

Behavior Recognition of the Elderly in Indoor Environment Based on Feature Fusion of Wi-Fi Perception and Videos

Yuebin Song, Chunling Fan[✉]

Abstract: With the intensifying aging of the population, the phenomenon of the elderly living alone is also increasing. Therefore, using modern internet of things technology to monitor the daily behavior of the elderly in indoors is a meaningful study. Video-based action recognition tasks are easily affected by object occlusion and weak ambient light, resulting in poor recognition performance. Therefore, this paper proposes an indoor human behavior recognition method based on wireless fidelity (Wi-Fi) perception and video feature fusion by utilizing the ability of Wi-Fi signals to carry environmental information during the propagation process. This paper uses the public WiFi-based activity recognition dataset (WIAR) containing Wi-Fi channel state information and essential action videos, and then extracts video feature vectors and Wi-Fi signal feature vectors in the datasets through the two-stream convolutional neural network and standard statistical algorithms, respectively. Then the two sets of feature vectors are fused, and finally, the action classification and recognition are performed by the support vector machine (SVM). The experiments in this paper contrast experiments between the two-stream network model and the methods in this paper under three different environments. And the accuracy of action recognition after adding Wi-Fi signal feature fusion is improved by 10% on average.

Keywords: human behavior recognition; two-stream convolution neural network; channel status information; feature fusion; support vector machine (SVM)

1 Introduction

With the intensifying aging of the population, the phenomenon of the elderly living alone has also gradually increased. At the same time, with the growth of age, the body functions of the elderly begin to degenerate, and the probability of accidents also gradually increases. The survey shows that the probability of the elderly getting help for indoor accidents is only about 65%,

among which falling and collision account for 80% of the accident types [1, 2]. They are also the main causes of injury or death. This is due to the lack of health supervision for the elderly [3]. Therefore, it is a meaningful study to improve the indoor detection technology of elderly activities and reduce the occurrence of accidents by utilizing advanced technologies of artificial intelligence and the internet of things in modern.

With the development of machine learning and artificial intelligence, computer vision has made good progress in the fields of human behavior recognition [4, 5], video event retrieval [6], abnormal behavior detection [7–9], and other video tasks. The research on human behavior recognition is also of great significance to the development of the above video recognition tasks. With the in-depth study of deep learning and the

Manuscript received Nov. 16, 2022; revised Dec. 27, 2022; accepted Jan. 30, 2023. The associate editor coordinating the review of this manuscript was Dr. Kun Qian. This work was supported by the National Natural Science Foundation of China (No. 62006135) and the Natural Science Foundation of Shandong Province (No. ZR2020QF116).

Yuebin Song and Chunling Fan are with College of Automation and Electronic Engineering, Qingdao University of Science and Technology, Qingdao 266100, China.

✉ Corresponding author. Email: chunlingfan@qust.edu.cn
DOI: [10.15918/j.jbit1004-0579.2022.131](https://doi.org/10.15918/j.jbit1004-0579.2022.131)

continuous expansion of the scale of various public datasets, the research on human behavior recognition has entered a diversified field, among which video-based human behavior recognition is the most widely used [10]. However, most videos have good quality and prominent motion targets in the general public training datasets. However, it is often difficult to obtain high-quality action videos in daily life due to the diverse and complex living environment characteristics. It also makes the video-based action recognition task vulnerable to the interference of environmental factors such as different target angles, camera shake, background clutter, target occlusion and so on, resulting in the loss of video action information and the degradation of recognition performance [11, 12]. Therefore, improving the performance of human behavior recognition, which is easily affected by environmental factors, is a research project of great significance.

There are many kinds of methods of learning for the depth of the human behavior recognition task [13–15], including the traditional convolutional neural networks (CNN), the 3 dimensional convolutional neural network (3DCNN), the time series models based on short-term memory network, the two-stream neural network, etc. Among them, Tran et al. proposed the convolutional 3D network (C3D) [16] in which 3D convolution is used. 3D convolution network and 3D pooling operations can simultaneously obtain temporal and spatial feature information from continuous video frames, so as to realize behavior recognition of behavior objects in video. Kiwon Rhee et al. used depth visual guidance to apply electromyography (EMG) signals to gesture recognition tasks [17]. This method uses information other than video to complete gesture recognition, making it possible for more multi-feature fusion recognition tasks. Hochreiter et al. proposed the long short-term memory (LSTM) [18], which uses three different doors to preserve and forget information, making up for the shortcomings of gradient explosion and loss in the initial recursive neural network (RNN).

Wen Qi et al. used recursive neural networks to study the use of multi-sensor-guided gesture recognition [19]. Based on the multisensor fusion model, a multilayer RNN consisting of an LSTM module and a dropout layer (LSTM-RNN) is presented, which is used for multi-gesture classification and shows strong anti-interference ability. And Ref. [20] used a multi-modal wearable remote control robot to complete the task of breath pattern detection. Simonyan and others put forward the innovation of the two-stream convolutional network for human behavior recognition tasks [21]. The network consists of spatial and temporal convolution networks that do not interfere with each other. The two networks extract their features, fuse them in a certain way, and finally perform classification and recognition. The network fully uses the spatiotemporal information in the action video and improves the performance of video-based action recognition. However, in practical application, the action video used to extract optical flow information in two flow convolutional neural networks is easily affected by environmental factors, resulting in the loss of action information, incomplete optical flow information extraction, and action recognition performance degradation.

With the application of orthogonal frequency division multiplexing (OFDM) [22] technology, researchers have found that channel state information (CSI) signals have a higher sensitivity to the environment in the propagation process and can provide more specific and accurate grained information [23, 24]. In 2014, Halperin et al. released a CSI measurement tool based on an Intel-5300 network adapter (CSI-Tool) [25]. A large number of research works on CSI-based perception have emerged. Among them, Wang et al. took the CSI as the detection signal of human fall behavior [26]. This method provides a reference for CSI signals used in human behavior recognition tasks. In [27], Guo et al. constructed a Wi-Fi based activity recognition dataset (WIAR) and proposed a Human Activity

(HuAc) system, proposed a subcarrier selection mechanism based on the sensitivity of subcarriers to human behavior, and established a relationship library between CSI and human joint activity to achieve pattern matching. [28] presented a Wi-Fi-enabled gesture recognition using dual-attention network (WIGRUNT) model using residual network (ResNet), which used a dual-attention mechanism to distinguish whether there are gesture movements and to identify the categories of actions by dynamically focusing on gesture movements in a time dimension and evaluating the correlation of signal CSI sequences.

