(Received 16 January 2024, Accepted 12 April 2024, Scheduled 22 April 2024)

Human Motion Recognition Based on Feature Fusion and Transfer Learning

Xiaoyu Luo[1, 2] and Qiusheng Li[1, 2, *]

[1] *Research Center of Intelligent Control Engineering Technology, Gannan Normal University, Ganzhou 341000, Jiangxi, China*
[2] *School of Physics and Electronic Information, Gannan Normal University, Ganzhou 341000, Jiangxi, China*

ABSTRACT: In order to solve the problem that the recognition accuracy of human motion is not high when a single feature is used, a feature fusion human motion recognition method based on Frequency Modulated Continuous Wave (FMCW) radar is proposed. By preprocessing the FMCW radar echo data, the range and Doppler parameters of human motions are obtained, and the range-time feature map and Doppler-time feature map datasets are constructed. In order to fully extract and accurately identify the human motion features, the two features are fused, and then the two features maps and feature fusion spectrograms are put into the VGG16 network model based on transfer learning for identification and classification. Experimental results show that this method can effectively solve the problem of lack of information and recognition rate of single feature motion recognition, and the recognition accuracy is more than 1% higher than that of the single feature recognition method.

Comparison Table between Abbreviations and Full Names in English

Abbreviations	Full names in English
Bi-LSTM	Bidirectional Long Short-Term Memory Network
DTM	Doppler-Time Map
FFT	Fast Fourier Transform
FMCW	Frequency Modulated Continuous Wave
MTI	Moving Target Indication
ResNet	Residual Network
RTM	Range-Time Map
VGG-Net	Visual Geometry Group Network

1. INTRODUCTION

One of the most important technologies in the evolution of human-computer interaction is human motion recognition [1]. The research of human motion recognition has been widely used in the fields of public security, intelligent aging, human-computer interaction, etc. [2–4]. Although the vision-based posture technology has been developed and matured [5–8], it has certain limitations in building human motion recognition system which will bring the problems of privacy exposure and being affected by conditions such as light and occlusion [9]. The use of radio frequency sensors such as radar to extract information related to human posture from the wireless electromagnetic wave signals reflected from the human body makes up for the shortcomings of the vision-based human motion recognition method that is vulnerable to light and object sight occlusion, and simultaneously gives greater consideration to privacy protection, and it has currently gained popularity as a research area in the realm of human motion recognition [10].

In recent years, scholars have continued to use radar for human motion recognition, Ref. [11] describes the principles and methods of Frequency Modulated Continuous Wave (FMCW) radar target angle, velocity, and range estimation. Ref. [12] proposes a Bi-LSTM network structure for six human movements, processed FMCW radar data into time series containing micro-Doppler features and range features, and used two features to train the network model for the recognition of Doppler features and range features. Ref. [13] fuses micro-Doppler features and range-time features to propose an end-to-end convolutional neural network with dual-channel inputs and demonstrated that its recognition capability is higher than that of a single-input network. Ref. [14] obtains the range, Doppler, and angle multidimensional parameters of the gesture target by time-frequency analysis of the FMCW radar information, and used convolutional neural network and feature tandem fusion method for gesture recognition. Ref. [15] extends the 2-dimensional data time-range, time-Doppler and range-Doppler features of a millimetre wave radar by jointly expanding them into a 3-dimensional data model and then performing the recognition of human motion.

It is evident from the above study that millimetre wave radar-based human motion recognition has great advantages and has received widespread attention, and the obtained results are very significant. Radar picks up non-stationary signals because of the motion of the target, changes in the environment, and the dynamic nature of the radar system itself. A non-stationary signal is a signal whose statistical characteristics change within a certain time frame. Ref. [16] proposes a novel system for person identification by ultrasonic acquisition of hand posture information, which introduces five ways of processing

* Corresponding author: Qiusheng Li (liqiusheng@gnnu.edu.cn).

non-stationary signals: Short-time Fourier Transform (STFT), Wavelet Transform (WT), Stokewell Transform (ST), Hilbert-Huang Transform (HHT) and Constant Q Transform (CQT). In this paper, the Fourier transform is applied to the non-stationary signals acquired by the radar to obtain the range and velocity features of the human body movements, and feature fusion of these two features. The two-dimensional features and fused parameter maps are put into the VGG16 network based on the transfer learning for recognition and classification. Using the analysis of the measured data, it is demonstrated that the use of feature fusion can effectively improve the accuracy of human motion recognition in FMCW radar, and the generalization ability of the model is high.

2. THEORETICAL ANALYSIS

2.1. Range and Doppler Feature Extraction

FMCW radar, because of its wide frequency bandwidth to obtain high range resolution, can measure the range information and micro-Doppler information of human motions, and its transmitted wave can be expressed as

$$x_t(t) = \cos\left(2\pi f_0 t + \pi \frac{B}{T} t^2\right), \quad 0 \le t \le T_c \tag{1}$$

where T_c represents the repetition period of a linear FM signal, B the frequency bandwidth, and f_0 the lowest frequency within the bandwidth. The incident wave at the receiver is the sum of the delayed attenuation of the transmitted wave returned by the P scatterers, which can be expressed as

$$x_r(t) = \sum_{i=1}^{P} a_i \cos\left[2\pi f_0(t-\tau_i) + \pi \frac{B}{T}(t-\tau_i)^2\right] \tag{2}$$

where i denotes the serial number of the scatterer. Since the ratio of the frequency point f_b corresponding to the FMCW radar target to the delay τ_i of the return signal at that range is equal to the ratio of the signal bandwidth B to the duration of T_c, which is expressed as $f_b/\tau_i = B/T_c$, according to the basic principle of radar, the delay τ_i of the return time can be expressed in terms of the range R_i as $\tau_i = 2R_i/c$, and thus the range estimation can be expressed as $R_i = cT_a f_b/2B$, with a range resolution of $\Delta R = c/2B$.

The orthogonal mixing of $x_r(t)$ and $x_t(t)$ is performed, and the beat signal is obtained by low-pass filtering. The fundamental frequency of the mixing signal generated by the i-th scatterer is $f_b = \tau_i B/T_c$, and RTM can be obtained by using range-dimension fast Fourier transform (FFT) against the beat In-phase/Quadrature (I/Q) signal. Figure 1 shows the flowchart of human motion radar echo processing.

