Research Article

Dawei Zhang*

Behavior recognition algorithm based on a dual-stream residual convolutional neural network

https://doi.org/10.1515/jisys-2024-0239
received April 29, 2024; accepted July 22, 2024

Abstract: In the process of behavior recognition, the recognition operation may be carried out in various environments such as sunny, cloudy, and night. Since traditional recognition algorithms are judged by identifying the pixels of the image, the intensity of the light will affect the image. The brightness and contrast of the display thus interfere with the recognition results. Therefore, traditional algorithms are easily affected by the lighting environment around the recognition object. To improve the accuracy and recognition rate of the behavior recognition algorithm in different lighting environments, a convolutional neural network (CNN) algorithm using a dual-stream method of time flow and spatial flow is studied here. First, we collect behavioral action data sets and preprocess the data. The core of the behavior recognition algorithm of the dual-stream residual CNN is to use the time stream and the spatial stream to fuse behavioral features and eliminate meaningless data features. After processing Perform feature selection on the data, select the acoustic wave and light-sensing features of the data, and finally, use the extracted features to classify and identify using the two-stream residual CNN and the traditional behavior recognition method. The behavior recognition algorithm based on the dual-stream residual CNN was tested on the data of four groups of people. For the behavioral feature map with a data volume of 50, the behavior recognition algorithm of the dual-stream residual CNN was effective in various environments under different lighting conditions. The recognition accuracy can reach 83.5%, which is 12.3% higher than the traditional. The behavior recognition algorithm of the dual-stream residual CNN takes 17.25 s less than the conventional recognition algorithm. It is concluded that behavior recognition based on dual-stream residual CNNs can indeed improve the recognition accuracy and recognition speed in environments with different lighting conditions than traditional behavior recognition.

Keywords: dual-stream residual convolutional neural network, behavior recognition, feature fusion, grayscale processing, deep learning

1 Introduction

In the information age, behavior recognition technology plays a vital role in enhancing the interactivity of intelligent systems. Advances in computer vision and deep learning have opened up new opportunities in areas such as intelligent monitoring, security, health monitoring, and human–computer interactions. However, the existing behavior recognition algorithms still face challenges in terms of accuracy, real-time performance, and adaptability to environmental changes. These challenges include the algorithm's lack of robustness under different lighting conditions and complex backgrounds, as well as its limited ability to

*** Corresponding author: Dawei Zhang,** College of Information Engineering, Liaodong University, Dandong, 118000, Liaoning, China, e-mail: zdw@liaodongu.edu.cn

respond in real time when dealing with highly dynamic behavior. To improve the performance of the algorithm, researchers are exploring more efficient data processing methods and advanced model architectures to achieve more accurate behavior analysis and faster recognition speed. In addition, the generalization ability of the algorithm is also the focus of the research, and the purpose is to make the algorithm run stably in various practical environments to improve the overall efficiency and user experience of the intelligent system.

In the process of designing the dual residual convolutional neural network (CNN) recognition algorithm, it is necessary to prepare the corresponding test equipment, including high-definition cameras and stopwatch computers. These devices provide the necessary hardware support for data acquisition and time recording. Then, the original information is collected, including action video and sound signal. The acquired original data will then go through a series of pre-processing operations, such as clipping, segmentation, denoising, and grayscale transformation, to improve the quality of the data and reduce the complexity of subsequent processing.

The pre-processed data will be input into the two-flow residual CNN model for training and processing. The core innovation of the model is that it can simultaneously process the time and space information flow, and extract the time-series feature and the space feature of the action through two parallel subnetworks. The fusion of these two subnetworks using residual connection not only deepens the network structure but also helps to solve the problem of gradient disappearance in deep network training, thus improving the efficiency of model training and the speed of behavior recognition.

In terms of feature selection, before feature extraction, the algorithm will first fuse multiple features, integrate different behavior feature data, and then select the behavior feature data with the least difference to achieve more accurate recognition. This method not only improves the ability of feature expression but also enhances the ability of the model to recognize different behavior features.

Finally, the valid information identified by the two-flow residual CNN system model will be classified and summarized in a table for easy analysis and evaluation. The evaluation process of the model includes the evaluation of recognition accuracy and speed, which are key indicators to measure the performance of the algorithm. Compared with traditional behavior recognition methods, double-flow residual CNNs show significant advantages in feature fusion and parallel processing, which can adapt to different lighting environments and improve the accuracy and speed of behavior recognition.