Therefore, the main contributions and innovations of this paper are as follows. This paper proposes the method of using the action information carried by the wireless fidelity (Wi-Fi) channel state information to compensate for the lost action information when the video action recognition task is affected by environmental factors. This paper mainly designs an indoor human behavior recognition method based on Wi-Fi signal perception and video feature fusion. This method extracts Wi-Fi signal features from a standard statistical algorithm, and two streams of convolutional neural network extract video

action features. The two groups of feature vectors are fused [29, 30] and input into a Support Vector Machine (SVM) [31, 32] for classification training. Finally, through the comparison experiment in three different environments, it is verified that the recognition model after feature fusion has higher accuracy.

2 Overall Scheme Design

The indoor human behavior recognition method based on Wi-Fi signal perception and video feature fusion is mainly divided into three parts: video feature extraction, Wi-Fi signal CSI feature extraction, feature fusion and classification. Fig. 1 shows the workflow of the human behavior recognition method based on Wi-Fi perception and video feature fusion. In the video feature extraction part, two streams of a convolutional neural network are used to extract the action features in the video. The standard statistical algorithm is used to remove the Wi-Fi signal features in the Wi-Fi signal acquisition. Finally, the two groups of feature vectors are fused and input into SVM for classification training.

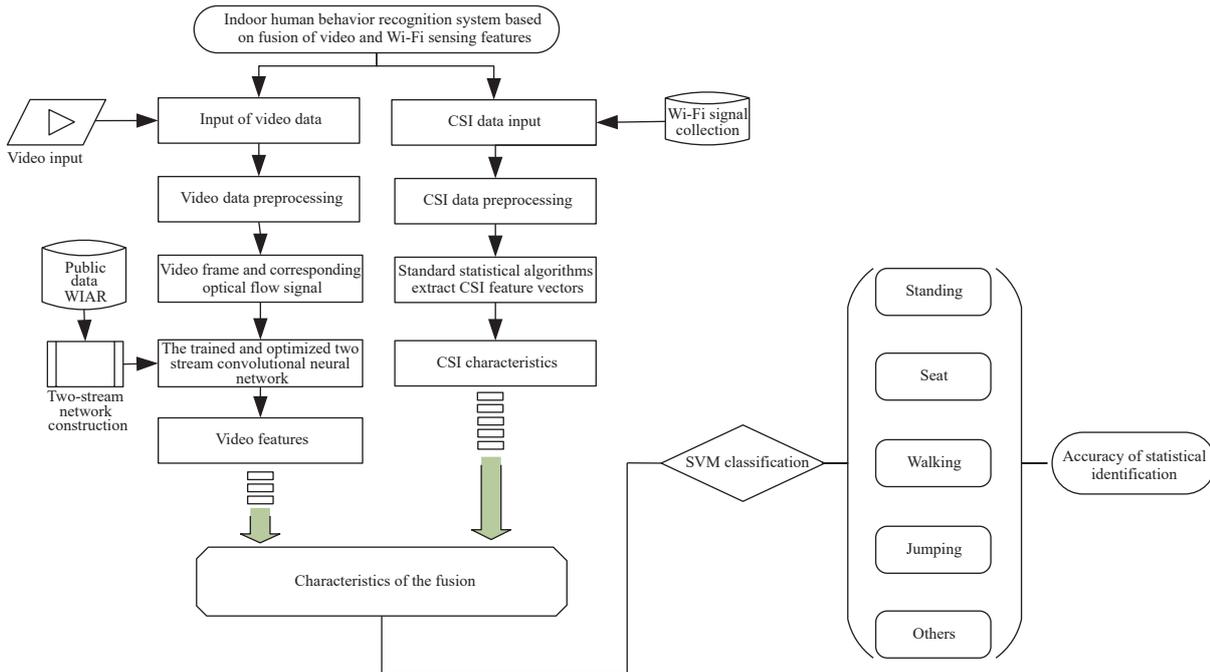


Fig. 1 Human behavior recognition process design based on feature fusion of Wi-Fi perception and video

3 Methodology

3.1 Video Feature Extraction

The information contained in the video can be divided into two dimensions: space and time. The spatial information mainly includes the background and object information in the video. Time information refers to the relative displacement of the object's position along the time axis in a continuous video frame. The two-stream network-based human behavior recognition framework proposed by Simanyan et al. makes full use of the spatiotemporal information of action videos [21]. The framework includes two independent convolutional networks, one of which is used to extract

the features of a single video image, that is, extract spatial information, which is called a spatial convolutional network. The other one is used to extract the optical flow information of the video, that is, to extract the time information, which is called the temporal convolutional network. The two network structures are trained on the single frame image of the sample video and the optical flow image extracted from the continuous video frame, respectively. Finally, the feature vectors extracted from the two networks are fused through model fusion to obtain the global feature representation of the sample. Fig. 2 shows the basic structure of a two-stream network.

