector. Table 2 illustrates the classification accuracies of the different fusion methods. The experimental results in Table 2 show that the learning weight matrix-based fusion method has the highest classification accuracy. Due to the neglection of significance of each domain data, both average-based fusion and concatenation fusion are inferior to learning weight matrix-based fusion.

The scatter diagrams of feature space visualization based on t-distributed stochastic neighbor embedding (t-SNE) [17] for different fusion methods are shown in Figs. 10–12. It is observed from three figures that the overlap among the features extracted by the leaning weight matrix-based fusion is the smallest, which makes it easy to classify the eight gesture classes accurately and reliably. For the features extracted by the average-based fusion, the confusion between the different gestures is severe, especially for the gestures of pinch, circle, and snap, as shown in Fig. 11. It is seen from Fig. 12 that the two gestures of pinch and snap have the serious overlap, which indicates that the extracted features with the concatenation fusion make it difficult to separate two gestures due to the small-scale motion amplitude of gestures.

The effects of learning rate, minibatch size, and optimizer on the performance of the proposed network are also investigated. First, the minibatch size is set to 16 and the adaptive moment estimation (Adam) optimizer is used for optimization, the values of learning rates are set to 0.0005, 0.001, 0.002, and 0.003, respectively. The classification accuracies and convergence epochs of network training for different learning rates are shown in Table 3. If the learning rate is too small, it will take much time for the network to complete the convergence in the training process and the classification accuracy will be relatively low. However, if the learning rate is too large to learn subtle features, the classification accuracy will not reach the optimal level. When the learning rate is set to 0.001, the classification performance is the best. Secondly, the learning rate is fixed to 0.001 and the Adam optimizer is used, the classification accuracies and convergence epochs for different minibatch sizes are shown in Table 4. It is observed from Table 4 that the network can make a good tradeoff between the classification accuracy and the convergence epochs when the minibatch size is equal to 16. Finally, the effects of different optimizers such as adaptive gradient (Adagrad), stochastic gradient descent (SGD), and Adam on the classification accuracy and convergence epochs are tested. The experimental results are shown in Table 5. It can be observed that the Adam optimizer can achieve the highest classification accuracy and the fastest convergence among the three optimizers.

The hand gesture recognition of unknown people based on trained data from known people is very important for practical application. In order to evaluate the generalization ability of the proposed network, the leave one subject out cross-validation method is used to separate the training and testing data, where the data from one of the five subjects are selected for evaluating

https://doi.org/10.1017/S1759078723000995 Published online by Cambridge University Press