The structure of the article is as follows: Section 1 is the introduction, introducing the background and research motivation of behavior recognition technology; Section 2 is related work, which reviews the existing techniques and methods in the field of behavior recognition. Section 3 describes in detail the design and implementation of the proposed two-flow residual CNN model, including key steps such as data preparation, model construction, feature selection, and fusion. Section 4 is the experiment section, which verifies the effectiveness of the proposed algorithm and shows its advantages by comparing it with the traditional method. Finally, in the conclusion section, the research results are summarized, and the future research direction is prospected.

2 Related work

With the popularization of informatization, behavior recognition is used more widely in life. In terms of intelligent monitoring and security, human behavior in the environment is automatically detected and identified. The system can automatically detect abnormal behavior and provide early warning. In smart homes and smart offices, behavior recognition technology [1] can automatically adjust environmental parameters according to the user's behavior pattern, improving the quality of life and comfort. The widespread use of behavior recognition makes people's lives more convenient. The traditional behavior recognition method is to use sensors to transmit information and use machine learning or pattern recognition methods to conduct behavior detection and analysis. This method uses the data transmitted by the sensor to change with the environment. Changes in data lead to inaccurate data, and the transmission rate is slow, which cannot meet the current real-time requirements. To use machine learning algorithms, it is necessary to first learn machine behavior, and then make judgments on behavior recognition through the learned behavior recognition library. Therefore, traditional behavior recognition methods cannot meet daily needs in terms of recognition accuracy or speed. Many scholars have studied the optimization of behavior recognition algorithms. For example, Jia proposed a behavior recognition algorithm based on global frequency domain pooling (GFDP) [2]. The characteristic of this algorithm is that it proposed a behavior recognition algorithm based on GFDP and introduced a convolutional layer. The batch normalization strategy is extended to the fully connected layer of the behavior recognition model with ERB (efficient residual block)-Res3D as the skeleton, and the behavior type of the object is judged by identifying the skeletal shape of the character. Some people have also proposed a behavior recognition method using ultrasonic signals [3]. This method takes advantage of the fact that ultrasonic waves propagate quickly in air and are not easily interfered with. Ultrasonic waves are used as a continuous sound source to conduct to the human body. According to the collision area of the human body movement, the sound waves are returned. Through the computer's sound wave processing algorithm, an object motion trend model is obtained, and the behavior of the object is judged based on the object's motion trajectory trend. In the process of identifying behaviors, these two behavior recognition algorithms [4] have higher recognition speed and accuracy than traditional recognition algorithms, and the behavior recognition effects of these two algorithms are not affected by the light intensity as compared to conventional algorithms. Affected greatly by illumination, the behavior recognition algorithm and ultrasonic recognition algorithm based on GFDP can indeed improve the accuracy and recognition efficiency of traditional behavior recognition, which can meet daily needs. However, there are also shortcomings to varying degrees. GFDP algorithm recognition is to extract and identify spatial vector features, thereby outlining the behavioral skeleton of the detection object. If the space where the detection object is located is cluttered, it will take longer to identify spatial vector features, and the recognition accuracy will become worse. The behavior recognition algorithm of ultrasonic signals mainly uses the conduction of acoustic signals. If the external sound is too high, it will confuse the transmission of acoustic waves, thus affecting the accuracy of recognition.

The behavior recognition algorithm studied by dual-stream residual CNN models temporal and spatial information through CNNs [5], which can better capture the dynamic and static characteristics of behavior. Dual steam refers to one processing time data stream and one processing spatial data stream, and each stream can have different architectures and parameter settings to better adapt to different environments. Within each flow, residual connections are typically used [6] to construct deeper networks. This connection method can help the network train and optimize more easily, as well as better learn the features of the data. The method of deep learning can perform feature recognition, automatically identify detected and non-detected objects, segment irrelevant data, and preserve detected data. Choosing appropriate features for extraction in complex environments and fusing the extracted features can effectively improve the efficiency of behavior recognition in different environments. The two-dimensional CNN framework based on video plant recognition in the dark [7] confirms that the dual-stream residual CNN model also has high accuracy in recognition under different lighting conditions, greatly reducing the interference of the surrounding environment on detection data.

This kind of neural network uses different data streams to perform convolution operations. The algorithm is widely representative of pattern learning, and its application in daily life generation is the most popular and practical algorithm among all learning patterns. It has become a hot application in the fields of image recognition and speech analysis. In the field of human behavior recognition, there have been many new developments based on CNNs. Some people have added temporal information to the traditional CNN to form a three-dimensional CNN image [8]. The grayscale average, vertical angle and horizontal distance gradient, and vertical and horizontal optical flow information are used as multiple channels for information transmission. For many consecutive frames, the neural convolution system operates to calculate the features of video data in both temporal and spatial dimensions. Some people have proposed using a dual-stream system for preprocessing CNN models, using high- and low-resolution video image data as inputs and outputs for two CNN models, respectively. Finally, data fusion is implemented in two processing layers [9] to achieve the final feature description of the video for final recognition results. The dual-stream residual CNN model is widely used in future applications.