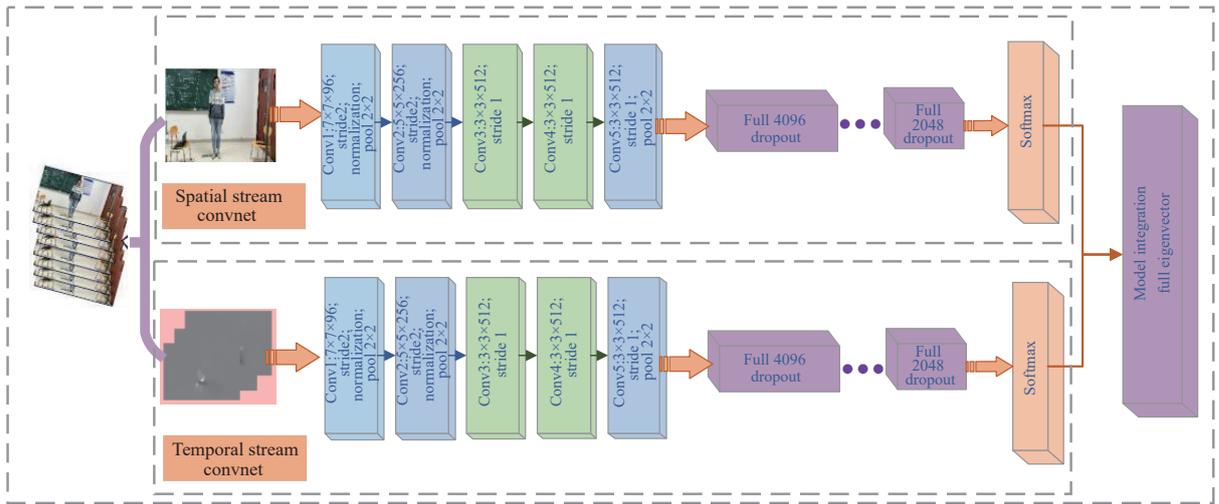


Fig. 2 Two-stream convolution neural network framework

The upper part of Fig. 2 is the structure of a spatial convolutional network. A spatial convolutional network is a classification model that trains the target features of samples based on a single video frame and generates the classification model of such images. When extracting spatial features, the extracted video frames need to be pre-processed. Firstly, all video frames are shortened to 256×256 , and then 25 video frames are selected with equal spacing. Subsequently, a 224×224 sub-image is obtained by cropping from the top left corner, then the video frame is inverted 90° counterclockwise, and again a 224×224 sub-image is obtained by cropping from

the top left corner. Repeated four times, cropping and flipping the four corners and the center of each frame, i.e., we can get 5 sub-images per frame. Finally, input the spatial network to extract features. The spatial network in the figure has a total of 8 steps, covering five convolution layers, three pooling layers, two fully connected layers, and a classification layer (Softmax), in which the size and stride of the convolution kernel are set for each convolution layer. For example, in the first layer, the size of the convolution kernel is 7×7 , and the number of convolution kernels is 96. In the convolutional network, the size of the pooling layer is 2×2 , two

fully connected layers are set behind the last pooling layer, and a feature vector of 2 048 dimensions is output in the seventh layer. A few training samples often lead to an overfitting phenomenon in the training process. The overfitting phenomenon refers to the high recognition rate during training due to the small number of training sets but the low recognition rate during experimental testing. To reduce the overfitting phenomenon, the dropout function is added to the fully connected layer so that some neuronal features do not participate in the training. Still, their feature values are retained in the training model, which improves the robustness of the neural network. The values of the two-layer dropout function in this paper are 0.5 and 0.9, respectively.

The lower part of Fig. 2 shows the temporal convolutional network structure. When extracting time characteristics, video frames are converted to optical flow maps, which range from dense status to sparse status to reduce optical flow storage space, specifically to compress red, green and blue system (RGB)-like images. Save all rescaled optical flow values to $[0, 255]$ integer as JPEG pictures. In this way, each optical flow diagram will be more than 10 kbit, and the storage space will be greatly reduced. The resulting optical flow map is then rotated and clipped, similar to video frame preprocessing in the spatial network. Finally, $224 \times 224 \times 2L$ light flow information for each sub-image is calculated and fed into the temporal convolutional network to extract the time characteristics. The temporal convolutional network calculates the displacement of pixels on the time axis in a continuous video frame image. Finally, the expulsion of all pixels forms an optical flow field, which is decomposed into horizontal and vertical vectors. The vector information of these two directions is the convolutional network's channel input sample, and the time feature is extracted by convolution. The concrete implementation step is to regard

the optical flow information of pixels as a set of displacement vectors \mathbf{d}_t between adjacent video frames t and $(t + 1)$. Random in a pixel (u, v) , by point derivative $\mathbf{d}_t(u, v)$, can be said that the pixels between t and $(t + 1)$ frame, the displacement vector of the image, and the image sample can be regarded as a vector field, including horizontal and vertical components and \mathbf{d}_{yt} can be regarded as the input channel of the image. To show movement in continuous video frames, the time convolution input L adjacent frames of the optical flow vector field of the channel \mathbf{d}_{xt} and \mathbf{d}_{yt} , and formed $2L$ input channels. Set the width and height of the video size of w and h , respectively. For any frame n input convolution network capacity $\mathbf{I}_m(u, v, c) \in \mathbf{R}^{2^{whL}}$ by the following to achieve

$$\mathbf{I}_m(u, v, (2k - 1)) = \mathbf{d}_{n+k-1}^x(u, v) \quad (1)$$

$$\mathbf{I}_m(u, v, 2k) = \mathbf{d}_{n+k-1}^y(u, v) \quad (2)$$

$$(u \in [1, w]; v \in [1, h]; k \in [1, L])$$

where $\mathbf{d}_{n+k-1}^x(u, v)$ and $\mathbf{d}_{n+k-1}^y(u, v)$ indicate that the pixel point is in a specific direction, and x and y represent the horizontal and vertical directions of the coordinates established for the image frame. So for a bit (u, v) , the channel $\mathbf{I}_m(u, v, c)$, $c \in [1, 2L]$ on L frames is used to encode point of displacement, and then the encoded information is input for convolution network feature extraction.

Finally, the two features are fused through model fusion, as shown on the right side of Fig. 2. The feature vectors extracted from spatial and temporal convolutional networks are arranged in a particular order to obtain the global features of the samples. This paper uses the two-stream convolution neural network framework displayed in Tab. 1. In the two-stream network framework used in this paper, characteristics of 2 048 dimensions are extracted from spatial and temporal networks, respectively. After the classification by softmax layer, the classification results are fused,

Tab. 1 Two-stream convolution network parameters

Parameter	The spatial network value	The temporal network value
Input layer	224×224	224×224×2L
Convolution layer	7×7×96	7×7×96
	stride 2;	stride 2;
	5×5×256	5×5×256
	stride 2;	stride 2;
	3×3×512	3×3×512
	stride 1;	stride 1;
	3×3×512	3×3×512
	stride 1;	stride 1;
Pooling layer	2×2	2×2
Full connection layer	4096; 2048	4096; 2048
Activation function	Rectified linear unit (ReLU)	ReLU
Loss function	Cross-entropy loss function	Cross-entropy loss function
Optimizer	Adam	Adam
Dropout	0.5; 0.9	0.5; 0.9
Learning rate	0.001	0.001

and finally, a vector descriptor with the video feature of 4096 dimensions is obtained.

3.2 Wi-Fi Signal Feature Extraction

CSI is a kind of information that can carry the variation characteristics of its transmission communication link. This information can measure the variation of channel state and the weakness degree of Wi-Fi signal on multiple transmission paths.

