Contents lists available at YXpublications

International Journal of Applied Mathematics in Control Engineering

Journal homepage: http://www.ijamce.com

Human Motion Recognition Method Based on Two-stream Feature Fusion of Millimeter Wave Radar

Yongqiang Zhang[a,b,c], Ziqiang Zhang[a], Yuejian Shen[d], Jinlong Ma[a,b,c], Weidong Wu[a,c]*

[a] *School of Information Science and Engineering, Hebei University of Science and Technology, Shijiazhuang, Hebei 050018, China*

[b] *Hebei Technology Innovation Centre of Intelligent IoT, Shijiazhuang, Hebei 050018, China*

[c] *Shijiazhuang Intelligent Communication IoT Industrial Technology Research Institute, Shijiazhuang, Hebei 050018, China*

[d] *School of Information Engineering, Shandong Huayu University of Technology, Dezhou, Shandong 253000, China*

ARTICLE INFO

Article history:
Received 3 January 2024
Accepted 11 March 2024
Available online 12 March 2024

Keywords:
Electromagnetic interference
Millimeter wave radar
Spatial pyramid pooling algorithm
Channel attention mechanism

ABSTRACT

With the rapid development of science, technology and electronic products, more and more intelligent products have complex structures and high integration levels. These products have extremely harsh electromagnetic environments due to factors such as mixed frequency bands, large power, and dense distribution. In order to solve the problem that millimeter-wave radar cannot effectively capture human body motion characteristics in an electromagnetic interference and noise environment, an anti-interference recognition algorithm based on two-stream feature fusion is proposed. The algorithm consists of three modules: distance feature extraction, Doppler feature extraction and feature fusion. The spatial pyramid pooling algorithm and channel attention mechanism are used to improve the VGG16 network and ResNet50 network respectively, improving the feature extraction capabilities of the network. Experimental results show that the recognition accuracy of our algorithm is 98.5%, and comparison with typical algorithms in the same field verifies the robustness of the algorithm in electromagnetic interference noise scenarios.

1. Introduction

The rapid development of artificial intelligence technology has given human action recognition technology broad application prospects in the fields of smart homes, medical monitoring, intelligent security, and assisted driving [1, 2]. In recent years, human action recognition using wearable devices, cameras, radar and other technologies has become a research hotspot [3, 4]. However, wearable methods mainly have problems such as physical discomfort, poor convenience and low robustness [5]. Cameras have the advantages of low cost, long detection range, and mature algorithms. They are currently the mainstream method in the field of human action recognition. However, the recognition rate of optical cameras is extremely low in scenes with insufficient light at night or exposure to direct sunlight. At the same time, there is a risk of privacy leakage in some special occasions [6]. In contrast, millimeter wave radar has high stability and privacy, and is not affected by weather, light and other environments. Car-grade millimeter-wave radar is generally a 77GHz FMCW. It only needs to extract and characterize the motion information of the Doppler frequency shift in the human body echo signal to identify the type of human movement detected by the radar at this time [7]. Therefore, there are an increasing number of human action recognition algorithms based on millimeter wave radar at home and abroad.

The traditional human action recognition method first performs manual feature extraction on the original millimeter wave radar signal, and then uses the extracted features as a classifier. T. Yang used millimeter-wave radar to collect human action echo signals, he used signal processing methods to manually extract features from the radar images and then further classify them. Although the method of manual feature extraction is simple, manual feature extraction is more affected by human subjective factors[8]. Fan Zhengguang and others used FMCW radar to collect echo signals of human body movements and converted them into micro-Doppler time-frequency images. Finally, they used support vector machines and long short-term memory networks for classification and recognition. The average recognition accuracy reached 80.7%, but currently Only in the laboratory stage[9]. Zhao considered that interference from non-target micro-motion in practical applications would lead to changes in the micro-Doppler characteristics of the target human body movement and affect the recognition effect, so they proposed a new radar signal pre-processing architecture, using the empirical model eliminate radar signal interference caused by non-target motion. This

* Corresponding author.
E-mail addresses: wwd@hebust.edu.cn (W Wu)
Digital Object Identifiers: https://doi.org/10.62953/IJAMCE.548712

method is better than existing recognition methods in non-target motion interference environment [10]. Ding C and others used the peak search method to extract micro-Doppler features from millimeter wave radar signals, and finally used k-Nearest Neighbor (KNN) to classify them. The calculation efficiency is high, but the accuracy of the algorithm needs to be improve[11]. Sakagami F and others converted radar signals into three dimensional features of time Doppler, time range and range Doppler as human action data, and then used Convolutional Neural Network (CNN) for classification. The results showed that multi-dimensional features It can effectively improve the classification accuracy, and the final accuracy rate on the self-built data set is as high as 95%[12].

However, most studies on human motion recognition in an ideal laboratory environment do not take into account the complex environment in practical applications. When two radars on the same frequency work together, signal interference will occur. Eventually, the human motion characteristics in the radar echo signal are not obvious or submerged, resulting in a significant decrease in the accuracy of the recognition algorithm in practical applications. Electromagnetic interference in practical applications is diverse, random, and uncertain, and traditional radar signal denoising methods or image denoising methods alone cannot remove all noise.

This paper proposes a human action recognition algorithm based on distance-time diagrams and Doppler diagrams. First, single-dimensional features are used to develop and improve a single algorithm, and a suitable network is found to extract feature information of range-time diagrams and Doppler diagrams respectively. After both networks are trained, transfer learning is used to combine them into a multi-feature multi-branch network, and then the feature information extracted by the two networks is fused. Finally, only the final fully connected layer and classifier need to be simply trained to complete the classification task, which greatly shortens the training time of multi-branch networks.