For the design of the dual-stream residual CNN recognition algorithm, relevant testing equipment, high-definition cameras, stopwatches, and computers are first prepared. After preparing these, the required raw information is collected, and then the collected data are subjected to a series of preprocessing operations such as cropping, segmentation, denoising, and grayscale transformation. The model is trained and processed, and effective information is recognized using a dual-stream residual CNN system model. The results are classified and summarized in Table 1. Finally, the model is evaluated. One of the innovative points of the neural network behavior algorithm and traditional behavior recognition is that the algorithm fuses multiple features before feature selection, integrating multiple behavioral feature data. In the process of selecting features, the behavioral feature data with the smallest difference is selected to achieve accurate recognition. The second is to process spatial and temporal information separately through two parallel subnets, connecting the two subnets in a residual manner to improve the speed of deep neural training and accelerate the speed of behavior recognition.

Table 1 summarizes the literature in related work.

Table 1: Literature overview of related work

Document number	Method/technique description	Application field/environment
Liu [1]	Behavior recognition technology based on CNN	Educational environment, classroom behavior monitoring
Jia et al. [2]	GFDP and the convolutional layer are used for behavior recognition	Multi-environment, improve the robustness of behavior recognition
Yang and Zhang [3]	Human behavior is detected and analyzed by ultrasonic signal	Multi-environment, especially suitable for noise interference environment
Wang et al. [4]	The YOLOv5x algorithm is applied to poultry behavior recognition	Agricultural environment, behavioral analysis of chicken
Chen et al. [5]	Image segmentation based on nonlocal information and subspace fuzzy C-ordered mean clustering	Image processing domain
Liu et al. [6]	The resolution of the noise image is reconstructed by using a dense residual connection U-shaped network.	Image denoising and enhancement
Du et al. [7]	A two-dimensional CNN for plant recognition in a dark/light environment	Dark/light environment, plant identification
Yang et al. [8]	X-ray fluorescence method for assessing heavy metals in soil, exceeding standard	Soil pollution analysis, especially rapid detection and assessment of heavy metal content
Zhang et al. [9]	Method of integrating multi-source data to evaluate bridge technical condition index	Bridge engineering, for real-time monitoring and evaluation of bridge technical condition

3 Methods

3.1 Preparation and collection of data

These data are captured using a handheld camera, and four testers who match their height and weight are selected and divided into numbers 1 to 4. They are then placed in a confined space under bright, normal, weak, and dark conditions to perform four types of exercise: walking, running, jumping, and squatting. Figure 1 shows the squatting behavior of number 1 under four different lighting environments. The above operation is repeated for the remaining numbers; high-definition cameras are used to capture motion information, and sound wave sensors are used to collect motion information during the movement. The devices used are Asus laptops, with hardware information: CPU (Central Processing Unit) processor model i7-9688H, graphics card model GTX3060, 8G, cuda v8.0, and industrial shooting cameras and millisecond level timers produced by Hikvision. Finally, according to the timer, the time used for each experiment is recorded. Each type of behavioral motion takes 12 videos, each lasting 10 s. After preprocessing operations such as video splitting, cropping, grayscale, and denoising, 210 video frames are formed by filtering out images with poor quality. Each video segment is 3 s long, and all videos are adjusted to a resolution of 1,024 × 680, with a conversion processing rate of 10 frames/s. The collected image and video data are manually labeled, with the target labels being walked, run, squat, and jump to generate label files in the format of text. Pycharm is used to divide label files evenly using classification functions.

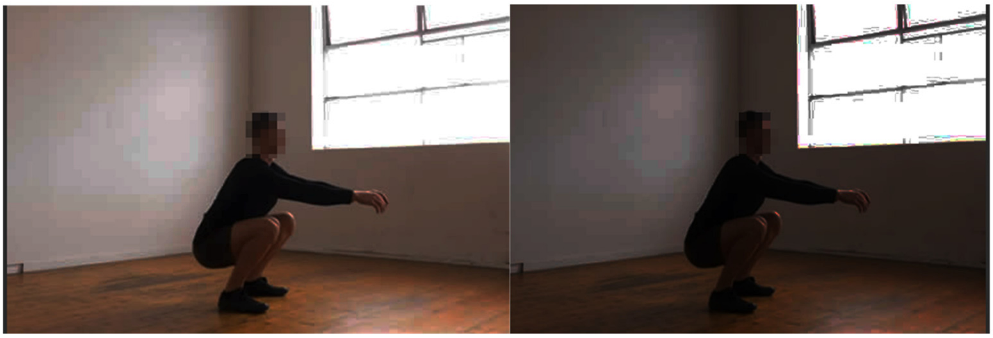

Figure 1: Number 1 squatting in four different lighting environments. Source: Youth Pitchers (Release Posture).