The CSI measures the channel by portraying each multipath component (multipath transmission) with time and frequency domain information. In the time domain, CSI uses the channel impulse response to represent the energy value of the signal arriving at the receiver. CSI is more sensitive to the environment and better portrays the environmental changes caused by the actions, as it can describe the channel changes from time and frequency domain information through subcarriers respectively. The principle of using CSI in Wi-Fi signals for action recognition is that Wi-Fi signals will form different multipath reflections when they encounter moving targets during propagation, which makes

the CSI parameters at the receiving end change and thus form different CSI waveforms. Therefore, different movements can be identified according to the different CSI waveforms.

CSI signals are generally described as channel impulse response (CIR). The frequency domain expression of CSI signal is transformed by International Football Friendship Tournament (IFFT) to obtain CIR.

CIR can be expressed as

$$h(t) = \sum_{i=1}^N \alpha_i e^{-j\theta_i} \delta(t - \tau_i) \quad (3)$$

where α_i is the amplitude decay of the i th path, θ_i is the phase offset of the i th path, τ_i is the time delay of the i th path, N is the total number of paths, and $\delta(t - \tau_i)$ is the Dirac δ function.

After Fourier transform, $H(f_i)$ is the CSI response value with center frequency f_i , where $|H(f_i)|$ is the amplitude value and $\angle H(f_i)$ is the phase value.

$$H(f_i) = |H(f_i)| e^{j\angle H(f_i)} \quad (4)$$

In the acquisition process, each packet containing a group of subcarriers is sent, and its expression is shown as

$$H = [H(f_1), H(f_2), \dots, H(f_i), \dots, H(f_{NS})] \quad (5)$$

where i represents the number of subcarriers, N represents the number of data packets received by each antenna, and S represents the number of antennas capable of receiving data.

After multipath transmission, the received signal at the receiver is expressed as

$$Y(f, t) = H(f, t) + X(f, t) + N_{\text{noise}} \quad (6)$$

where, $Y(f, t)$ is the frequency domain representation of the received signal, $X(f, t)$ is the frequency domain representation of the transmitted signal, $H(f, t)$ is the channel frequency response (CFR) at the time t , f is the center frequency of the subcarrier, and N_{noise} is the environmental noise carried in the propagation process. Thus, the representation of the CSI signal is obtained. For details, please refer to reference [33].

As CSI is a fine-grained physical layer signal, it is more sensitive to the environment. It can carry environmental information, so CSI signal is often used for recognition in the research of human behavior recognition. The data values of CSI are the amplitude and phase of each subcarrier corresponding to the frequency domain space after OFDM technology decoding. These values are the action information, and some noise carried in the propagation process. OFDM enables Wi-Fi signals to be transmitted in parallel

through multiple carrier channels, significantly improving communication efficiency. This technology is widely used in Wi-Fi wireless devices. The main working principle of OFDM is to convert the Wi-Fi signal into several subcarriers, which are orthogonal to each other, and then modulate the subcarriers to the sub-channels for parallel low-speed data stream transmission. The orthogonal feature of subcarriers can reduce the interference among transmission channels. Fig. 3 shows how it works.

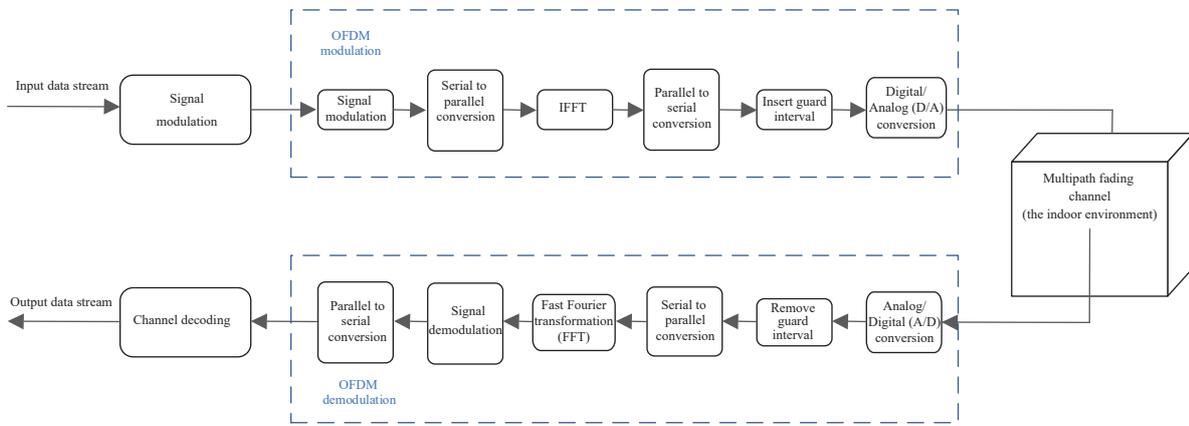


Fig. 3 Working principle of OFDM system

Fig. 4 is a schematic diagram of action acquisition. The transmitter of the signal (the left part of Fig. 4) is a Wi-Fi router; the receiver of the call (the right amount of Fig. 4) is a computer equipped with an Intel-5300 Network Interface Card (NIC); the middle part is the active area of the moving target, and the sender sends the signal. During the propagation process, multipath reflection occurs through static environments such as the mover or the ground, forming different propagation paths and finally collected by the receiver.

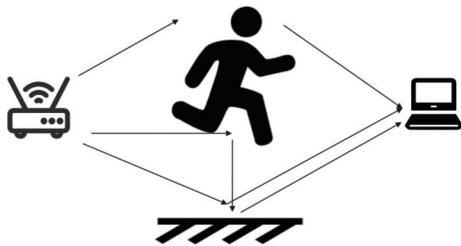


Fig. 4 Schematic diagram of CSI signal acquisition

In this paper, the public data set WIAR is used to extract CSI signal features for classification training, which contains 16 sets of CSI action data, such as standing, squatting, and sitting, completed by three volunteers. The data is in the form of a “.dat” file, which can be directly processed in matrix laboratory (MATLAB). During the action recognition and detection experiment, the Wi-Fi signal sending and receiving environment should be established in the room. The Wi-Fi router completes the Wi-Fi signal-sending terminal. The acquisition program is run first, and the router starts to send data. The movement process will cause a change in the transmission path of the Wi-Fi signal, so the movement information is carried in the transmission process. The computer equipped with the Intel-5300 NIC receives Wi-Fi signals and saves them. The CSI-tool saves the collected Wi-Fi signals as a “.dat” file. Because the receiver of the

Intel-5300 network adapter is equipped with three antennas, each antenna receives 30 subcarriers at a time. So each data packet is a data matrix of 3×30 . The collected CSI waveform is shown in Fig. 5.