2. Related works

2.1 Convolutional neural network

There are many vision-related neurons in the human brain, and these neurons are connected to each other through weak electrical signals. In neural networks, neuron is the most basic unit, which can be regarded as a logical computing unit. The working process of a single neuron is: for the given input $x_1,x_2,\ldots,x_n$, perform multiplication operations with the corresponding weights $w_1,w_2,\ldots,w_n$ respectively, and then add the offset b to the result. Finally, the activation function is used to convert and the corresponding output is obtained [13].

$$y = f\left(\sum_{i=1}^{n} x_i w_i + b\right) = f\left(x_1 w_1 + x_2 w_2 + \ldots + x_n w_n + b\right) \quad (1)$$

Multiple neurons connected to each other form a simple neural network, also known as a multilayer perceptron model. Assuming there are multiple inputs x_1,x_2,x_3, and the final output result is $f(x)$.

$$h_1^{(2)} = g\left(z_1^{(2)}\right) = g\left(\theta_{10}^{(1)} x_0 + \theta_{11}^{(1)} x_1 + \theta_{12}^{(1)} x_2 + \theta_{13}^{(1)} x_3\right) \quad (2)$$

$$h_2^{(2)} = g\left(z_2^{(2)}\right) = g\left(\theta_{20}^{(1)} x_0 + \theta_{21}^{(1)} x_1 + \theta_{22}^{(1)} x_2 + \theta_{23}^{(1)} x_3\right) \quad (3)$$

$$h_3^{(2)} = g\left(z_3^{(2)}\right) = g\left(\theta_{30}^{(1)} x_0 + \theta_{31}^{(1)} x_1 + \theta_{32}^{(1)} x_2 + \theta_{33}^{(1)} x_3\right) \quad (4)$$

$$h_3^{(2)} = g\left(z_3^{(2)}\right) = g\left(\theta_{30}^{(1)} x_0 + \theta_{31}^{(1)} x_1 + \theta_{32}^{(1)} x_2 + \theta_{33}^{(1)} x_3\right) \quad (5)$$

$$\begin{aligned} f(x) &= h_1^{(3)} \\ &= g(\theta_{10}^{(2)} h_0^{(2)} + \theta_{11}^{(2)} h_1^{(2)} + \theta_{12}^{(2)} h_2^{(2)} + \theta_{13}^{(2)} h_3^{(2)} + \theta_{14}^{(2)} h_4^{(2)}) \end{aligned} \quad (6)$$

where θ represents the matrix that controls the mapping parameters, $\theta^{(1)} \in \mathbf{R}^{4\times4}$, $\theta^{(2)} \in \mathbf{R}^{1\times5}$, $\mathbf{Z}_i^{(j)}$ represents the result of the weighted linear combination on the input, g represents the activation function, and $\mathbf{h}_i^{(j)}$ represents the calculated value of the *i*-th neuron in layer *j*[14].

CNN is a deep learning model that has wide applications in the fields of image processing and computer vision[15]. The convolutional layer is the most important part. The output feature map *out* size of the convolutional layer is determined by multiple factors, including the input image *in* size, convolution *kernel* size, *stride*, *padding*, and *bias*.

$$height_{out} = \frac{height_{in} - height_{kernel} + 2 \times padding}{stride} + bias + 1 \quad (7)$$

$$width_{out} = \frac{width_{in} - width_{kernel} + 2 \times padding}{stride} + bias + 1 \quad (8)$$

The fully connected layer is usually located at the end of the CNN, and the neurons between two adjacent fully connected layers are fully connected to synthesize the features extracted by the previous convolutional network and output the final classification result. Therefore, in order to ensure that the output value satisfies the probability distribution, the normalized exponential (Softmax) function is usually used behind the fully connected layer. Assume that a certain output value after passing through the fully connected layer is $y_j \in (y_1, y_2, \ldots, y_n)$, and the output value after passing through the Softmax function is $o_i \in (o_1, o_2, \ldots, o_n)$.

$$o_i = \frac{e^{y_i}}{\sum_j e^{y_i}} \quad (9)$$

2.2 Millimeter wave radar ranging and speed measurement principle

The radar transmit signal is：

$$\mathrm{S_T}(t,\mathrm{n}) = A_T \exp(j2\pi((f_c - 0.5B)t + \int_0^t \frac{B}{T}\tau \mathrm{d}\tau) + j\varphi_T) \quad (0 \le t < T, 0 \le n < N) \quad (10)$$

Where A_T is the amplitude of the radar transmission signal, f_c is the radar carrier frequency, B is the maximum operating bandwidth of the radar, T is the frequency modulation period of the signal, φ_T is the initial phase of the transmission signal, t is the time on the fast time axis, N is the frame chirps signal repeatedly transmitted by the radar transmitting antenna , n is the chirp signal index of each frame.

After the transmitted signal is reflected from the target object, it is ` by the receiving antenna. The phase of the received signal is:

$$\varphi_{\mathrm{R}}(t,n) = (f_c - 0.5B + f_D)(t - \Delta\mathrm{t}) + \frac{B}{2T}(t - \Delta\mathrm{t})^2 \quad (11)$$

$$f_D = \frac{2 f_c v}{c} \quad (12)$$

$$\Delta\mathrm{t} = \frac{2d}{c} \qquad (\Delta\mathrm{t} \le t < T, 0 \le n < N) \quad (13)$$

There are two phase differences in the received signal due to the movement of the target object. The first is the Doppler frequency shift f_D, v is the relative speed between the target and the radar, and c is the speed of light. Δt is the time delay between transmitting the signal

and receiving the signal, and d is the straight-line distance between the radar and the target.