2.2. Construction of RTM and DTM Datasets

Since the chirp signal periods and sampling intervals in the radar are on the same time axis but in different time scales, a distinction can be made between the slow time axis and fast time axis. Each chirp signal period and sampling interval are

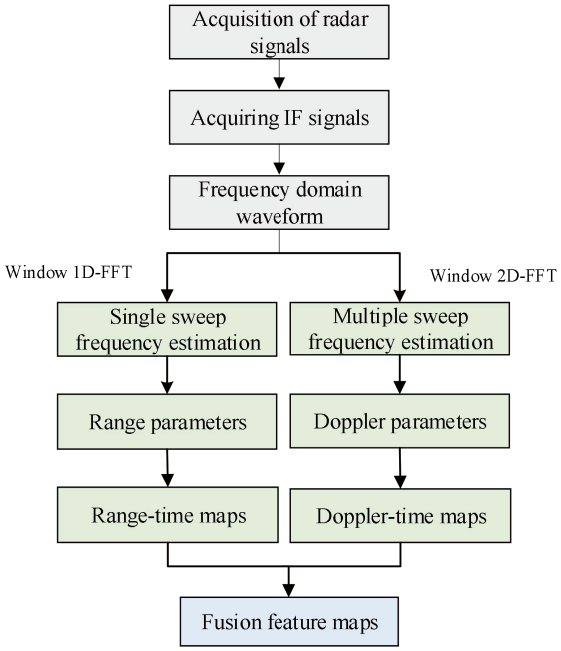

FIGURE 1. Flow chart of radar signal processing.

divided into two dimensions: the vertical axis is the slow time axis, which represents the different Chirp signals, and the horizontal axis is the fast time axis, which represents the sampling interval of each chirp signal. In actual radar testing work, the speed of human movement is far less than the radar scanning frequency. It can be considered that there is no change in range within a chirp signal period, and the change in range is reflected in the adjacent chirp signals. Therefore, the fast time contains the range information of human movement, and the slow time contains the micro-Doppler information of the human body.

2.2.1. Construction of RTM Datasets

According to the principle of radar ranging, the range information of human motion is stored in the pulse sequence, and 128 FFTs can be performed on 128 pulses in fast time dimension. Meanwhile, in order to prevent spectrum leakage, the Hamming window is added to each column of data, and the result is taken as the range information of the frame signal. The range information obtained is integrated into the time domain, and the range characteristic map of human motions in continuous time, namely RTM, can be obtained. Figure 2 is the RTM diagram of 6 motions.

2.2.2. Construction of DTM Datasets

Speed information describes the speed of the target in the process of movement, and the speed of movement is different for different motions. After one-dimensional FFT is performed on the original signal, the spectrum of each pulse is obtained, and then two-dimensional FFT is performed on multiple pulse echoes. Doppler information is estimated from the phase changes of the same frequency, and the range-Doppler spectrum is obtained. Spectrum peak search is performed and accumulated in the time domain, and the Doppler-time map of

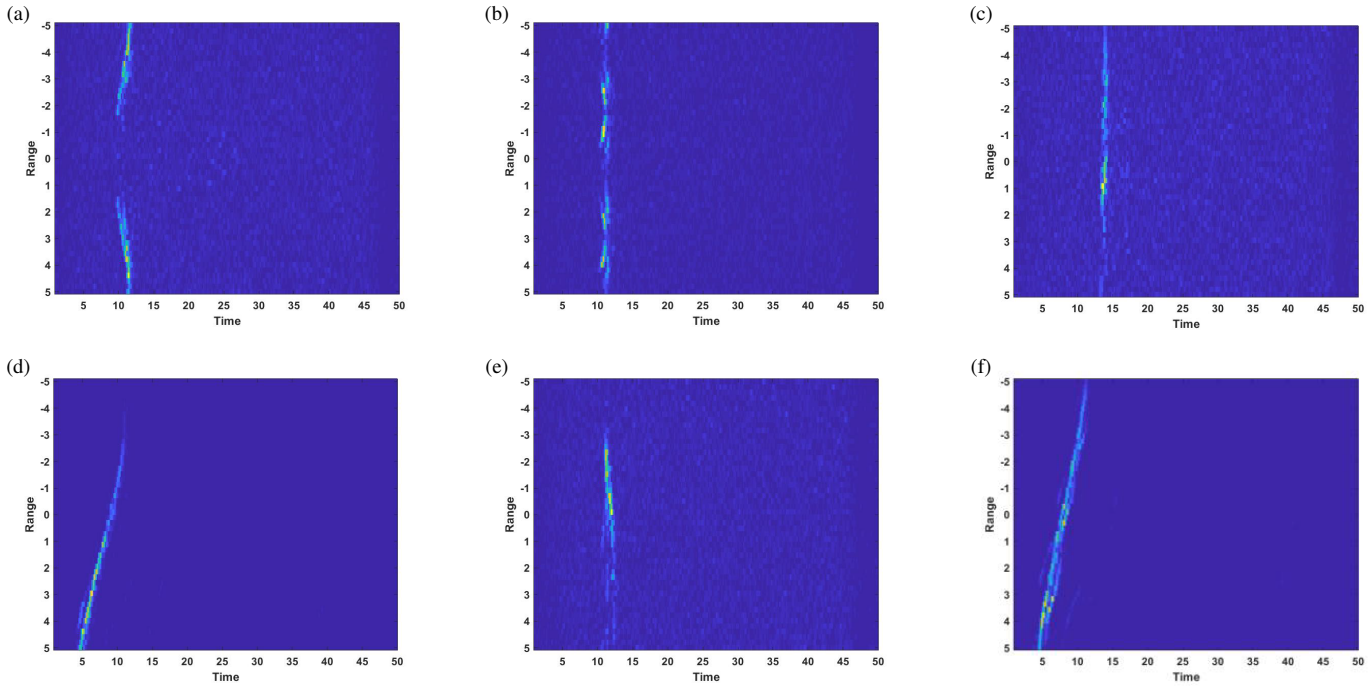

FIGURE 2. Schematic diagram of RTM for different motions, (a) bend, (b) clap, (c) jump, (d) run, (e) squat, (f) walk.