The collected CSI data needs to be reprocessed. Abnormal sample points will inevitably appear in the process of data collection. This paper uses Hampel outlier detection to eliminate the data values with large differences. MATLAB

was used to write an outlier detection program, which stipulated that the median value of the corresponding sample point and the standard deviation of the pair median value were calculated for 30 subcarriers of an input data packet. If the sample point exceeded or was equal to three standard deviations below the median value, the sample point would be an outlier. The median value would be used to replace the sample point.

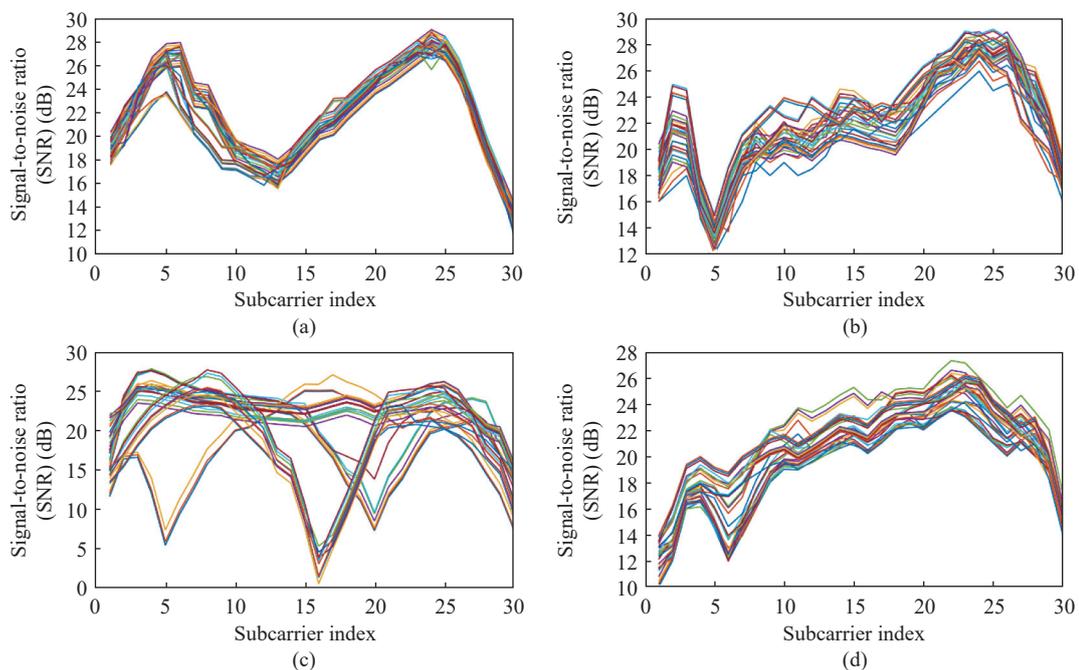


Fig. 5 CSI waveform diagram: (a) standing; (b) sitting; (c) jumping; (d) walking

Since Wi-Fi signals are susceptible to an indoor environment, changes in the indoor environment (such as the unintentional actions of collectors, ambient temperature, etc.) will affect the information carried by the channel state information CSI. Therefore, Wi-Fi signals inevitably have noise, which makes it impossible to extract Wi-Fi signal features directly. The frequency of CSI waveform changes caused by moving targets belongs to the low-frequency part. By contrast, the frequency of environmental noise carried by moving targets belongs to the high-frequency domain. Therefore, low pass filtering is adopted in this paper to reduce noise after outliers are removed. And the PAC principal compo-

nent analysis is used to extract the CSI waveform features.

Finally, the waveform is interpolated. Interpolation is the numerical estimation of points where data records do not yet exist, based on a known data sequence, according to some law. Since the CSI signal is subject to some loss on some links due to absorption by various furniture, equipment and walls, it can happen that the collected packets will have a small loss of deviation. Therefore, to ensure the integrity of the experiment, each subcarrier stream amplitude is interpolated according to the actual waveform to reduce and offset the loss. Because only a small amount of data needs to be estimated, the

amount of data is small and the computational effort is small. So linear interpolation is chosen. In other words, the two data adjacent to the left and right of the packet that needs to be interpolated in the sequence are estimated numerically for filling, and the signal segments containing action information are marked out to reduce interference for subsequent feature fusion and action classification.

Finally, the standard statistical algorithm is used to extract the features of the image data. This paper selects the average value, maximum value, minimum value, standard deviation, amplitude, average absolute deviation, variance, and eight mode feature values. There are three antennas at the signal receiver, each antenna has 30 subcarriers at a time, and each waveform extracts eight eigenvalues. Therefore, the feature vector of 720 dimensions is finally obtained and saved in the file of the “.mat” type, namely the feature vector of the Wi-Fi signal.

3.3 Feature Fusion and Classification

Feature fusion is the method of extracting the feature information of the same research object in different ways and fusing the feature information to obtain new features. The feature fusion method can compensate for the shortcomings of different techniques and complement each other. This paper adopts the approach of early fusion.

Prefusion is usually used in traditional machine learning, a relatively simple and convenient method. The feature information extracted differently will be spliced and fused into a new feature vector with a specific length and the sum of multidimensions as the final feature representation of the research object [34]. When multi-feature fusion occurs, it is unavoidable that the dimension of feature data is not uniform. The eigenvectors can then be dimensionally adjusted or reduced [35]. Then the corresponding elements can be fused into a new feature vector through accumulation and multiplication. Finally, classification learning is committed to completing the recognition task.

In this paper, the video feature vector and CSI feature vector are fused in the early stage, and the two groups of feature vectors are directly spliced together. The dimension of the video feature vector is 4 096 dimensions, the CSI signal feature vector is 720 dimensions, and the dimension of each sample feature vector after fusion is 4 816 dimensions. Fig. 6 shows the implementation process of early fusion. It can be seen from Fig. 6 that CSI feature values are directly spliced together with video feature vectors after adjusting the dimension of feature vectors. Ensure that all eigenvalues are equally involved in classification calculation and classification recognition.