The mixer is responsible for multiplying the received signal and the transmitted signal, and then passing the low-pass filter to obtain the intermediate frequency signal:

$$X = [x(1), x(2), \ldots, x(n)] \tag{14}$$

$$x(n) = [x(1,n), \ldots, x(m,n)]^T \tag{15}$$

$$x(m,n) = A\exp(j2p(\frac{2f_c d}{c} + \frac{2f_c v}{c}Tn + \frac{2Bd}{Tc}\frac{T}{M}m) + j\Phi_s) \tag{16}$$

Where X is the intermediate frequency signal matrix output by the mixer, $x(m,n)$ is the intermediate frequency signal in m rows and n columns in the matrix, A is the amplitude of the intermediate frequency signal, and Φ_s is the initial phase.

By performing two Fourier transforms on the matrix X, the distance and speed information of the target can be estimated. Column-level Fourier transform is performed first, followed by row-level Fourier transform. When the column-level Fourier transform is performed on the matrix X, a peak will appear at a specific frequency, corresponding to $2Bd/Tc$ in equation (16). The frequency at this time represents the distance between the millimeter wave radar and the target. The corresponding frequency response is:

$$X_d = [x_d(1), x_d(2), \ldots, x_d(N)] \tag{17}$$

Perform Fourier transform on the row vector of the X_d matrix to obtain the relative velocity between the target and the millimeter-wave radar:

$$X_v = [x_v(1), x_v(2), \ldots, x_v(M)] \tag{18}$$

3. Two-stream Feature Fusion Algorithm

Existing radar recognition algorithms can be divided into two categories. One type is a method based on single-dimensional information extraction, such as extracting the distance-Doppler map of human movement and inputting it into a neural network for recognition and classification. However, a single feature has major flaws in human action recognition. A neural network using only a single feature cannot identify two actions with similar features, such as walking left and right, walking forward and walking backward, etc. The other method is to extract multi-dimensional features of human actions. This method extracts different dimensional features of human actions through different networks, and stacks the output of each network to enter subsequent processing and prediction steps. Compared with a single convolutional neural network, a multi-branch network can process various human body information and has better recognition capabilities. However, its multi-branch network structure also greatly increases network parameters and calculation volume, which is not conducive to network training.

The overall architecture of the algorithm is shown in Figure 1. Based on the VGG16 network, a spatial pyramid pooling algorithm(SPP) is added to better extract the Doppler features of human body movements. In addition, in the ResNet50 network, the use of channel attention mechanism(SENet) can more effectively extract the distance features of human actions. In order to better fuse the two features of human movement distance and Doppler frequency extracted by the feature extraction network, a self-attention mechanism is introduced and a weighted method is used for fusion. Finally, the SoftMax classifier is used to identify and classify human actions.

3.1 Doppler Feature Extraction

On the basis of the VGG16 network, the SSP is added to better extract the Doppler features of human body movements .

VGG Net is an improvement based on the AlexNet network model, which is expanded in depth and width. This is because deep networks have stronger expressive capabilities than shallow networks, and the deeper the network, the stronger the learning ability [16]. The VGG16 network is a 16-layer network composed of 5 convolutional blocks and 3 fully connected layers. Among them, block1 and block2 contain two convolutional layers, while block3, block4 and block5 contain three convolutional layers. Each convolutional layer uses an activation function and is followed by a max pooling layer for dimensionality reduction. Finally, the output vector of the fully connected layer is passed to the Softmax layer for classification.

SPP is often added to convolutional neural networks for algorithm improvement. SPP uses the multi-scale information of the pooling area to map input images of different sizes to fixed feature dimensions, thus avoiding scaling of the input image and reducing the loss of spatial information of the input image, which can effectively improve the feature expression ability of the image.

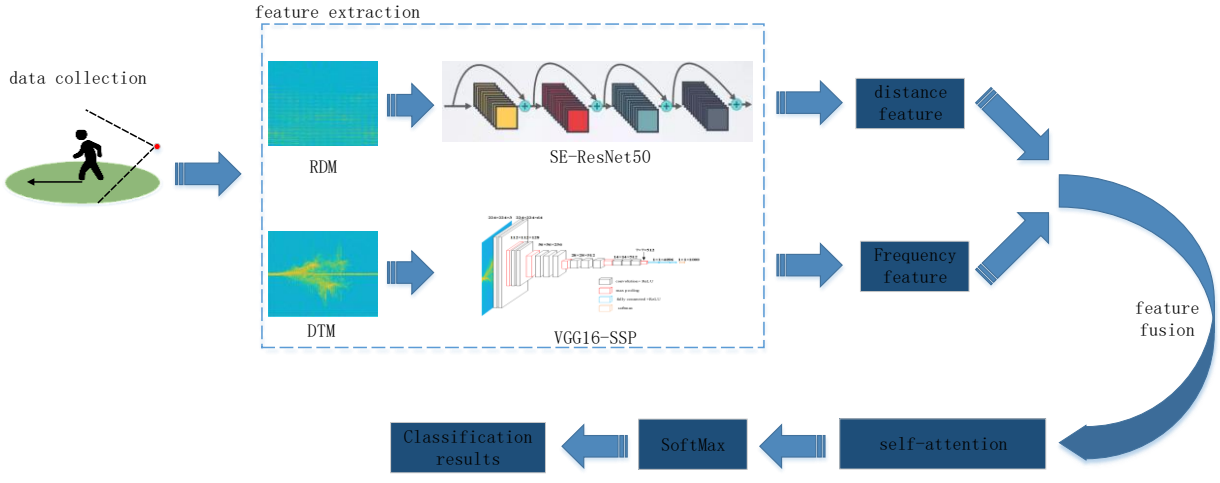


Fig.1. The Overall Architecture of the Algorithm.