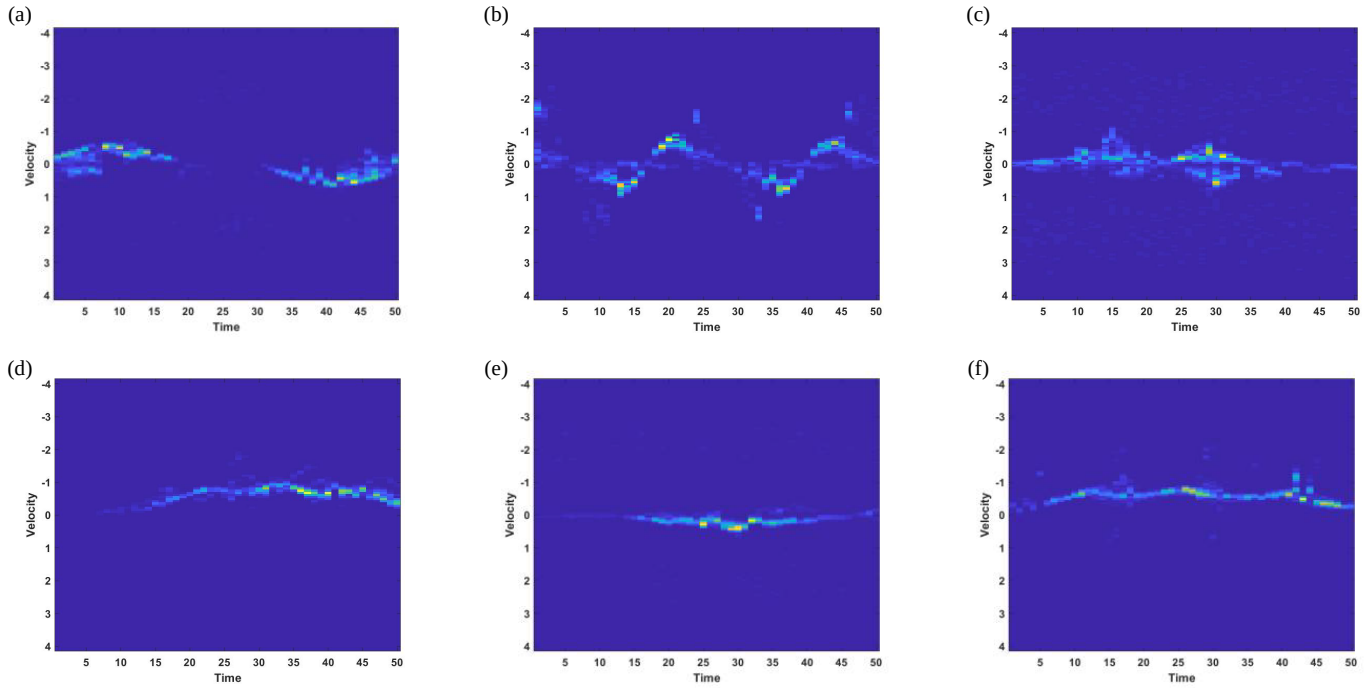

FIGURE 3. Schematic diagram of DTM for different motions, (a) bend, (b) clap, (c) jump, (d) run, (e) squat, (f) walk.

human motions, namely DTM, is obtained. Figure 3 is the DTM diagram of 6 motions.

2.2.3. Static Clutter Filtering

When the radar transmits signals, other objects in the external environment will also generate radar echoes, which will interfere with the detection of human motion. In the FMCW radar system, clutter concentrates at zero frequency. Therefore, this paper adopts Moving Target Indication (MTI) filter clutter suppression processing. MTI filter uses the difference between the Doppler frequency of clutter and moving targets to make the frequency response of the filter have a deep stopband at the integer multiple of DC and PRF (pulse repetition frequency), while the suppression at other frequency points is weak, so as to suppress the static target and still object clutter through a deep

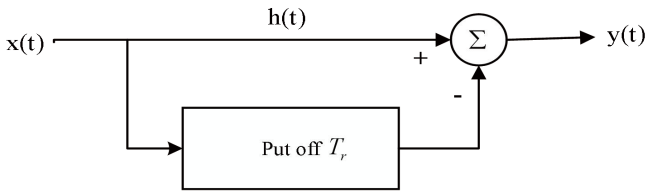

FIGURE 4. Single delay line pair canceller.

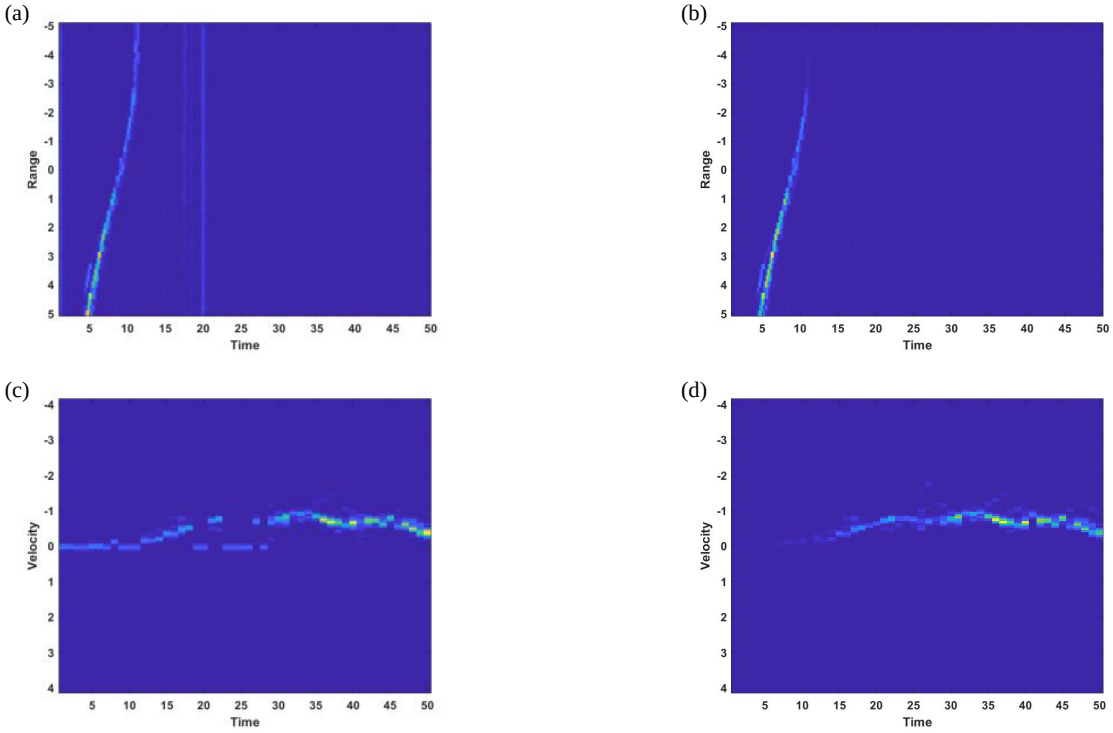

FIGURE 5. MTI filter input and output comparison result. (a) RTM before MTI filter input, (b) RTM after MTI filter input, (c) DTM before MTI filter input, (d) DTM after MTI filter input.

"notch". In the experiment, a two-pulse canceller is usually used, also known as a one-time canceller, and its filter structure is shown in Figure 4 below, where $x(t)$ is the input signal, $y(t)$ the output signal, $h(t)$ the system shock response, and T_r the pulse interval frequency. Figure 5 illustrates the comparison of RTM and DTM of human running motion before and after passing through the MTI filter.