To ensure the adaptability of the algorithm to different user behavioral characteristics, the collected data set covers a diverse sample of behaviors, including individuals of different ages, genders, and physical conditions, thus training a more generalized model.

In the process of data collection, the standards of privacy protection and data security are strictly observed. All individuals involved in data collection provide informed consent, ensuring lawful use of the data. In addition, encrypted storage and secure transmission measures are in place to protect the privacy of participants.

3.2 Image preprocessing

3.2.1 Cutting

The video is cropped to remove unnecessary parts of the image, such as borders or irrelevant backgrounds. Cropping can be done based on image content or predefined regions, and a large amount of image data often increases the amount of data processing. By cropping irrelevant backgrounds, the detection rate can be greatly accelerated while also preventing interference from other background factors, as shown in Figure 2.

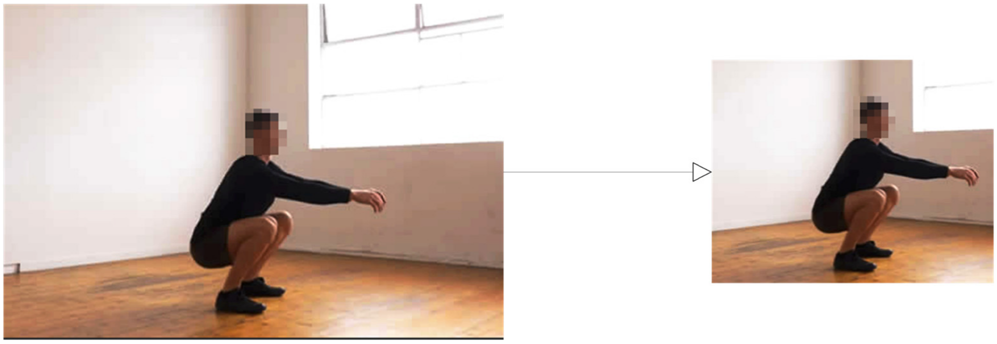

Figure 2: Cropping at the edges of the image. Source: Youth Pitchers (Release Posture).

3.2.2 Grayscale processing

The detection of video image information or video data in this article is mostly based on the three primary colors: red, green, blue, and RGB. Excessive information from the three primary colors increases the input of image data to the computer, resulting in a significant increase in the amount of calculated data, making the data prone to anomalies and color loss. Moreover, the clearer the image, the more pixels there are, and the more significant the increase in data volume. However, in the process of dual-stream CNN behavior recognition, the information recognized by the algorithm is extracted through the detection of the target limb space vector, and the color depth of the image does not affect the extraction, which is considered invalid information. Therefore, the three-dimensional RGB color space can be reduced to a one-dimensional grayscale space through the grayscale processing of images [10]. This not only reduces the processing load of neural networks on data but also reduces the interference of irrelevant pixels. The schematic diagram of the image grayscale processing results is shown in Figure 3:

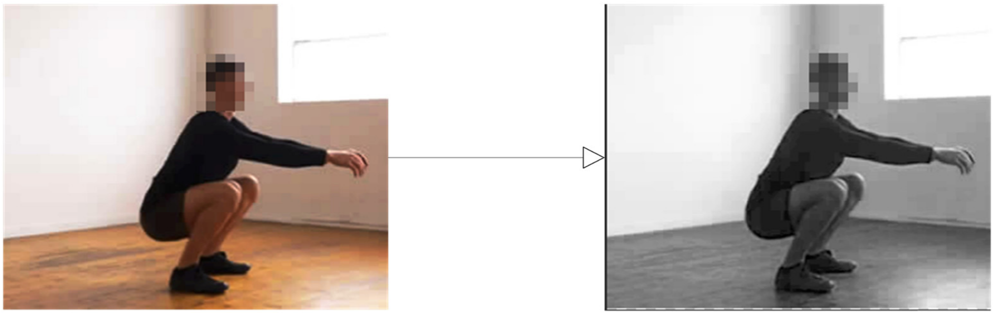

Figure 3: Images before and after grayscale processing.