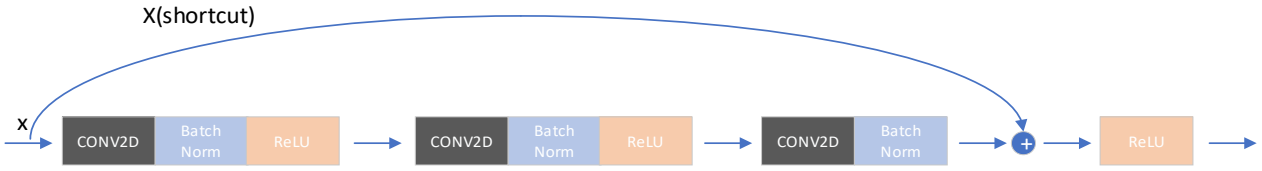


Fig.2. Identity block.

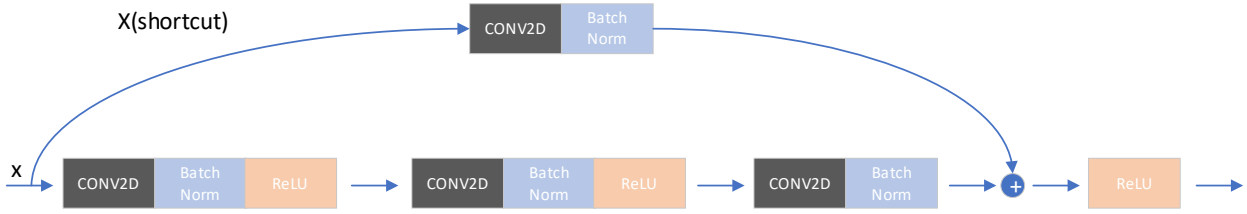


Fig.3. Conv block.

In this algorithm, the pooling size and stride are flexibly set based on the feature map obtained from the last convolutional layer. For example, assuming that the size of the last layer of convolution output feature map is $m \times m$, then the number of layers is $n = 3$ spatial pyramid pooling, and the calculation of each pooling window size w and step size t is as follows:

$$\begin{cases} w_1 = ceil(m/n), & t_1 = floor(m/n), & n = 1 \\ w_2 = ceil(m/n), & t_2 = floor(m/n), & n = 2 \\ w_3 = ceil(m/n), & t_3 = floor(m/n), & n = 3 \end{cases} \quad (19)$$

Among them, *ceil* means rounding up, and *floor* means rounding down.

In order to effectively extract the feature information of the Doppler map, SPP is added to the VGG16 network for improvement. While maintaining the same number of layers of the traditional VGG16 network, SPP is used to replace the final maximum pooling layer. This facilitates information "aggregation" in the deep stages of the network, while avoiding scaling of the input image, thereby reducing the loss of spatial information in the image.

3.2 Distance feature extraction

A large amount of research has shown that although the deeper the network depth, the more information can be obtained, which may improve the model optimization effect, but when the network depth reaches a certain level, the error rate of the model will increase. This is not due to overfitting, but because deepening the network depth will lead to gradient explosion and gradient disappearance. Therefore, the deeper the network, the more difficult it is to optimize the model, rather than being unable to learn more features. In order to deal with the above-mentioned "network degradation" problem and enable the deep network to be better trained, He Kaiming and others proposed the deep residual network (ResNet) [17].

ResNet includes two types of blocks, one of which is the Identity block, as shown in Figure 2. This block does not change the dimensions of the input, but instead learns the input residual and then adds it to the input to produce the output such that the dimensions of the input and output are consistent.

The other is the Conv block, as shown in Figure 3. This block also learns the residual part after convolution mapping, and finally adds the output of the convolution part and the output of the residual part to get the final output. At this time, the input Different from the output dimensions.

In order to enhance the deep feature extraction and expression capabilities of the ResNet50 network, a channel attention mechanism [18] is used to enhance the feature channels related to the classification task while suppressing the feature channels irrelevant to the task. The network structure is shown in Figure 4.

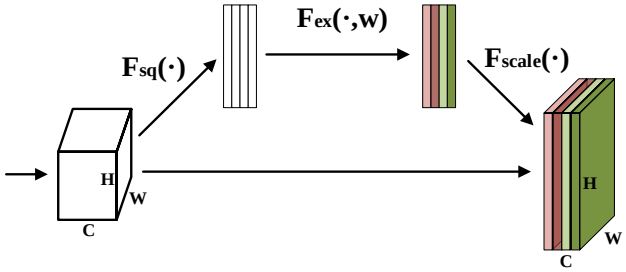


Fig.4. Network Structure of Channel Attention Mechanism.

3.3 Feature fusion

In order to fuse the features extracted by VGG16 and ResNet50, it is necessary to delete the last two fully connected layers and Softmax classifier of the VGG16 network, and delete the Softmax classifier of the ResNet50 network. In this way, the Doppler features and range-time features extracted by the two networks can be obtained.

The operation process of the feature fusion module is as follows. Assume that the input feature map is as follows:

$$\boldsymbol{F} \in \mathbb{R}^{H \times W \times C} \quad (20)$$

Among them, H and W represent the height and width of the feature map respectively, and C represents the number of channels of the feature map. Streamline the data from each channel for subsequent processing.

$$\boldsymbol{F}_c \in \mathbb{R}^{N \times C} \quad (21)$$

Among them, N represents the number of pixels of the feature map, that is $N = H \times W$. Flatten the channel data of the feature map to obtain a matrix F_C with a dimension of $N \times C$. Then perform matrix multiplication on F_C and its transpose to obtain the inter-channel correlation matrix G .

$$\boldsymbol{G} = \boldsymbol{F}_c^T \cdot \boldsymbol{F}_c \in \mathbb{R}^{C\times C} \tag{22}$$

Perform the Softmax operation on each column of the matrix G. Assume that G_{ij} represents the element of the i-th row and j-th column of G, then the matrix $\mathbf{M} \in \mathbb{R}^{C\times C}$ obtained after the operation.