2.3. Feature Map Preprocessing

After obtaining the range-time and Doppler-time graphs of human movements, the feature spectra of individual movements are not obvious enough, and the numerical differences are large, making it difficult for the training of convolutional neural networks to converge. Therefore, the feature maps are normalized and fused.

2.3.1. Normalization

Normalization serves to convert data with different magnitudes into a uniform range of criteria, speeding up the convergence of the training network. Firstly, each feature spectrum is discretized and standardized, and the range-time feature graph is taken as an example to scale it numerically according to the formula:

$$\bar{r}_{m,n} = \frac{r_{m,n} - \left[\sum_{x=1}^{X}\sum_{y=1}^{Y} r_{x,y}/(X \cdot Y)\right]}{\max\{R\} - \min\{R\}} \tag{3}$$

where $r_{m,n}$ represents the initial image's pixel value; $\bar{r}_{m,n}$ represents the value of image pixel after normalization; X, Y are the row and column of the image.

2.3.2. Feature Map Fusion

After the normalization of the feature maps, the pixel values of all feature maps are in the same interval, which reduces the internal differences of each feature map. Then the image fusion method based on local energy features and Laplacian pyramid is used to fuse the features of range-time maps and Doppler-time maps.

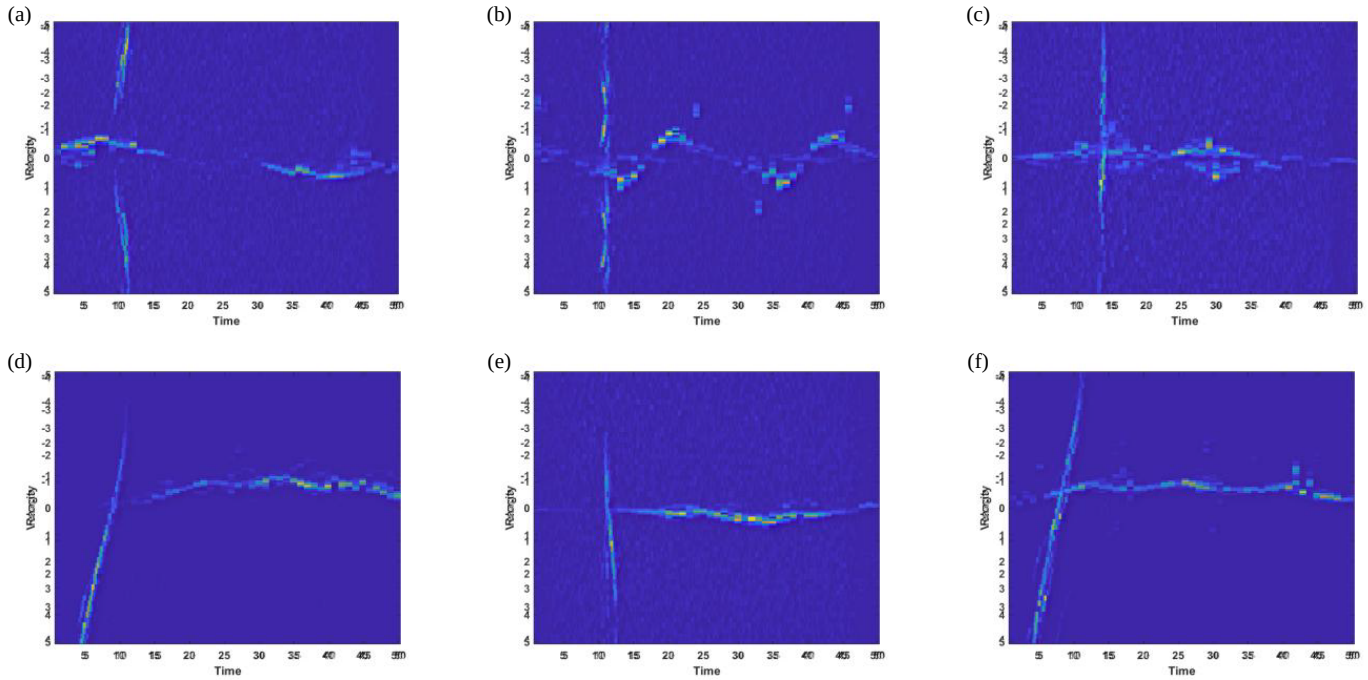

FIGURE 6. Feature fusion maps for different motions, (a) bend, (b) clap, (c) jump, (d) run, (e) squat, (f) walk.

Local energy features are defined as:

$$s(a,b) = \sum_x \sum_y C(a+x, b+y)^2 \tag{4}$$

where a and b represent the position coordinates of the image; x and y represent the window positions; and $C(a+x, b+y)$ is the pixel value of the image.

The steps of the local energy feature algorithm are:

(1) Select a threshold e.

(2) Calculate local energy maps for multi-source images.

(3) Calculate match degree

$$M_{IJ}(a,b) = \frac{\left(\sum_x \sum_y C_I(a+x, b+y) \cdot C_J(a+x, b+y)\right)^2}{S_I(a,b) \cdot S_J(a,b)}$$

(4) If the matching degree of the point $M < e$, select the graph with high energy of the point and discard the others.

(5) If the matching degree of the point $M > e$, the energy size determines how much weight is allocated. The weight of the small energy is $W_{\min} = 0.5 * \left(1 - \frac{1-M_{IJ}}{1-e}\right)$, and the weight of the large energy is $W_{\max} = 1 - W_{\min}$.

Since the Laplacian pyramid can extract more detailed information at multiple scales, it can be used to extract details at multiple scales by combining the energy feature algorithm with the Laplacian pyramid to achieve better results. The specific approach is as follows: decompose the input image into the Laplacian pyramid to obtain two image pyramids. Then each layer of the two pyramids is fused with the local energy feature algorithm. Finally, reconstruct the original image from the pyramids.

This fusion method can bring out the obvious feature pixels in both images, improve the signal-to-noise ratio of the fused image. The implementation method is not complicated, and the fusion speed is fast. Compared with the original image, the contrast of the fused image will be reduced, but the features are more obvious, which can better express each motion. Figure 6 shows the feature fusion image of the six motions after processing.