3.2.3 Image thresholding

The main purpose of image thresholding is to further compress video images to reduce data volume, remove unnecessary pixels, and avoid the background color of the processed grayscale image being the same as the recognized object, resulting in biased results. The overall motion framework of behavioral athletes is extracted, and the threshold calculation formula is as follows:

$$T = T_1[x, y, f(x, y), p(x, y)], \tag{1}$$

where $f(x, y)$ represents the horizontal pixel point, (x, y) represents the grayscale value, $p(x, y)$ represents pixels in the vertical direction, and T_1 represents the grayscale gradient value. How to determine accurate threshold information is a problem in the recognition process. If the threshold is set too high, other background images would be included in the detection image of the moving object, increasing the calculation value of the motion behavior recognition algorithm or a loss of some target motion images. Both of these situations may cause errors in subsequent action estimation. The threshold determination method used in this article [11] is the three-dimensional minimum entropy method, and the specific steps are as follows.

Assuming that the behavioral motion image has L grayscale levels after grayscale processing, and the image resolution is $M \times N$, the grayscale probability distribution function is

$$P_i = \frac{n_i}{M \times N}. \tag{2}$$

If the boundary value of background pixels is T, the probability function of the grayscale distribution of the background is

$$W_1(t) = \sum_{i=0}^{T} P_i. \tag{3}$$

The probability function of the grayscale distribution of behavioral athletes is

$$W_2 = \sum_{i=T+1}^{L-1} P_i. \tag{4}$$

According to Formulas (3) and (4), the entropy values of background pixels and the entropy values of behavior movers can be obtained as follows:

$$H_1 = -\sum_{i=0}^{P_i} \frac{P_i}{W_1(t)} \times \log\left(\frac{P_i}{W_1(t)}\right), \tag{5}$$

$$H_2 = -\sum_{i=T+1}^{L-1} \frac{p_i}{W_2(t)} \times \log\left(\frac{p_i}{W_2(t)}\right). \tag{6}$$

Here, H_1 is the entropy value of background pixels, and H_2 is the entropy value of behavioral motion.

Based on the information entropy of background and behavioral movements, the information entropy value of the entire video can be obtained:

$$H = H_1 + H_2. \tag{7}$$

After obtaining the determined threshold, the image data can be divided based on the threshold to highlight the behavior of the movers from the image. The image thresholding effect is shown in Figure 4.

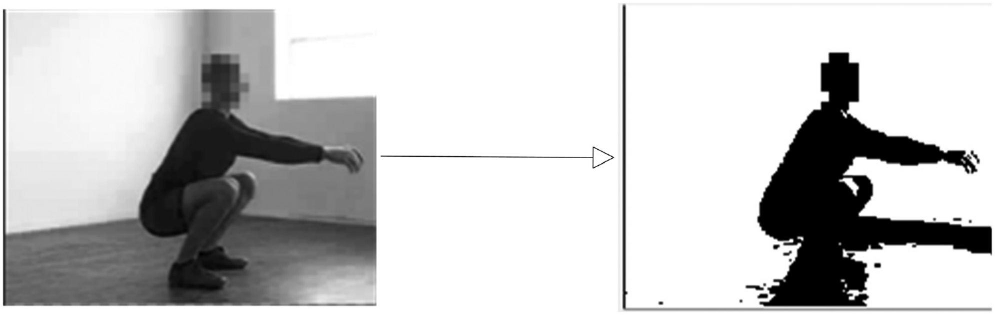

Figure 4: Images before and after thresholding.

3.2.4 Noise reduction

Removing external noise during the transmission of sound wave information can be transmitted to the receiver in the form of sound waves, which can confuse the measured data and easily cause anomalies in the recognition data. The commonly used denoising methods include Gaussian filtering [12], median filtering, and bilateral filtering. A two-dimensional Gaussian kernel of $n \times n$ is defined, with the center of the kernel being a two-dimensional Gaussian function:

$$G(x, y) = \frac{1}{2\pi\sigma^2} e^{\left(-\frac{x^2+y^2}{2\sigma^2}\right)_{\times N}}. \tag{8}$$

Here, (x, y) represents the offset of the kernel center, and σ is the standard deviation of the Gaussian distribution. Each element value in the Gaussian kernel represents the weight at that position, and the larger the weight, the greater the impact of the corresponding pixel at that position during the filtering process. The

Gaussian kernel is convolved with the image, and a filter is applied to each pixel in the image [13]. For each pixel position (i, j), the convolution operation is used to weight-average the Gaussian kernel with the surrounding pixels in the image to obtain the new pixel value at that position in the output image. The specific processing effect is as follows.