$$M_{ij} = \frac{exp\left(G_{ij}\right)}{\sum_{j=1}^{C} exp\left(G_{ij}\right)} \tag{23}$$

Apply it to the feature matrix to obtain the attention spectrum:

$$\boldsymbol{F}_{c1} = \boldsymbol{F}_c \cdot \boldsymbol{M}^T \tag{24}$$

Restore the fused attention spectrum to its original size:

$$\boldsymbol{F}_{c1} \in \mathbb{R}^{N\times C} \rightarrow \boldsymbol{F}_{c2} \in \mathbb{R}^{H\times W\times C} \tag{25}$$

In order to better extract features under the attention mechanism, a common method is to fuse the attention spectrum and the original feature map. However, when performing fusion, there may be problems in directly multiplying the two. It may make the value of the original feature map too small, making it difficult for the neural network to learn later, and the output feature map will rely too much on the information of the attention spectrum. Therefore, a more appropriate way is to perform fusion through weighted summation. Fusion is performed by weighted summation:

$$\boldsymbol{F}_f = \alpha \cdot \boldsymbol{F}_{c2} + \boldsymbol{F} \tag{26}$$

4. Experiments

First, the algorithm proposed in this paper and the comparison algorithm are trained and tested using a data set without co-frequency interference noise. Next, a data set containing co- frequency interference noise will be used to verify the robustness of the algorithm. In order to verify the feasibility of the anti-interference recognition algorithm for two-stream feature fusion proposed in this article, it will be compared with typical algorithms.

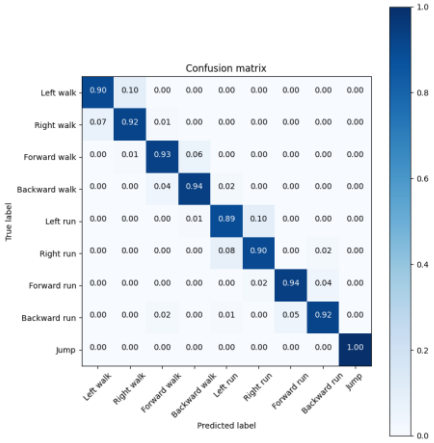


Fig.5. VGG16-SSP Algorithm Confusion Matrix.

4.1 . Results on data sets without co-channel interference

Use the Doppler map data set to train and test the VGG16-SSP network algorithm. The confusion matrix of the VGG16-SSP network algorithm is shown in Figure 5. It can be found that the recognition accuracy of the VGG16 recognition algorithm based on SSP is 92%. However, it can also be seen that the VGG16-SSP single-branch algorithm cannot easily distinguish movements in the left and right directions.

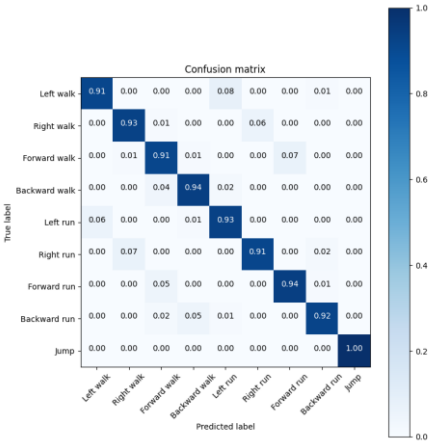
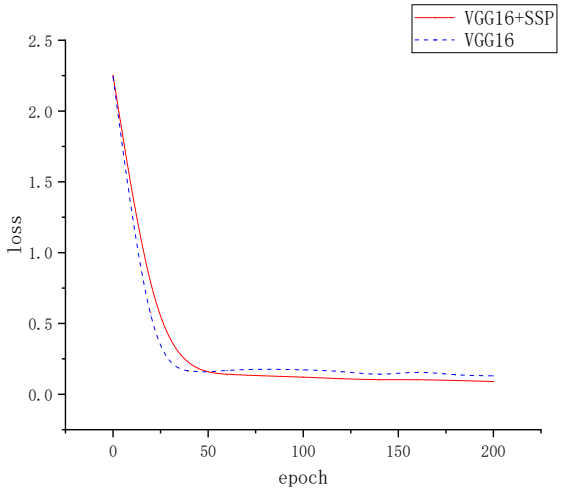
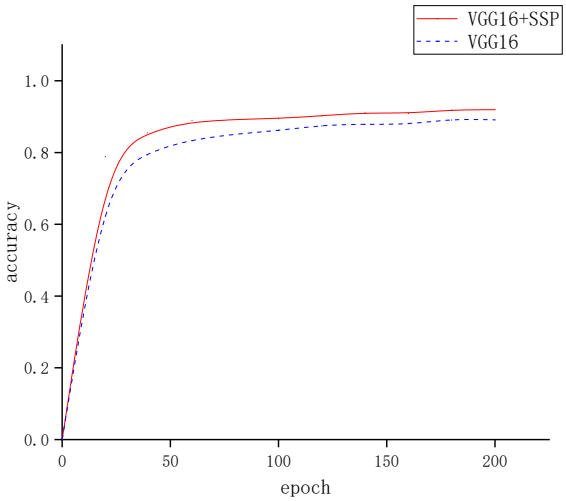


Fig.6. SE-ResNet50 Algorithm Confusion Matrix.



(a) Loss changes of two algorithms.



(b) Accuracy changes of two algorithms

Fig.7. VGG16+SPP algorithm loss and accuracy.

Then, the distance-time map data set is used to train and test the SE-ResNet50 network algorithm. The confusion matrix of the SE-ResNet50 network algorithm is shown in Figure 6. The ResNet recognition algorithm based on SENet designed in this paper has a recognition accuracy of 93%. Compared with the VGG16-SSP

network algorithm that recognizes Doppler features, it can identify human bodies more effectively. However, it can also be seen that the SE-ResNet50 single-branch algorithm cannot easily distinguish between running and walking.