3. TRANSFER LEARNING BASED ON VGG16 MODEL

3.1. VGG16 Model

VGG-Net is a deep convolutional neural network, developed by the Visual Geometry Group of Oxford University and Google DeepMind [17]. The VGG16 convolutional neural network model has 16 weight layers, including 13 convolution layers and 3 fully connected layers. The 13 convolution layers are segmented by a maximal pooling layer at layers 2, 4, 7, 10, and 13, respectively. A three-channel image with a resolution of 224×224 is input. After the convolution operation is completed, the input data batch is normalized, which serves to bring the value intervals closer to the limit saturation region after a nonlinear function mapping [18–22]. In the convolution layer, a 3×3 filter is used, with every 2 or 3 filters stacked consecutively to form a convolution sequence to mimic the effect of a larger sensory field, with a sliding step size of 1, and boundary padding is used to keep the data dimensions before and after invariant. In the pooling layer, a 2×2 pooling window is adopted, and the step size is set to 2, which is used to halve the size of the convolution feature image while retaining important feature information, the fully-connected layer is composed of three consecutive fully-connected combinations, with

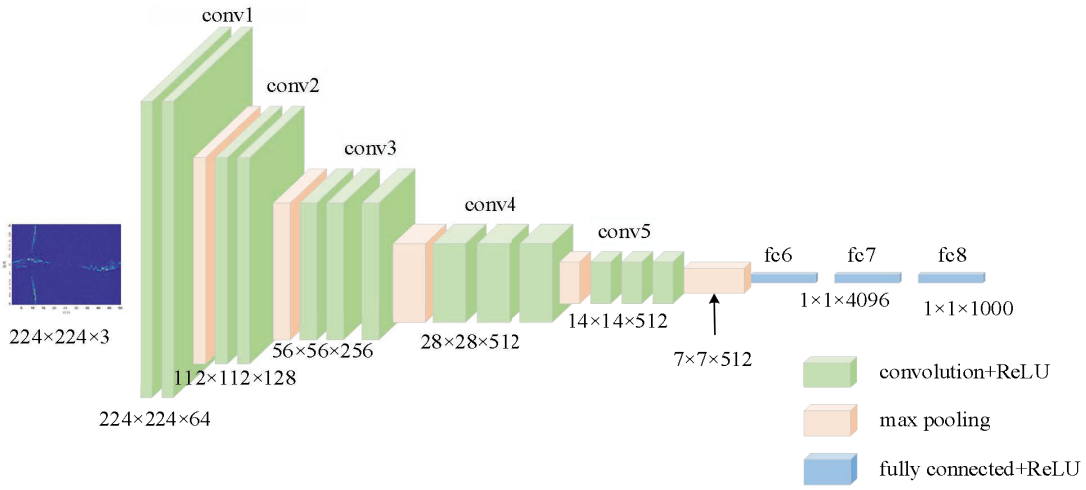

FIGURE 7. Structure of VGG16.

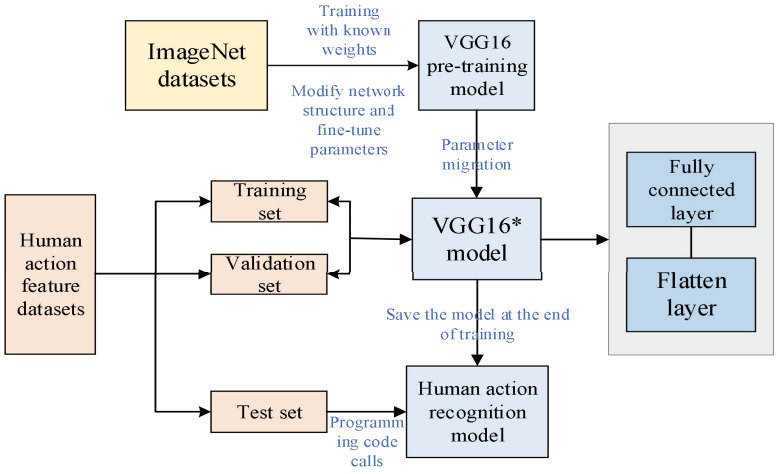

FIGURE 8. Schematic diagram of transfer learning process.

the number of channels of 4096, 4096, and 1000, respectively; and lastly, a 1,000-label Softmax classifier is used to classify the output. The network model of VGG16 is shown in Figure 7.

3.2. Transfer Learning

Transfer Learning [23] refers to improving the generalization of models by transferring models or knowledge trained in one domain to another domain or problem. Compared with the original model, its advantages are more obvious. The network structure using transfer learning is based on the original trained model, by fine-tuning the existing deep network to adapt to a specific task, to achieve the transfer of the model or parameters, and the process of transfer learning is shown in Figure 8. Denote the VGG16 network model after transfer learning as VGG16∗.

After transfer learning the network structure, the initial performance of the model is higher, the rate of model enhancement faster, and the convergence effect , more obvious. Firstly, the VGG16 network model is pre-trained using the ImageNet dataset, and the trained network model is saved. On this basis, the method of transfer learning is introduced to transfer the parameters of the pre-trained network, and the weight ratio of the trained network model is used to replace the original network weight random initialization operation. The last 3 layers of the model are replaced by a flatten layer and a new fully connected layer. The function of the flatten layer is to flatten the output of the convolution layer into a one-dimensional vector and change the output value of the fully connected layer to 6, so as to realize the recognition of human motions.

3.3. Hyper-Parameterisation

During deep learning training, the choice of hyper parameters affects the accuracy of the network after training. In this paper, when using the VGG16 model for training, the other initial parameter weights are adopted from the weight values trained on the ImageNet dataset. A training batch size of 32 and a maximum number of iterations (Epochs) of 100 are selected, and the network model parameters are optimized using a variant of stochastic gradient descent (SGD), the RMSprop optimizer, with the learning rate set to 2e-5.

FIGURE 9. RTM accuracy change curve.

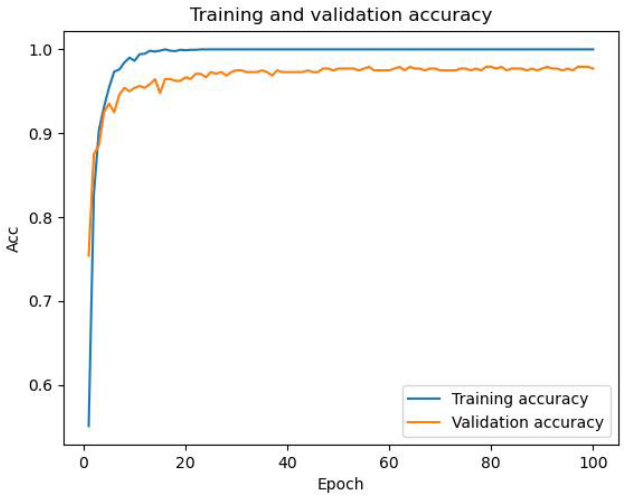

FIGURE 10. DTM accuracy change curve.