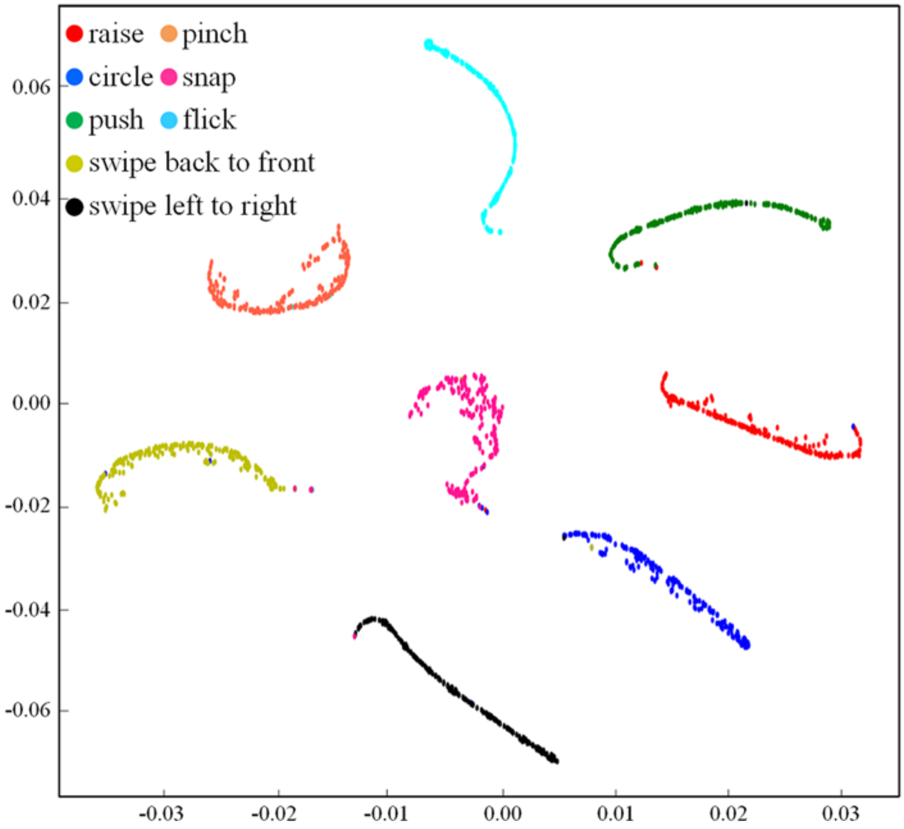


Figure 10. Scatter diagram of learning weight matrix-based fusion.

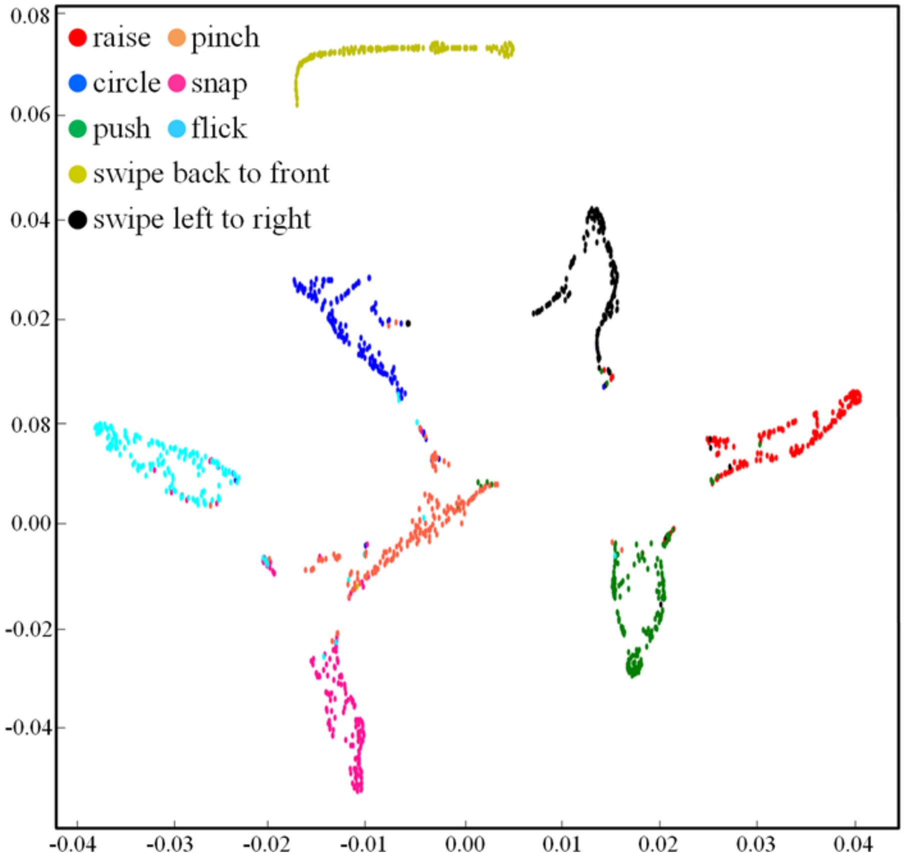


Figure 11. Scatter diagram of average-based fusion.

https://doi.org/10.1017/S1759078723000995 Published online by Cambridge University Press