In order to verify the beneficial effect of the addition of the SSP module on reducing the loss value and improving the accuracy of the VGG16 algorithm, the SSP module in the algorithm was removed to form a control experimental group. The comparison of loss values and accuracy during the training process of the two algorithms is shown in Figure 7. It can be seen that the introduction of the SSP module can improve the receptive field of the model and better capture the global information in the image, thus improving the performance of the model.

After training the above two network algorithms, VGG16-SPP and SE-ResNet50, migration learning is used to form the anti-interference algorithm for two-stream feature fusion proposed in this paper. The confusion matrix of the anti-interference algorithm of two-stream feature fusion is shown in Figure 8. It can be seen that compared with the single-branch network algorithm, the algorithm proposed in this article has better recognition performance for human actions, and the algorithm recognition accuracy is 98.5%. At the same time, this algorithm is more accurate in terms of recognition of direction and action analogies.

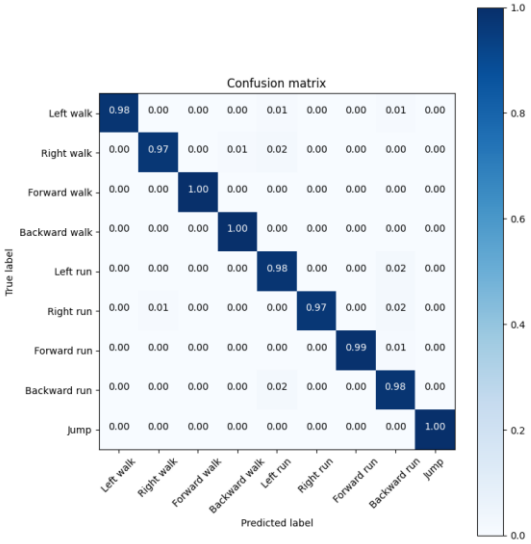
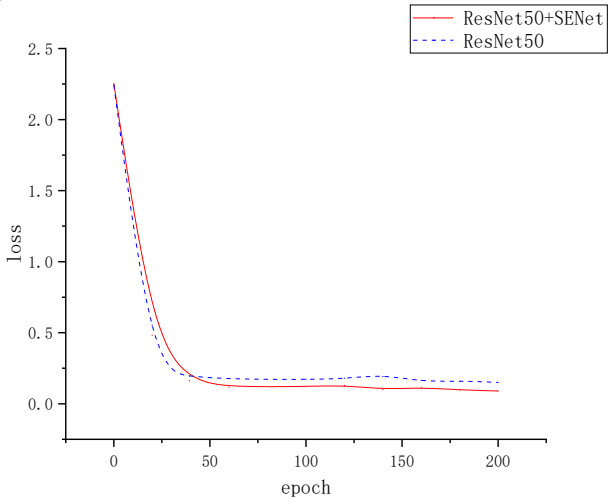


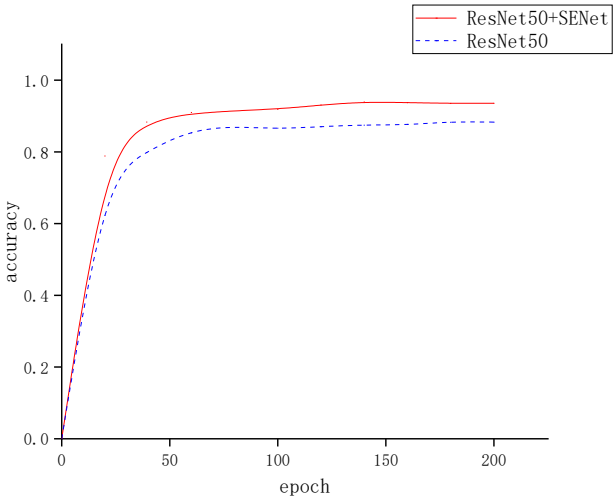
Fig.8. Two-stream Feature Fusion Algorithm Confusion Matrix.

In order to verify the beneficial effect of the SENet module on reducing the loss value and improving the accuracy of the ResNet50 algorithm, the SENet module was removed to form a control group. The comparison of LOSS values and accuracy during the training process of the two algorithms is shown in Figure 9. It can be seen that the introduction of SENet The module can enhance the model's attention to important features to learn the importance of different channels in the feature map, thereby improving the accuracy and robustness of the model.

In order to evaluate the effectiveness of the algorithm proposed in this paper, algorithm 1 and algorithm 2 proposed in literature [19] and literature [12] are used for comparison. Algorithm 1 is a method based on a single CNN. In the article, the Doppler time maps of nine types of human movements are directly processed; Algorithm 2 used two identical CNN networks to extract the features of the range-time map and the Doppler time map respectively. Then simply stacked and fused these features, and used a classifier for classification. For comparison, the three algorithms all use the same data set for training and testing. The recognition accuracy of the three algorithms for nine types of human actions is shown in table 1.



(a) Loss changes of two algorithms.



(b) Accuracy changes of two algorithms

Fig.9. ResNet50+SENet algorithm loss and accuracy..

As can be seen from the table, the relative accuracy of Algorithm 1 using a single branch CNN is relatively low. Compared with the algorithm proposed in this paper, Algorithm 1 only processes data through a single branch CNN, so it cannot fully extract millimeter wave radar human action information, resulting in a decrease in classification accuracy. Multi-branch Algorithm 2 inputs two types of features into the branch network separately for feature extraction processing. The network structure is more complex and can effectively utilize feature information. The accuracy is higher than Algorithm 1. Our algorithm uses single-branch network training optimization and forms a multi-branch network through transfer learning. The self-attention mechanism is introduced during feature fusion to make full use of Doppler and distance features. Has the highest accuracy in data sets without co-frequency interference noise.