4. COMPARATIVE EXPERIMENTALS AND ANALYSIS

4.1. Data Collection

TI's IWR1642 development board and DCA1000 data acquisition card are used as the radar hardware platform to capture human motions. The IWR1642 device operates in the 77–81 GHz band and is an integrated single-chip FMCW radar sensor. The device has the advantages of low power consumption, high integration, and a maximum operating bandwidth of 4 GHz. DCA1000 data acquisition card can achieve real-time capture of radar data and real-time transmission of radar data through Ethernet transmission. During the experiment, the IWR1642 and DCA1000 were fixed on a tripod about 1 m from the ground, with no other moving targets interfering except for the experimental subject and no static objects placed in the straight-line range between the radar and the experimenter. The human target was 2.5 m away from the radar, and the parameters of the radar were set as shown in Table 1.

TABLE 1. Radar parameter settings.

Parameter	Numerical value
Bandwidths (GHz)	4
FM slope (MHz/μs)	64
Sampling points	256
Number of chirps	128
Frame rate	50
Sampling Rate (ksps)	5120

Eight subjects and six types of movements were tested: bending, clapping, jumping, running, squatting, and walking. The acquisition time for each motion was 2 s and repeated 50 times. The total amount of data collected was 2400, and the obtained RTM, DTM, and fusion feature maps were divided into training set, validation set, and test set in the ratio of $6:2:2$.

4.2. Evaluation Indicators and Performance Analysis

The designed model is evaluated using the confusion matrix and classification algorithm evaluation metrics, which include the following:

(1) Accuracy: The number of correctly categorized samples as a proportion of the total number of samples, defined as:

$$acc = \frac{P_{true}}{P_n} \tag{5}$$

(2) Precision: The proportion of samples predicted to be in the positive category that are actually positive, expressed as:

$$precision = \frac{T_P}{T_P + F_P} \tag{6}$$

(3) Recall: The proportion of samples that are actually in the positive category and predicted to be positive, expressed as:

$$recall = \frac{T_P}{T_P + F_N} \tag{7}$$

where P_{true} is the number of all correctly classified samples, P_n the total number of samples, T_P the number of positive classes predicted to be positive, F_P the number of negative classes predicted to be positive, and F_N the number of positive classes predicted to be negative.

The two feature maps that were extracted Range-Time Map (RTM) and Doppler-Time Map (DTM) were then added to the network for training and verification, respectively, in order to confirm the authenticity and accuracy of the extracted data. Figures 9 and 10 show the accuracy variation curves of RTM and DTM, respectively, and Figures 11 and 12 show the confusion matrices of RTM and DTM, respectively. Tables 2 and 3 show the evaluation metrics for RTM and DTM, respectively.

By analyzing the confusion matrix and evaluation index information of the above two single feature maps, it is evident

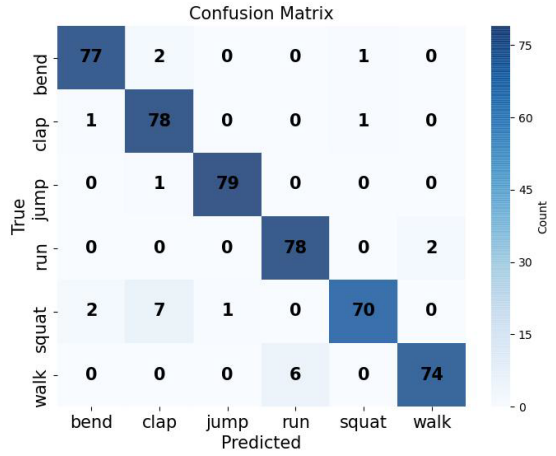

FIGURE 11. RTM confusion matrix.

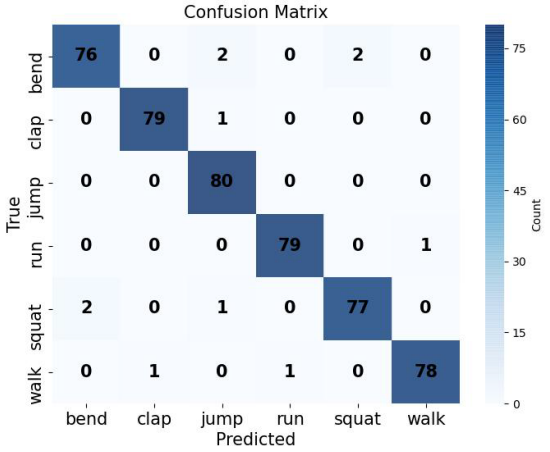

FIGURE 12. DTM confusion matrix.

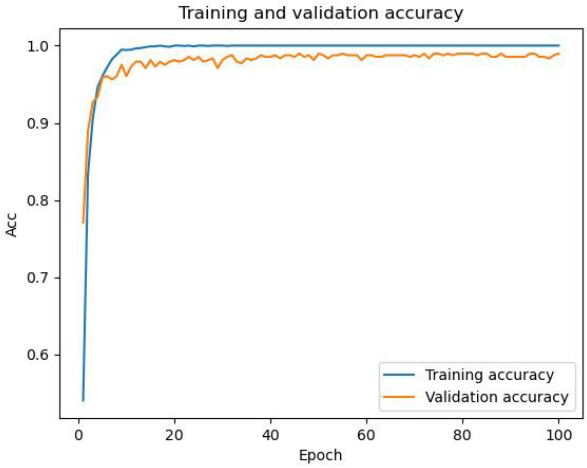

FIGURE 13. Accuracy curve of fusion feature map.

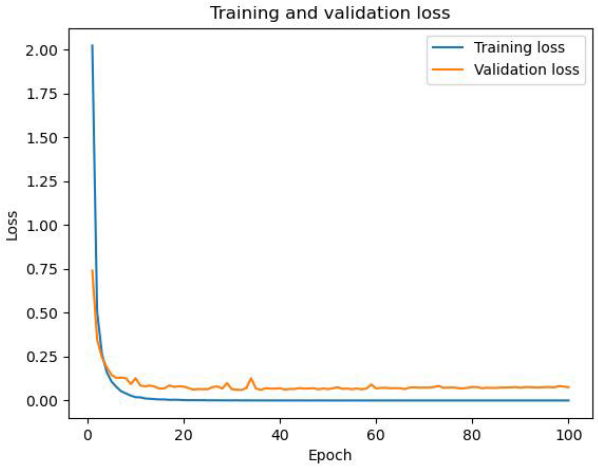

FIGURE 14. Change curve of loss value in fusion feature map.

TABLE 2. Evaluation indicators of RTM.