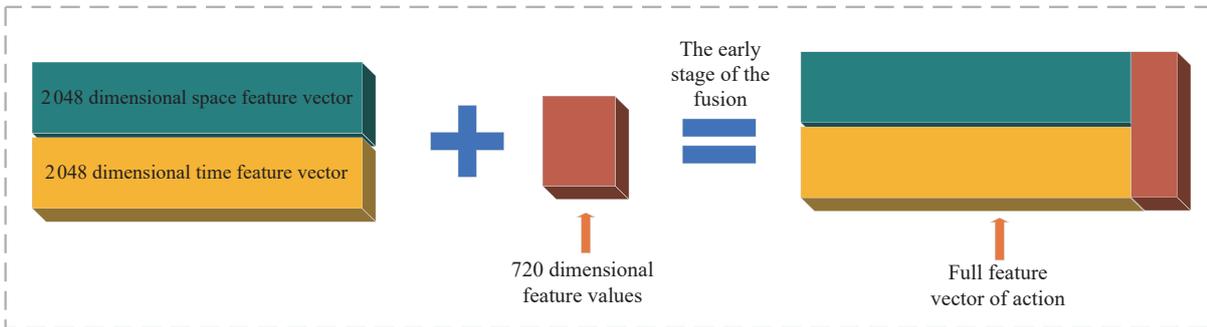


Fig. 6 Schematic diagram of feature fusion

When the new feature vector after feature fusion is obtained, it can be input to the classifier for action classification and recognition. The commonly used classifier is the SVM algorithm, which is usually called a classifier [36].

At first, SVM was mainly used to solve binary classification problems, and then it was gradually developed and applied in multi-classification tasks and became the mainstream algorithm in the field of traditional machine learning.

After the known feature information samples are input to SVM, the relationship between the feature data and the sample label is found through training, and a function model for classification is finally generated. The generated function model is used to classify and predict the unknown feature information. According to the classification method, it can be divided into linear classification and linear non-classification. In this paper, the linear function is selected as the kernel function, and its expression is as follows

$$f(x, x_i) = kxx_i \quad (7)$$

where x is the sample to be identified, x_i is the sample for training classification, and $f(x, x_i)$ is used to calculate the similarity between sample x and training sample x_i .

Its linear classification principle is shown in Fig. 7. The essence of SVM's classification task is to find an optimal classification decision hyperplane for similar features by repeatedly training the features in the data set. When the new feature information is input, the feature information can be accurately classified according to the linear kernel function model corresponding to the hyperplane. It can be seen from Fig. 7 that the circle and the cross represent two different categories of feature data sets, namely samples. The two samples are separated by a straight line, and this classification is linearly separable. Among them, the distance d between the circle and the cross is called the interval. SVM obtains the classification decision hyperplane by solving the maximum interval d , which can accurately separate the sample points of different categories.

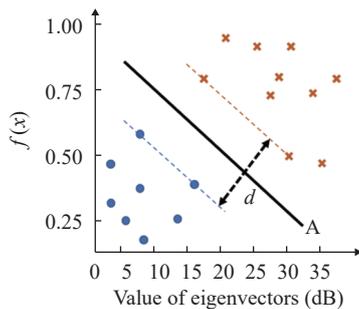


Fig. 7 Linear separable optimal hyperplane

Assuming hyperplane classifier expressions for $f(x) = \omega x + b$ (ω is parameter matrix, b is intercepting), can separate the two kinds of feature information thoroughly, when $f(x) = 0$, and sample x is on the hyperplane. When the sample point function value $f(x) \leq 0$, its category label is (-1) , and when $f(x) > 0$, the sample label is 1, which can realize the recognition of samples.

According to the experimental environment designed in this paper, Library for Support Vector Machines (LibSVM) is used as a classifier for the fused feature vectors. LibSVM is an open-source SVM software package developed in a MATLAB environment, which can be compiled and used directly in MATLAB software. This package provides many parameter settings for the function model, and different settings of these parameters can solve various classification problems. Facing the requirements of other classification problems, it only needs to complete the required parameter settings for the selected kernel function model, which reduces the frequent training links of feature information and the learning difficulty.

LibSVM provides multiple learning methods, such as one-to-one and one-to-many, which is suitable for numerous dichotomous or multi-classification tasks. In this paper, cross-validation is used to evaluate classification accuracy. The specific implementation steps are as follows.

1) Firstly, the training dataset is divided into five action sub-datasets.

2) Then, the sub-datasets of one action are selected as one class, and the four remaining action data sets are selected as the other class. This is equivalent to constituting a dichotomous classifier.

3) Take 5 consecutive times, construct 5 classifiers, and record the recognition accuracy of each classifier.

4) Finally, the average value of the results of the five classifiers is calculated as the final recognition accuracy.

4 Experimental Design and Analysis of Results

4.1 Experimental Environment

According to the experimental requirements, the testing equipment used in this paper is as follows.

1) Hardware

Wi-Fi router with TP-Link, camera, a laptop computer equipped with Intel-5300 NIC.

2) Software

MATLAB 2019, Pycharm 2020, LibSVM.

3) Experimental Environment

During the experiment, a quiet and closed environment was maintained to minimize the interference of environmental noise. The Wi-Fi router is the transmitter of Wi-Fi signals, and the Intel-5300 NIC is the receiver. The distance between the transmitter and receiver is 3 m, and the middle is the motion range of the mover. In front of the athlete there is a camera to record video of the action.

In order to verify the change in action recognition performance after feature fusion, the experiment is divided into three parts. The two-stream network framework and the feature fusion model are used to conduct comparison tests in a normal environment, dark environment, and partially occluded environment.

4.2 Analysis of Experimental Data

During the experiment, the five actions were recognized and detected by the two-flow convolutional neural network and the proposed feature fusion method in three environments, namely normal environment, dark light and local occlusion. The recognition results were statistically analyzed in the following section.

Tab. 2 shows the statistics of action recognition results in the comparison test under a normal environment. As can be seen from Tab. 2, under normal circumstances, both the two-stream convolutional network behavior recognition and the feature fusion method proposed in this paper have good performance. And the recognition rate

of the five types of movements is above 80%, with the recognition accuracy of standing and walking adding CSI feature fusion reaching 100%. While other action recognition accuracy remains the same. The reason is that the recognition rate is low due to the different jumping heights and jumping postures of different moving targets.

It is not difficult to see from Tab. 2 that human behavior recognition based on the two-stream network framework has a good performance in a similar typical environment. At the same time, the feature fusion method with Wi-Fi signal can only improve the accuracy of behavior recognition for standing and walking, which have a small range of actions and similar actions of different targets.