Tab.1. Accuracy of three algorithms for nine categories of human actions.

movement	algorithm1	algorithm2	ours
Go left	93%	94%	98%
Go right	92%	92%	97%
Go forward	94%	93%	100%
Go back	94%	95%	100%
Run left	91%	97%	98%

Run right	92%	94%	97%
Run forward	93%	95%	99%
Run back	92%	96%	98%
Jump	95%	99%	100%

4.2 . Results on co-frequency interference data set

In order to verify the beneficial effect of adding the SSP module to complex electromagnetic signals to improve the robustness of the VGG16 algorithm, the SSP module was removed from the VGG16 algorithm and the two algorithms were trained using data without co-frequency interference. After training, data containing co- frequency interference are used for testing to form a control experimental group. The accuracy of the two algorithms under different signal-to-noise ratios(SIR) is shown in Figure 10. The results show that as the SIR decreases, the accuracy of both algorithms decreases. However, the decrease in accuracy of the VGG16 algorithm with the SSP module added is much smaller than that of the VGG16 algorithm without the SSP module. This shows that adding the SSP module can improve the robustness of the VGG16 algorithm in complex electromagnetic signals.

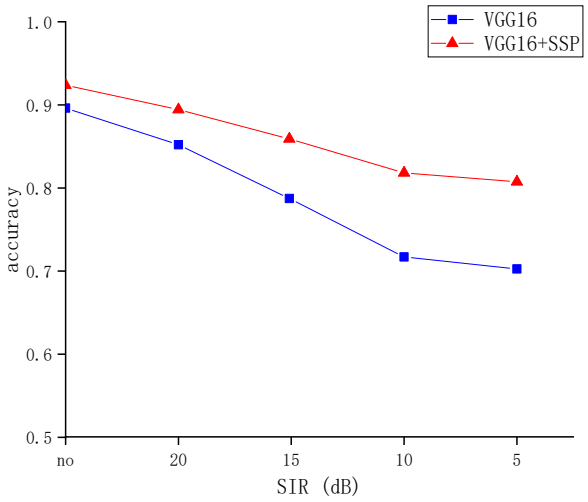


Fig.10. Accuracy of three algorithms under different SIR interferences.

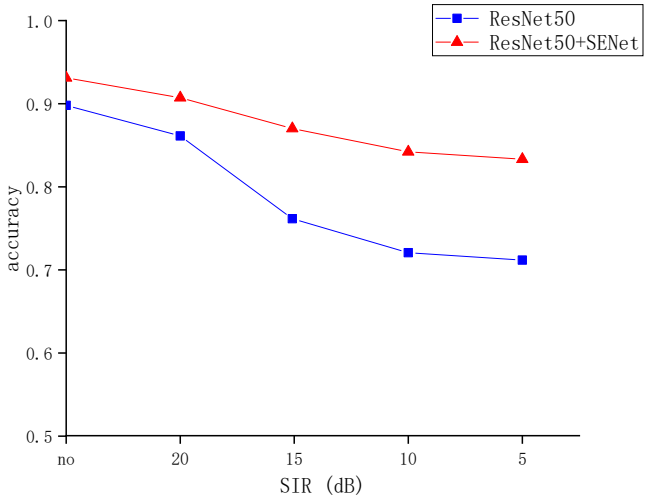


Fig.11. Accuracy of three algorithms under different SIR interferences.

In order to verify the role of the SENet module in the ResNet50 algorithm in complex electromagnetic signals, the SENet module was removed from the ResNet50 algorithm to form a control group, and the accuracy of the two algorithms under different SIR was compared. The results are shown in Figure 11. The accuracy drop of the ResNet50 algorithm with the SENet module added is much smaller than that of the ResNet50 algorithm without the SENet module. This shows that the introduction of the SENet module can effectively learn the importance of different channels in the feature map. This improves the robustness of the ResNet50 algorithm in complex electromagnetic signals. This result shows that the introduction of the SENet module plays an important role in improving the performance of the ResNet50 algorithm in practical applications.

In order to evaluate the effectiveness of the two-stream feature fusion algorithm in complex electromagnetic signals, the above-mentioned Algorithm 1 and Algorithm 2 are used as comparison algorithms, and different degrees of co-frequency interference data are used as verification sets for verification, as shown in Figure 12.

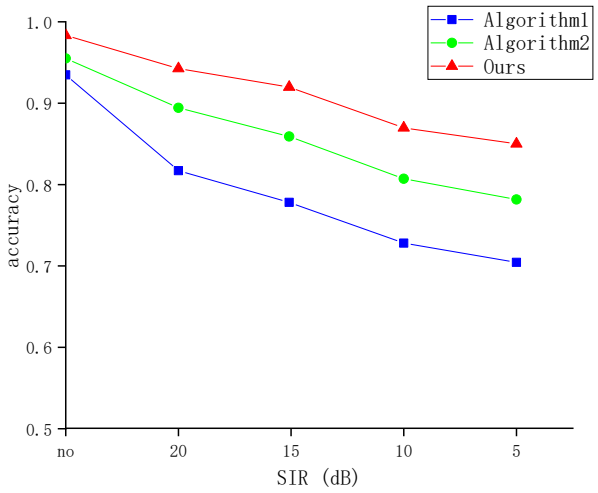


Fig.12. Accuracy of three algorithms under different SIR interferences.

The single-branch algorithm 1 is the most sensitive to co-frequency interference. As the degree of interference increases, the accuracy drops significantly, from 93% to 63%. Multi-branch algorithm 2 is less affected by interference than algorithm 1, but the accuracy still drops from 95% to 75%. The two-stream feature fusion anti-interference algorithm proposed in this article, after introducing the SPP module and SENet module, can more effectively extract human action features containing co-frequency interference noise signals. Our algorithm still guarantees high accuracy and robustness even in data sets containing co-frequency interference noise.