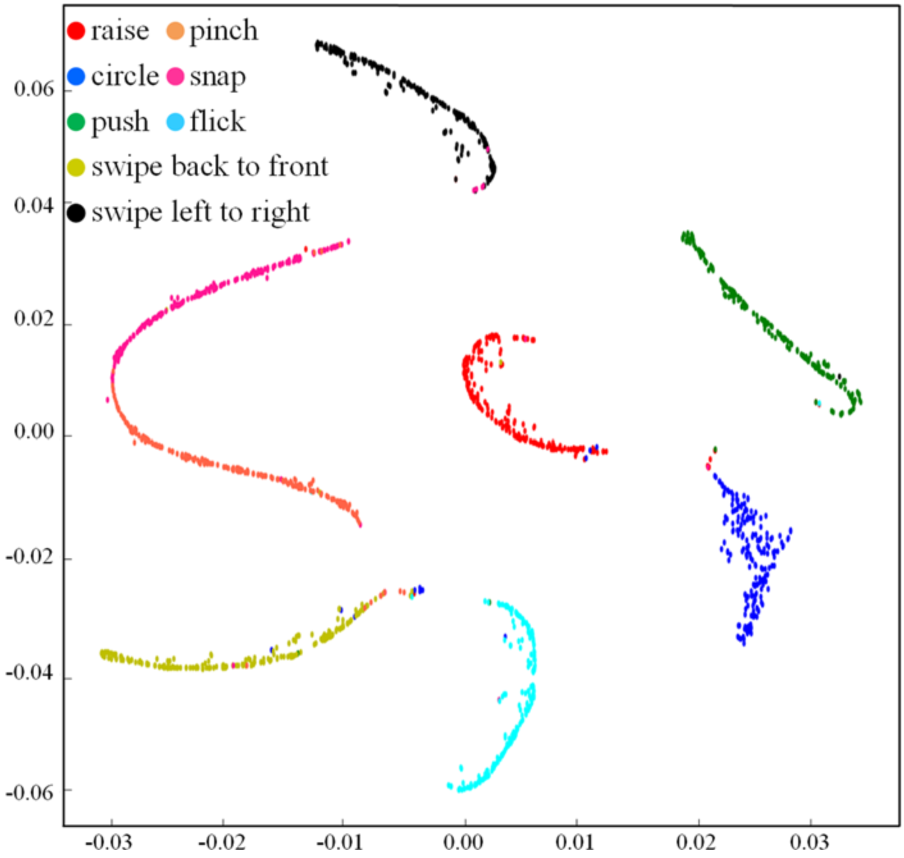


Figure 12. Scatter diagram of concatenation fusion.

Table 3. Classification accuracies and convergence epochs of different learning rates

Learning rate	0.0005	0.001	0.002	0.003
Accuracy (%)	92.18	98.45	96.87	94.68
Convergence epochs	400	180	120	80

Table 4. Classification accuracies and convergence epochs of different minibatch sizes

Minibatch size	8	16	32	64
Accuracy (%)	94.92	98.45	98.75	98.13
Convergence epochs	170	180	350	420

Table 5. Classification accuracies and convergence epochs of different optimizers

Optimizer	Adagrad	SGD	Adam
Accuracy (%)	95.70	94.06	98.75
Convergence epochs	270	350	180

the performance and the data of the remaining four subjects are used for training the network model. Each of five subjects labeled from A to E is tested when the training data and testing data change sequentially. Table 6 shows the classification accuracy for each subject. It is observed from Table 6 that the gesture classification accuracy is higher than 97% for each subject. Hence, the proposed network model has the robust classification performance for different subjects.

Table 6. Classification accuracy of each subject

Subject	A	B	C	D	E
Accuracy (%)	98.13	97.19	99.06	98.75	98.44

Conclusion

Gesture recognition has wide application in human–machine interaction. In this article, we propose a multidomain fusion network architecture for hand gesture recognition method based on FMCW radar system. The proposed method uses two 2-D-CNNs and one 3-D-CNN to extract feature information from the range–time matrix, Doppler–time matrix, and range–Doppler–frame tensor. Specifically, the learning weight matrix-based fusion method is developed to fuse the features of multidomain radar data. The experimental results show that the recognition accuracy of multidomain input is 98.45%, which is higher than that of single-domain input. The learning weight matrix-based fusion can fuse multidomain features more effectively and improve the hand gesture classification performance. In addition, the experiment results of the leave one subject out cross-validation prove that the proposed network has the satisfactory generalization ability. In the future, we try to develop the network model with smaller size and higher accuracy for embedded implementation.

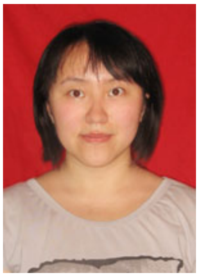
Acknowledgements. This work was support in part by the Aeronautical Science Foundation of China under Grant 2019ZC054004, and in part by the Scientific Research Project of the Education Department of Liaoning Province under Grant JYT2020015.