Tab. 2 Data statistics under normal environment

Action category	Two-stream network identification		Video and CSI feature fusion identification	
	Identification number	Accuracy	Identification number	Accuracy
	Standing	10	90.0%	10
Squatting	10	90.0%	10	90.0%
Sitting	10	90.0%	10	90.0%
Walking	10	90.0%	10	100.0%
Jumping	10	80.0%	10	80.0%

Tab. 3 shows the statistical results of action recognition in the comparison test under a low light environment. From Tab. 3 analysis, in the case of dimmed environmental light, caused by the change of the optical flow information, activity recognition based on the two-stream convolutional network performance descends, and recognition accuracy of five kinds of action are down. Although the performance of activity recognition based on video and CSI feature fusion falls slightly, its recognition accuracy is still higher than the former under the same environmental conditions. The recognition performance is much higher than action recognition based on two-stream network. This also proves that motion recognition incorporating Wi-Fi signal fusion can overcome environmental interference caused by light changes.

Tab. 3 Data statistics in dark environment

Action category	Two-stream network identification		Video and CSI feature fusion identification	
	Identification number	Accuracy	Identification number	Accuracy
	Standing	10	80.0%	10
Squatting	10	80.0%	10	90.0%
Sitting	10	80.0%	10	90.0%
Walking	10	70.0%	10	90.0%
Jumping	10	60.0%	10	80.0%

In this experiment, the moving area of the moving target was partially set below the knee. As can be seen from Tab. 4, because the moving target in the video was blocked, the action recognition performance based on the two-stream network was significantly decreased, and the movements with obvious leg floating in squatting and walking were significantly affected by the occlusion. The motion recognition based on video and Wi-Fi signal feature fusion has little influence on the recognition performance due to CSI feature compensation. However, the motion posture and amplitude of different moving targets are different, and Wi-Fi carries environmental noise, which cannot be eliminated, so the recognition accuracy of squatting and walking actions is slightly reduced. It can be concluded from Tab. 4 that local occlusion has a great impact on the performance of action recognition based on a two-stream network. After adding Wi-Fi signal feature fusion, the performance of action recognition is improved due to the action information carried by Wi-Fi.

Tab. 4 Data statistics under partial occlusion

Action category	Two-stream network identification		Video and CSI feature fusion identification	
	Identification number	Accuracy	Identification number	Accuracy
	Standing	10	80.0%	10
Squatting	10	60.0%	10	80.0%
Sitting	10	70.0%	10	90.0%
Walking	10	60.0%	10	80.0%
Jumping	10	70.0%	10	80.0%

Through comparative tests in three environments, it can be seen that environmental factors

have a great impact on the performance of action recognition based on the two-stream network. After CSI feature fusion is added, although the recognition accuracy is slightly decreased due to environmental interference, the overall recognition accuracy is significantly improved compared with that of the two-stream network. After experimental verification, the indoor human behavior recognition method based on Wi-Fi perception and video feature fusion proposed in this paper can improve the performance of human behavior recognition based on the two-stream network when it is interfered by environmental factors.

5 Conclusion

This paper introduces action recognition based on video task easily affected by environmental factors such as light, and background. Therefore, in order to improve the performance of action video under the influence of environmental factors, this paper proposes a human behavior recognition solution for action video that utilizes information features carried by Wi-Fi signals to compensate for information loss due to environmental factors. An indoor human behavior recognition method based on Wi-Fi perception and video feature fusion is designed. The extracted video feature vectors and Wi-Fi channel state information feature vectors are fused and finally input into the support vector machine to complete classification and recognition.

Finally, after adding the two-stream convolutional networks and a feature fusion system under the environment of the three different contrast experiments, respectively, the experiment selected the five experiments, compared to the experimental data, after joining the Wi-Fi signal feature fusion of 5 kinds of human action recognition accuracy which were improved, and could overcome particular environmental interference. Considering that video and Wi-Fi signals have advantages and disadvantages, the research of

weighted feature fusion based on video and Wi-Fi signals has excellent potential. Looking forward to the future, the feature fusion of the video and Wi-Fi signals can overcome environmental interference and effectively improve the accuracy of indoor video tasks.

References:

- [1] B. Qiu and C. Wen, "Investigation on the status Quo and influencing factors of accidental injuries among the elderly in Suzhou," *General Nursing*, vol. 21, no. 14, pp. 2252-2253, 2016.
- [2] M. Yuan, S. Wei, J. Zhao, and M. Sun, "A systematic survey on human behavior recognition methods," *Springer Nature Computer Science*, vol. 3, no. 1, pp. 1-25, 2021.
- [3] S. Lee, "Falls associated with indoor and outdoor environmental hazards among community-dwelling older adults between men and women," *Biomed Central Geriatrics*, vol. 21, no. 1, pp. 134-146, 2021.
- [4] S. Hang, Q. Wen, Y. Schmirander, S. Ertug Ovrur, S. Cai, and X. Xiong, "A human activity-aware shared control solution for medical human-robot interaction," *Assembly Automation*, vol. 42, no. 3, pp. 388-394, 2022.
- [5] K. Qian, T. Koike, T. Nakamura, W. Schuller, and Y. Yamamoto, "Learning multimodal representations for drowsiness detection," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 8, pp. 11539-11548, 2022.
- [6] S. Leia, F. Reza, B. Terry, F. Tamim, K. Souraiya, N. Hoda, S. Sofija, T. James, K. Shehroz, and I. Andrea, "Indoor location data for tracking human behaviours: A scoping review," *Sensors*, vol. 22, no. 3, pp. 1220-1228, 2022.
- [7] J. Peng, Y. Zhao, and L. Wang, "Research on video abnormal behavior detection based on deep learning," *Laser & Optoelectronics Progress*, vol. 58, no. 6, pp. 6-14, 2021.
- [8] W. Qi and H. Su, "A Cybertwin based multimodal network for ECG patterns monitoring using deep learning," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 10, pp. 6663-6670, 2022.
- [9] H. Si, X. Hu, and Y. Wang, "Research on video-based human action behavior recognition algorithms," *Institute of Physics Conference Series: Earth and Environmental Science*, vol. 440, no. 3, pp. 32142-32142, 2020.
- [10] G. Yang and W. Zou, "Overview of video behavior classification methods based on deep learning," *Electronic Technology Application*, vol. 48, no. 7, pp. 7-12, 2022.
- [11] X. Wang and M. Zhi, "Summary of object detection based on convolutional neural network," in *Eleventh International Conference on Graphics and Image Processing (ICGIP)*, Hangzhou, China, pp. 151-158, 2020.
- [12] Y. Zhao, R. Li, X. Zhang, Y. Cao, and X. Chen, "Research on recognition algorithm of abnormal behavior of workers in two-stream convolutional network," *Journal of Physics: Conference Series*, vol. 1621, no. 1, pp. 188-195, 2020.
- [13] Q. Ye, H. Zhong, C. Qu, and Y. Zhang, "Human interaction recognition method based on parallel multi-feature fusion network," *Intelligent Data Analysis*, vol. 25, no. 4, pp. 809-823, 2021.
- [14] M. Yuan, S. Wei, and Q. Sun, "A systematic survey on human behavior recognition methods," *Springer Nature Computer Science*, vol. 55, no. 10, pp. 98-104, 2021.
- [15] X. Sun, T. Ren, Y. Zi, and G. Wu, "Video visual relation detection via multi-modal feature fusion," *Multimedia*, vol. 19, no. 1, pp. 2657-2661, 2019.
- [16] D. Tran, L. Bourdev, and R. Fergus, "Learning spatiotemporal features with 3D convolutional networks," in *IEEE 2015 IEEE International Conference on Computer Vision (ICCV)*, Santiago, Chile, pp. 4489-4497, 2015.
- [17] K. Rhee and H. C. Shin, "Electromyogram-based hand gesture recognition robust to various arm postures," *International Journal of Distributed Sensor Networks*, vol. 14, no. 7, pp. 1-15, 2018.
- [18] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735-1780, 1997.
- [19] W. Qi, O. S. Ertug, Z. Li, M. Aldo, and R. Song, "Multi-sensor guided hand gesture recognition for a teleoperated robot using a recurrent neural network," *IEEE Robotics and Automation Letters*, vol. 6, no. 3, pp. 6039-6045, 2021.
- [20] Y. Zhang, L. Shi, Y. Wu, K. Cheng, J. Cheng, and H. Lu, "Gesture recognition based on deep deformable 3D convolutional neural networks," *Pattern Recognition*, vol. 107, no. 1, pp. 107416-107431, 2020.
- [21] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," *Advances in Neural Information Processing Systems*, vol. 1, no. 1, pp. 568-567, 2014.