5. Summary

This paper proposes an anti-interference recognition algorithm for dual-stream feature fusion. It uses the spatial pyramid pooling algorithm and the channel attention mechanism to improve the VGG16 network and ResNet50 network respectively, improving the feature extraction capabilities of the network. Then a feature fusion module based on the self-attention mechanism is adopted to fuse the distance features and Doppler frequency features output by the feature extraction network to make the features more significant. Experimental results show that our algorithm has a high accuracy on both data sets without co-frequency interference noise and data sets containing co-frequency interference noise, which illustrates the effectiveness and robustness of this algorithm in scenarios with electromagnetic interference.

Acknowledgements

This research was funded by National Defense Basic Research Plan (grant number. JCKYS2020DC202), Natural Science Found

ation of Hebei Province (grant number. F2022208002), Science a nd Technology Project of Hebei Education Department (Key pro gram) (grant number. ZD2021048).

References

[1] A. MALVIYA, R. KALA. Social Robot Motion Planning Using Contextual Distances Observed from 3d Human Motion Tracking. Expert Systems with Applications, 2021, 184: 115515

[2] Y. CHEN, F. YANG, T. LANG, et al. Real-Time Street Human Motion Capture. arXiv preprint arXiv:2112.11543, 2021

[3] X. ZHOU. Wearable Health Monitoring System Based on Human Motion State Recognition. Computer Communications, 2020, 150: 62-71

[4] Z. WU. Human Motion Tracking Algorithm Based on Image Segmentation Algorithm and Kinect Depth Information. Mathematical Problems in Engineering, 2021, 2021: 1-10

[5] Q. WANG, K. WANG, W. CHEN. Clgnet: A New Network for Human Pose Estimation Using Commodity Millimeter Wave Radar. 2020 3rd International Conference on Algorithms, Computing and Artificial Intelligence, 2020: 1-5

[6] O. STEVEN EYOBU, D. S. HAN. Feature Representation and Data Augmentation for Human Activity Classification Based on Wearable Imu Sensor Data Using a Deep Lstm Neural Network. Sensors, 2018, 18(9): 2892

[7] S. MOHAMMADI, S. G. MAJELAN, S. B. SHOKOUHI. Ensembles of Deep Neural Networks for Action Recognition in Still Images. 2019 9th International Conference on Computer and Knowledge Engineering (ICCKE), 2019: 315-318

[8] T. YANG, J. CAO, Y. GUO. Placement Selection of Millimeter Wave Fmcw Radar for Indoor Fall Detection. 2018 IEEE MTT-S International Wireless Symposium (IWS), 2018: 1-3

[9] R. ZHAO, X. MA, X. LIU, et al. Continuous Human Motion Recognition Using Micro-Doppler Signatures in the Scenario with Micro Motion Interference. IEEE Sensors Journal, 2020, 21(4): 5022-5034

[10] C. DING, H. HONG, Y. ZOU, et al. Continuous Human Motion Recognition with a Dynamic Range-Doppler Trajectory Method Based on Fmcw Radar. IEEE Transactions on Geoscience and Remote Sensing, 2019, 57(9): 6821-6831

[11] F. SAKAGAMI, H. YAMADA, S. MURAMATSU. Accuracy Improvement of Human Motion Recognition with Mw-Fmcw Radar Using Cnn. 2020 International Symposium on Antennas and Propagation (ISAP), Osaka, Japan, IEEE, 2021: 173-174

[12] M. AOYAGI, S. WATANABE. Resolution of Singularities and the Generalization Error with Bayesian Estimation for Layered Neural Network. IEICE Trans, 2005, 88(10): 2112-2124

[13] J. FABER, G. A. GIRALDI. Quantum Models for Artificial Neural Networks. Electronically available: http://arquivosweb. lncc. br/pdfs/QNN-Review. pdf, 2002, 5(7.2): 5-7

[14] C. SZEGEDY, W. LIU, Y. JIA, et al. Going Deeper with Convolutions. Proceedings of the IEEE conference on computer vision and pattern recognition, 2015: 1-9

[15] HAOLI. Traffic Classification Algorithm Using Cnn and Multi-Head Attention Mechanism for Representation Learning. IOP Publishing Ltd

[16] K. HE, X. ZHANG, S. REN, et al. Deep Residual Learning for Image Recognition. Proceedings of the IEEE conference on computer vision and pattern recognition, 2016: 770-778

[17] Z. WU, N. MA, Y. GAO, et al. Attention Mechanism Based on Improved Spatial-Temporal Convolutional Neural Networks for Traffic Police Gesture Recognition. International Journal of Pattern Recognition and Artificial Intelligence, 2022, 36(08): 2256001

[18] R. ZHANG, S. CAO. Real-Time Human Motion Behavior Detection Via Cnn Using Mmwave Radar. IEEE Sensors Letters, 2018, 3(2): 1-4

Yongqiang Zhang born in 1981, Associate Professor and a master supervisor, received the Ph.D. degree in weapon system operation engineering from the Army Engineering University of PLA, Shijiazhuang, China, in 2023. His research interests include artificial intelligence, and Internet of Things technology.

Ziqiang Zhang born in 1998, studied in Hebei University of Science and Technology with a major in computer technology. The main research directions are: artificial intelligence.

Xiaohan Sun born in 1997, graduated from Hebei University of Science and Technology with a major in computer technology. The main research directions are: artificial intelligence.

Jinlong Ma born in 1981, Associate Professor, received the Ph.D. degree in information and communication engineering from the Harbin Institute of Technology. His research interests include information spreading dynamics in complex networks and data analysis of online social networks.

Weidong Wu born in 1973, Lecturer, and a master. Main research areas: embedded development, IoT.