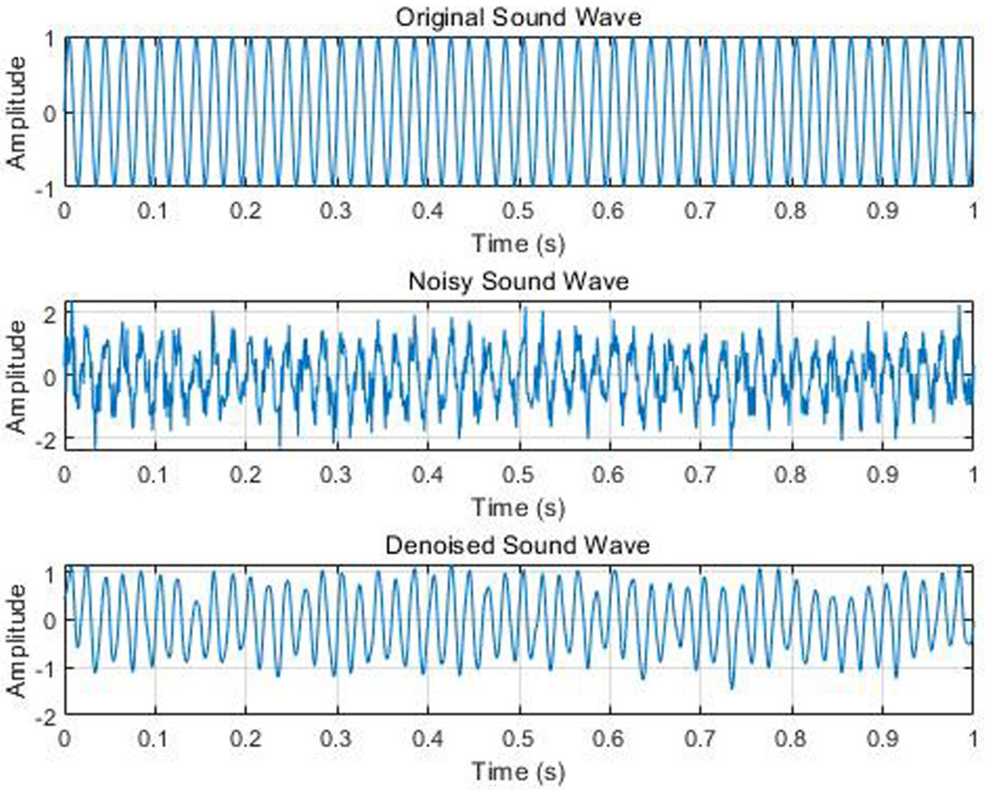

Figure 5: Sound wave information for denoising processing.

In Figure 5, the original sound wave represents the most primitive data; the noisy sound wave represents data that have been affected by noise interference; and the denoised sound wave represents the data that have been denoised. The original data are subjected to external interference to form a noisy sound wave image, which is denoised through median filtering. Although the processed sound wave data have some differences, the frequency and wavelength of the sound wave have not changed significantly.

To evaluate the impact of noise on the accuracy of behavior recognition, this study will use a quantitative analysis method to compare the results of behavior recognition under different noise levels. Based on this analysis, non-local mean filtering is used to reduce the negative impact of noise on model performance and to ensure high accuracy even in complex environments.

3.3 Constructing dual-stream residual CNN models

The dual-stream residual CNN mainly extracts information through spatial and temporal streams. First, the structures of spatial and temporal flows are defined. For spatial flow, 2D convolutional layers are usually used to process static image frames to extract spatial information from the image. For time flow, 3D residual layers are usually used to process video sequences to extract temporal information from the video. In each flow, residual modules are used to deepen the network structure, better train the deep network, and alleviate the problem of gradient vanishing. A typical residual module consists of multiple convolutional layers, where each convolutional layer is followed by an identity mapping [14], and the input is added to the output using skip connections. After constructing the residual module, the two streams are fused with features [15]. Finally, the merged features are fed into the fully connected layer, and appropriate activation and loss functions are selected based on the specific task. Figure 6 shows the structure of the neural network model.