- [22] C. Koo and R. Olesen, "Patent issued for quality control scheme for multiple-input multiple-output (MIMO) orthogonal frequency division multiplexing (OFDM) systems (USPTO 9509378)," *Computers, Networks & Communications*, vol. 60, no. 786, pp. 1-8, 2016.
- [23] Z. Wang, B. Guo, Z. Yu, and X. Zhou, "Wi-Fi CSI based behavior recognition: From signals, actions to activities," *Journal of Engineering*, vol. 56, no. 5, pp. 109-115, 2018.
- [24] M. Seifeldin, A. Saeed, and A. E. Kosba, "Nuzzer: A large-scale device-free passive localization system for wireless environments," *IEEE Transactions on Mobile Computing*, vol. 7, no. 12, pp. 1321-1334, 2013.
- [25] D. Halperin, W. Hu, and A. Sheth, "Predictable 802.11 packet delivery from wireless channel measurements," *Association for Computing Machinery Special Interest Group on Data Communication Computer Communication Review*, vol. 4, no. 20, pp. 159-170, 2010.
- [26] H. Wang, A. Klaser, and C. Schmid, "Action recognition by dense trajectories," *Computer Vision and Pattern Recognition IEEE*, vol. 10, no. 1109, pp. 3169-3176, 2011.
- [27] L. Guo, L. Wang, J. Liu, W. Zhou, and B. Lu, "HuAc: Human activity recognition using crowdsourced WiFi signals and skeleton data," *Wireless Communications and Mobile Computing*, vol. 2018, no. 1, pp. 1-15, 2018.
- [28] Y. Gu, X. Zhang, and Y. Wang, "WiGRUNT: WiFi-enabled gesture recognition using dual-attention network," *IEEE Transactions on Human-Machine Systems*, vol. 52, no. 4, pp. 736-746, 2021.
- [29] K. Qian, T. Koike, K. Yoshiuchi, B. W. Schuller, and Y. Yamamoto, "Can appliances understand the behaviour of elderly via machine learning? A feasibility study," *IEEE Internet of Things Journal*, vol. 8, no. 10, pp. 8343-8355, 2021.
- [30] B. Zhou, W. Wei, and H. Wu, "Vehicle re-identification method based on global and local feature fusion," *Journal of Physics: Conference Series*, vol. 2005, no. 1, pp. 1173-1178, 2021.
- [31] Y. Ren, "Human motion analysis and fall prevention detection technology based on SVM," *Information Technology*, vol. 43, no. 6, pp. 56-68, 2019.
- [32] H. Fan, "Application of SVM classification algorithm based on convolutional neural network in image classification," *Science and Technology Bulletin*, vol. 38, no. 8, pp. 24-28, 2022.
- [33] S. Eman, E. Nada, and S. Amany, "Utilizing deep learning models in CSI-based human activity recognition," *Neural Computing and Applications*, vol. 34, no. 8, pp. 5993-6010, 2022.
- [34] T. Zhang, S. Fan, J. Hu, X. Guo, Q. Li, Y. Zhang, and W. Aziguli, "A feature fusion method with Guided training for classification tasks," *Computational Intelligence and Neuroscience*, vol. 2021, pp. 6647220-6647235, 2021.
- [35] H. Su, W. Qi, J. Chen, and D. Zhang, "Fuzzy approximation-based task-space control of robot manipulators with remote center of motion constraint," *IEEE Transactions on Fuzzy Systems*, vol. 30, no. 6, pp. 1564-1573, 2022.
- [36] K. Qian, "A comparative study on human action recognition using multiple skeletal features and multiclass support vector machine," *Machine Learning and Applications: An International Journal*, vol. 5, no. 1, pp. 1-15, 2018.



Yuebin Song is a postgraduate who is currently studying at the Qingdao University of Science and Technology. At present, the main research directions are deep learning and human behavior recognition.



Chunling Fan is a Professor, Master's Supervisor, Director of the Laboratory Branch of China Instrument Society, and the member of the Shandong Engineering Education Professional Certification Expert Group. In November 2004, she received her Ph.D. degree in precision instruments and machinery from Shanghai Jiaotong University. From February to March 2019, she was a Visiting Scholar at Taiwan Yishou University. Her research interests include multiphase flow sensing and fluid flow, depth learning, machine vision, and information processing.