Motion category	Precision	Recall
bend	0.963	0.963
clap	0.886	0.975
jump	0.988	0.988
run	0.929	0.975
squat	0.972	0.875
walk	0.974	0.925

TABLE 3. Evaluation indicators of DTM.

Motion category	Precision	Recall
bend	0.974	0.95
clap	0.988	0.988
jump	0.952	1
run	0.988	0.988
squat	0.975	0.963
walk	0.987	0.975

that in the range-time confusion matrix, the recognition accuracy of squatting and walking movements are both low. Squatting is easy to be misjudged as clapping, and walking is easy to be misjudged as running. In the Doppler-time confusion matrix, the recognition accuracy of squatting and bending movements is also low, and there are also misjudgments. Overall, it seems that the Doppler-temporal features have a higher correct recognition rate for each motion than the range-time features, so it is necessary to feature-fuse the two feature maps to show more feature information to improve the overall recognition accuracy. To attain increased recognition precision, the RTM and DTM were normalized, and feature fusion was performed, after which they were put into the VGG16* network for training. Figure 13, Figure 14, and Figure 15 show the accuracy change curve, loss value change curve, and confusion matrix after multi-feature fusion, respectively. Table 4 and Table 5 are the evaluation index of the fusion feature map and the accuracy comparison results of the three feature maps.

As illustrated in Figure 13 and Figure 14, the accuracy of the training set of the fused feature map reaches almost 100%, and the accuracy of the validation set reaches 98%. The model basically converges when the training is carried out up to 40 rounds; the loss value of the training and validation decreases to the final convergence; there is no overfitting phenomenon;

TABLE 4. Evaluation metrics for fusion feature maps.

Motion category	Precision	Recall
bend	0.988	0.988
clap	0.988	1
jump	1	1
run	0.964	1
squat	0.987	0.975
walk	1	0.963

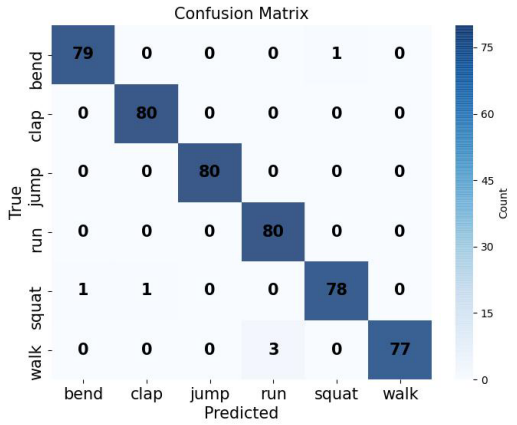

FIGURE 15. Fusion feature map confusion matrix.

and the recognition accuracy reaches a high level. Compared with the single-feature map, the accuracy is improved, and the misjudgment rate is significantly reduced. As indicated by Table 4, the recognition accuracy of bending, clapping, jumping, and running is higher than the average recognition accuracy. By comparing Table 2, Table 3, and Table 4, it is evident that squatting is easily misjudged as bending and clapping, and walking is easily misjudged as running. Except for walking, the recognition accuracy of all other human motions is improved after fusion of the features, which indicates that feature fusion not only obtains the key feature points of the range-time feature and Doppler-time feature separately, but also promotes the recognition ability of both features. After feature fusion, the overall recognition rate increases, which can prove that the fusion of range-time features and Doppler-time features can make up for the shortcomings of single-feature recognition. Additionally, it demonstrates how multi-feature fusion can effectively recognize human motion and fully depict the whole information of human motion. From the accuracy comparison in Table 5, it can be seen that the recognition accuracy of RTM is the lowest; the recognition accuracy after fusing RTM and DTM features is obviously improved; and the recognition effect of fused feature maps is better than that of single feature maps.

4.3. Effect of Image Normalization on Training Results

In order to compare the effect of normalization on the model training performance, both the datasets without image normalization and the datasets after normalization are put into the VGG16 pre-training model for training, and Table 6 displays the training outcomes.

TABLE 5. Comparison of the accuracy of the three feature maps.

	Accuracy
RTM	0.95
DTM	0.977
Fused feature maps	0.988

TABLE 6. Comparison table of normalization on training results.

	Accuracy	Number of rounds in which the model converges
Unnormalized	0.979	50
Normalized	0.988	40

It is evident from the table that following the normalization of the image, the recognition accuracy is higher, and the model converges faster, indicating that image normalization has a positive contribution to accelerating the convergence process of the neural network, making the training more stable and improving the recognition accuracy.

4.4. Comparative Evaluation of Recognition Accuracy with Various RTM and DTM Weight Ratios

The feature fusion method used in this paper is an image fusion method based on local energy features and Laplace pyramid, which calculates the weights based on local energy features and matching degree, and does not fix the weight ratios of RTM and DTM. In order to explore the effect of different weight ratios on the accuracy of human motion recognition, RTM and DTM are fused with different weight ratios, and the fused feature map is put into the VGG16 pre-training model for training. Table 7 shows the comparison results of recognition accuracy when RTM and DTM have different weight ratios.

TABLE 7. Comparison of model accuracy with different weight ratios.

Weight ratio	Accuracy
RTM : DTM $= 0.7 : 0.3$	0.981
RTM : DTM $= 0.5 : 0.5$	0.983
RTM : DTM $= 0.3 : 0.7$	0.985
Assigning weights based on local energy	0.988

From the table, it can be seen that the feature fusion method used in this paper, which calculates the weights based on the local energy features and matching degree, has a higher recognition accuracy than the feature fusion method with a fixed weight proportion, and at the same time, as the weight proportion of the DTM increases, the recognition accuracy increases, which indicates that the DTM has a higher confidence level than the RTM, and in combination with the results in Table 5, it can be shown that the DTM is more than RTM, can be utilized with higher value, and can bring better recognition effect.

4.5. Performance Analysis of VGG16 Pre-Trained Model against Other Models

For the purpose to illustrate how various convolutional neural network architectures affect the precision of motion recognition in this paper, at the same time, the traditional convolutional neural network models represented by VGG19, ResNet50, Inception V3, and Xception were constructed to train the fusion feature map, while each model had the same experimental parameters. The accuracy obtained from the VGG16 pre-trained model is compared with the other models, and Table 8 shows the comparison results of the recognition accuracy of various models.

TABLE 8. Comparison results of the accuracy of the models.

Model	Accuracy
VGG16	0.988
VGG19	0.981
ResNet50	0.985
Inception V3	0.927
Xception	0.971

The table shows that following pre-training by transfer learning, all convolutional neural network models are capable of successfully recognizing the extracted motion characteristics, and the recognition accuracy reaches more than 90%, among which the best recognition effect is the VGG16 model, with a recognition accuracy of 98.8%, which suggests that compared with the other four models, the VGG16 model is a better model for achieving the recognition of human motions.