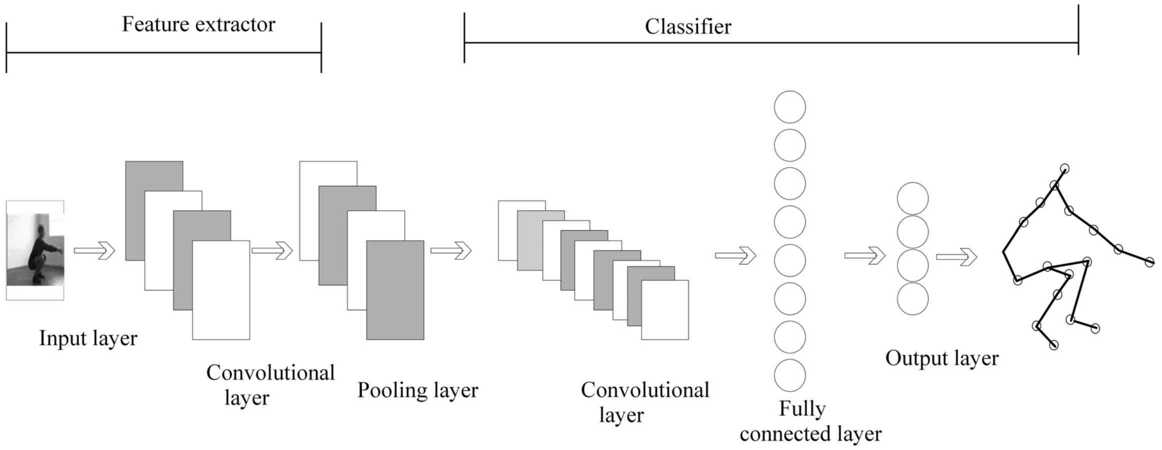

Figure 6: Algorithm structure of CNN model.

When designing a two-flow residual CNN, power consumption can be optimized through network simplification, parameter quantification, model pruning, dynamic resource adjustment, and hardware acceleration to accommodate mobile devices and enhance the user experience.

In the construction of the model, a lightweight design is adopted, an efficient network structure is selected, the number of parameters is reduced, and the model complexity is reduced by model compression and quantization techniques to adapt to resource-constrained devices. This helps to reduce computational requirements and memory footprint while maintaining recognition accuracy, improving the practicality and efficiency of the model on mobile devices.

When constructing the two-flow residual CNN model, regularization technique, data enhancement, and adaptive learning rate adjustment are considered to improve the generalization ability and robustness of the model in long-term tasks [16,17].

3.4 Feature selection

The differences in various behaviors and movements ultimately stem from the different angles between the joints during exercise and are not fundamentally related to the height, shortness, obesity, or thinness of the athletes. Therefore, feature extraction of behavioral movements is the extraction of neural node model features between human limbs in various postures [18]. This article adopts the method of constructing a dual-stream residual CNN model, dividing the human body model into 20 neural nodes. FAST feature selection is performed as shown in Figure 7:

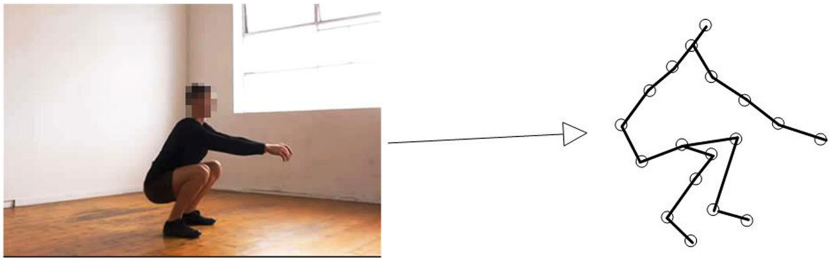

Figure 7: Performing neural node extraction.

The normal squatting neural node model accurately identifies the effective information in the data based on the collected image and acoustic information as shown in Figures 8 and 9, and presents it with a visual box.

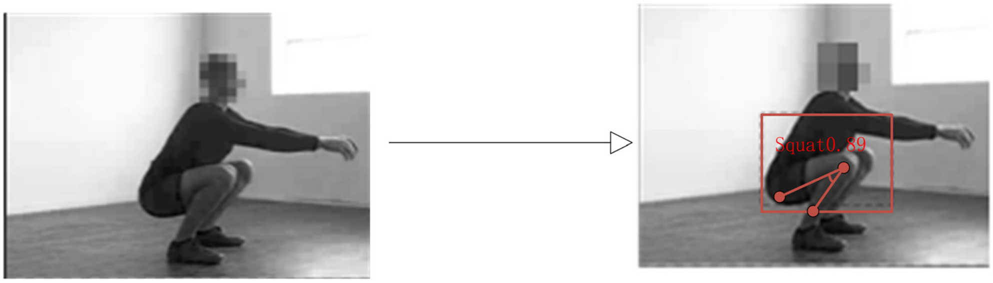

Figure 8: Extracting image information using visual boxes.

Figure 8 shows the extraction of the target area for squatting, which is the angle degree between the calf and thigh. Through feature extraction, it can be seen that the angle is less than 90°. By extracting key nodes, it can be determined that the behavior is squatting.