Competing interest. The authors report no conflict of interest.

https://doi.org/10.1017/S1759078723000995 Published online by Cambridge University Press

References

1. **Guo L, Lu Z and Yao L** (2021) Human-machine interaction sensing technology based on hand gesture recognition: A review. *IEEE Transactions on Human-Machine Systems* **51**(4), 300–309.
2. **Plouffe G and Cretu M** (2016) Static and dynamic hand gesture recognition in depth data using dynamic time warping. *IEEE Transactions on Instrumentation and Measurement* **65**(2), 305–316.
3. **Li G, Zhang R, Ritchie M and Griffiths H** (2018) Sparsity-driven micro-Doppler feature extraction for dynamic hand gesture recognition. *IEEE Transactions on Aerospace and Electronic Systems* **54**(2), 655–665.
4. **Khan F, Leem SK and Cho SH** (2017) Hand-based gesture recognition for vehicular applications using IR-UWB radar. *Sensors* **17**(4), 833.
5. **Li Y, Gu C and Mao J** (2022) 4-D gesture sensing using reconfigurable virtual array based on a 60-GHz FMCW MIMO radar sensor. *IEEE Transactions on Microwave Theory and Techniques* **70**(7), 3652–3665.
6. **Scherer M, Magno M, Erb J, Mayer P, Eggimann M and Benini L** (2021) TinyRadarNN: Combining spatial and temporal convolutional neural networks for embedded gesture recognition with short range radars. *IEEE Internet of Things Journal* **8**(13), 10336–10346.
7. **Zhang J, Tao J and Shi ZG** (2019) Doppler radar-based hand gesture recognition system using convolutional neural networks. In *International Conference on Communications, Signal Processing, and Systems*, 1096–1113.
8. **Kim Y and Toomajian B** (2016) Hand gesture recognition using micro-Doppler signatures with convolutional neural network. *IEEE Access* **4**, 7125–7130.
9. **Zhang X, Wu Q and Zhao D** (2018) Dynamic hand gesture recognition using FMCW radar sensor for driving assistance. In *2018 10th International Conference on Wireless Communications and Signal Processing (WCSP)*, 1–6.
10. **Sun Y, Tai F, Schliep F and Pohl N** (2018) Gesture classification with handcrafted micro-Doppler features using a FMCW radar. In *2018 IEEE MTT-S International Conference on Microwaves for Intelligent Mobility (ICMIM)*, 1–4.
11. **Zhang Z, Tian Z and Zhou M** (2018) Dynamic continuous hand gesture recognition using FMCW radar sensor. *IEEE Sensors Journal* **18**(8), 3278–3289.
12. **Wang Y, Shu Y, Jia X, Zhou M, Xie L and Guo L** (2022) Multifeature fusion-based hand gesture sensing and recognition system. *IEEE Geoscience and Remote Sensing Letters* **19**, 1–5.
13. **Xia X, Luomei Y, Zhou C and Xu F** (2021) Multidimensional feature representation and learning for robust hand-gesture recognition on commercial millimeter-wave radar. *IEEE Transactions on Geoscience and Remote Sensing* **59**(6), 4749–4764.
14. **Chen Z, Li G, Fioranelli F and Griffiths H** (2019) Dynamic hand gesture classification based on multistatic radar micro-Doppler signatures using convolutional neural network. In *IEEE Radar Conference (RadarConf19)*, 1–5.
15. **Skaria S, Hourani AA and Evans RJ** (2020) Deep-learning methods for hand-gesture recognition using ultra-wideband radar. *IEEE Access* **8**, 203580–203590.
16. **Lei W, Jiang X, Xu L, Luo J, Xu M and Hou F** (2020) Continuous gesture recognition based on time sequence fusion using MIMO radar sensor and deep learning. *Electronics* **9**(5), 869.
17. **van der Maaten L and Hinton G** (2008) Visualizing high-dimensional data using t-SNE. *Journal of Machine Learning Research* **9**(11), 2579–2605.



Tianhong Yang received the B.E. and M.E. degrees from Jilin University, Changchun, China, in 2005 and 2007, respectively. She is currently an assistant professor with the College of Electronic Information Engineering, Shenyang Aerospace University, Shenyang, China. Her current research interests include ultra-wideband radars signal processing and antenna design.

Hanxu Wu received the B.E. degree from Shenyang Aerospace University, Shenyang, China, in 2020. He is currently pursuing an M.E. degree with the College of Electronic Information Engineering, Shenyang Aerospace University. His main research interests include radar-based hand gesture recognition and deep learning.

https://doi.org/10.1017/S1759078723000995 Published online by Cambridge University Press