Utilizing the full range and speed features of human motions, the feature fusion map used in this paper achieves 98.8% recognition accuracy, 1%–3% higher than the single feature map. This fully realizes the mining of the feature information of human motions and allows for the recognition and classification of human motions.

5. CONCLUSION

In this paper, a human motion recognition method based on feature fusion and transfer learning is proposed. Firstly, an FMCW radar is used to collect the human motion datasets, and the data are processed and analyzed to calculate the range parameters and Doppler parameters of the human body relative to the radar; then, the range and Doppler parameters are accumulated in the time axis to obtain the human motion RTM and DTM datasets, and in order to adequately extract and identify the multiple human motion features, the two kinds of feature spectrograms are normalized and then fused for feature fusion; finally, the two feature spectrograms and fused feature maps are put into the VGG16 network model based on transfer learning for recognition and classification. The experimental results show that the recognition accuracy of the proposed method is higher than that of a single feature map, which proves the effectiveness of the method. However, the experimental scene in this paper is set in an ideal environment without interference from other moving targets, and the recognition and detection of human movements in complex environments should be considered in subsequent research and practical applications.

ACKNOWLEDGEMENT

The authors thank the National Natural Science Foundation of China (Grant: 61561004) and the Jiangxi Graduate Student Innovation Fund Project (Grant: YC2023-S843) for the support of this study, and thank the anonymous reviewers for their help and improvements to this paper.

REFERENCES

[1] Yang, L. M. and Z. H. Li, "Design of gesture recognition system towards human computer interaction," *Industrial Control Computer*, Vol. 33, No. 3, 18–20, Mar. 2020.

[2] Zhang, Y. Y. and X. Guo, "Research and realization of dynamical gesture recognition algorithm based on kinect," *Computer Technology and Development*, Vol. 27, No. 12, 11–15, Aug. 2017.

[3] Liu, Y., R. Y. Xie, Y. Feng, *et al.*, "Survey on resident's daily activity recognition in smart homes," *Computer Engineering and Applications*, Vol. 54, No. 7, 35–42, Jan. 2021.

[4] Gao, X. W., Z. Shen, G. Y. Xu, *et al.*, "Traffic anomaly detection based on multi-target tracking," *Application Research of Computers*, Vol. 38, No. 6, 1879–1883, Dec. 2021.

[5] Tran, D., H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6450–6459, Salt Lake City, USA, Jun. 2018.

[6] Xiong, X., Y. Zheng, and S. Zhang, "Fall detection and human behavior recognition system based on long and short time memory networks and variants," *Information Communications*, No. 2, 65–67, Feb. 2020.

[7] Sabokrou, M., M. Pourreza, M. Fayyaz, R. Entezari, M. Fathy, J. Gall, and E. Adeli, "AVID: Adversarial visual irregularity detection," in *14th Asian Conference on Computer Vision*, 488–505, Perth, Australia, 2018.

[8] Liu, T., Q. Qiao, J. Wang, X. Dai, and J. Luo, "Human action recognition via spatio-temporal dual network flow and visual attention fusion," *Journal of Electronics & Information Technology*, Vol. 40, No. 10, 2395–2401, Aug. 2018.

[9] Jiang, L. B., G. Y. Wei, and L. Che, "Human motion recognition by 77 GHz radar based on dictionary learning," *Science Technology and Engineering*, Vol. 20, No. 6, 2137–2324, Feb. 2020.

[10] Li, X., Y. He, and X. Jing, "A survey of deep learning-based human activity recognition in radar," *Remote Sensing*, Vol. 11, No. 9, 1068, 2019.

[11] Lee, J., S. Hwang, S. You, W.-J. Byun, and J. Park, "Joint angle, velocity, and range estimation using 2D MUSIC and successive interference cancellation in FMCW MIMO radar system," *IEICE Transactions on Communications*, Vol. 103, No. 3, 283–290, 2020.

[12] Shrestha, A., H. Li, J. L. Kernec, and F. Fioranelli, "Continuous human activity classification from FMCW radar with Bi-LSTM networks," *IEEE Sensors Journal*, Vol. 20, No. 22, 13 607–13 619, 2020.

[13] Zhang, L. L., B. Liu, L. L. Qu, *et al.*, "Human activity recognition with FMCW radar based on fusion feature convolutional neural network," *Telecommunication Engineering*, Vol. 62, No. 2, 147–154, Jul. 2022.

[14] Wang, Y., J. Wu, Z. Tian, M. Zhou, and S. Wang, "Gesture recognition with multi-dimensional parameter using FMCW radar," *Journal of Electronics & Information Technology*, Vol. 41, No. 4, 822–829, 2019.

[15] Zhao, Y., Z. Zhang, and Z. Zhang, "Multi-angle data cube action recognition based on millimeter wave radar," in *2020 Chinese Control and Decision Conference (CCDC)*, 749–753, Hefei, China, Aug. 2020.

[16] Franceschini, S., M. Ambrosanio, V. Pascazio, and F. Baselice, "Hand gesture signatures acquisition and processing by means of a novel ultrasound system," *Bioengineering*, Vol. 10, No. 1, 36, 2023.

[17] Simonyan, K. and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *Computer Science*, 2014.

[18] Hashemi, S., H. Emami, and A. B. Sangar, "A new comparison framework to survey neural networks-based vehicle detection and classification approaches," *International Journal of Communication Systems*, Vol. 34, No. 14, e4928, 2021.

[19] Ali, M. A., H. E. A. E. Munim, A. H. Yousef, and S. Hammad, "A deep learning approach for vehicle detection," in *2018 13th International Conference on Computer Engineering and Systems (ICCES)*, 98–102, Egypt, Dec. 2018.

[20] Qi, C., Y. Zuo, Z. Chen, and K. Chen, "Rice processing accuracy classification method based on improved VGG16 convolution neural network," *Transactions of the Chinese Society of Agricultural Machinery*, Vol. 52, No. 5, 301–307, Mar. 2021.

[21] Zhuang, F. Z., P. Luo, Q. He, *et al.*, "Survey on transfer learning research," *Journal of Software*, Vol. 26, No. 1, 26–39, Jul. 2015.

[22] Liu, W. and W. Q. Ning, "Research and application of face mask wear recognition based on transfer learning," *Journal of Jilin Normal University (Natural Science Edition)*, Vol. 44, No. 1, 96–103, Feb. 2023.

[23] Zhou, K. and M. Jiang, "Research progress and prospect of small sample target recognition based on transfer learning," *Aeronautical Science and Technology*, Vol. 34, No. 2, 1–9, Feb. 2023.