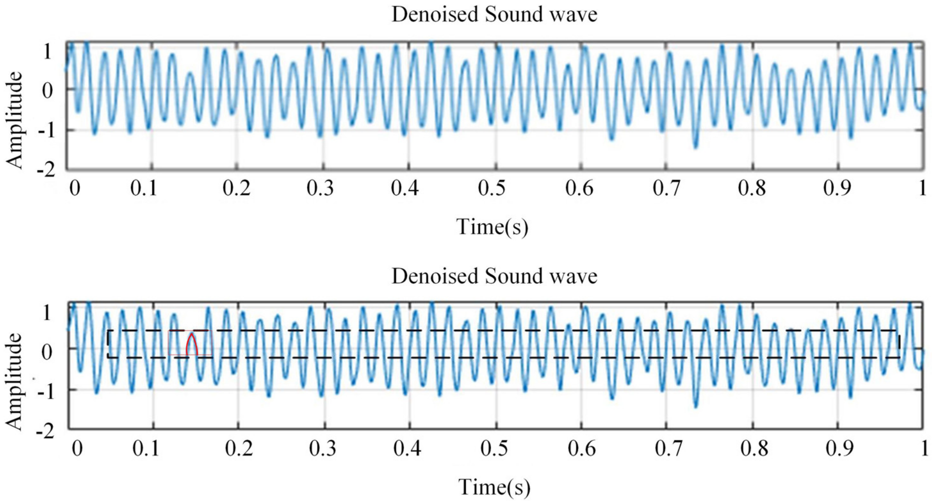

Figure 9: Extracting effective information about sound waves using visual frames.

Figure 9 is used to extract features from the sound wave data. It can be seen that a wave cycle occurs every 0.1 s. The frequency of the wave is 10. The figure shows that the maximum wavelength is 1, and the minimum is −1. It can be seen that human body conduction occurs through waves, and the upper and lower values are the equilibrium values. This shows that the behavior is moving in the horizontal direction and then passes through the minimum value of the wave, which is marked with a red mark in the figure. The angle between the two limbs of the human body can be calculated through the frequency of the minimum wave, and the angle is determined by the degree of the angle.

The video and sound data obtained through preprocessing are processed using a dual-stream residual CNN model to extract the position of neural nodes, determine their positions, and calculate the relative distance before the neural nodes. Moreover, a smooth, straight line is used to connect and determine the size of the angle between key nodes. The extracted squatting behavior is transformed into a neural node diagram, as shown in Figure 10.

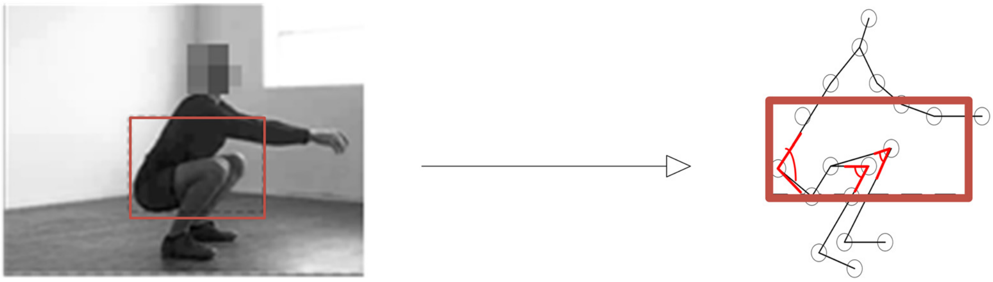

Figure 10: Extracting the angle between key nodes.

The template information of a behavioral action is represented by the neural node distance matrix [19], which becomes the template information matrix. The specific manifestation is as follows:

$$[\nabla d_{1.2}^{1} \ldots \Delta d_{k(k-1)}^{1}] = M, \tag{9}$$

where k represents the number of neural nodes extracted for this behavior feature, and the maximum number of neural nodes is 18; d represents the number of video frames in the behavior feature template; and 1 to K is the distance between each neural node.

In the feature selection phase, consider incorporating multi-task learning into the model design, learning common features across tasks by sharing the initial layer of the network while customizing the output layer for each task to enhance the generalization and efficiency of the model.

To enrich the feature representation and improve the recognition accuracy, the feature extraction method based on orthogonal polynomials is adopted in this research. These techniques include fast and accurate calculation of high-order Tchebichef polynomials and fast overlapping block processing algorithms, which can extract strong descriptive features from behavior data and provide richer information for behavior recognition [20,21].

3.5 Design of CNN model

First, in the dual-stream residual CNN database, templates for various behaviors are inputted, and through convolutional neural algorithms, neural node templates are formed one by one. Then, the neural nodes of the preprocessed test image data frames [22] are extracted one by one, and the extracted neural nodes are compared with the behavior templates of the neural network database. After selecting the node with the smallest difference and determining its position, the distance between each node is calculated to form a distance vector for the neural network nodes [23]. The specific flowchart is shown in Figure 11.

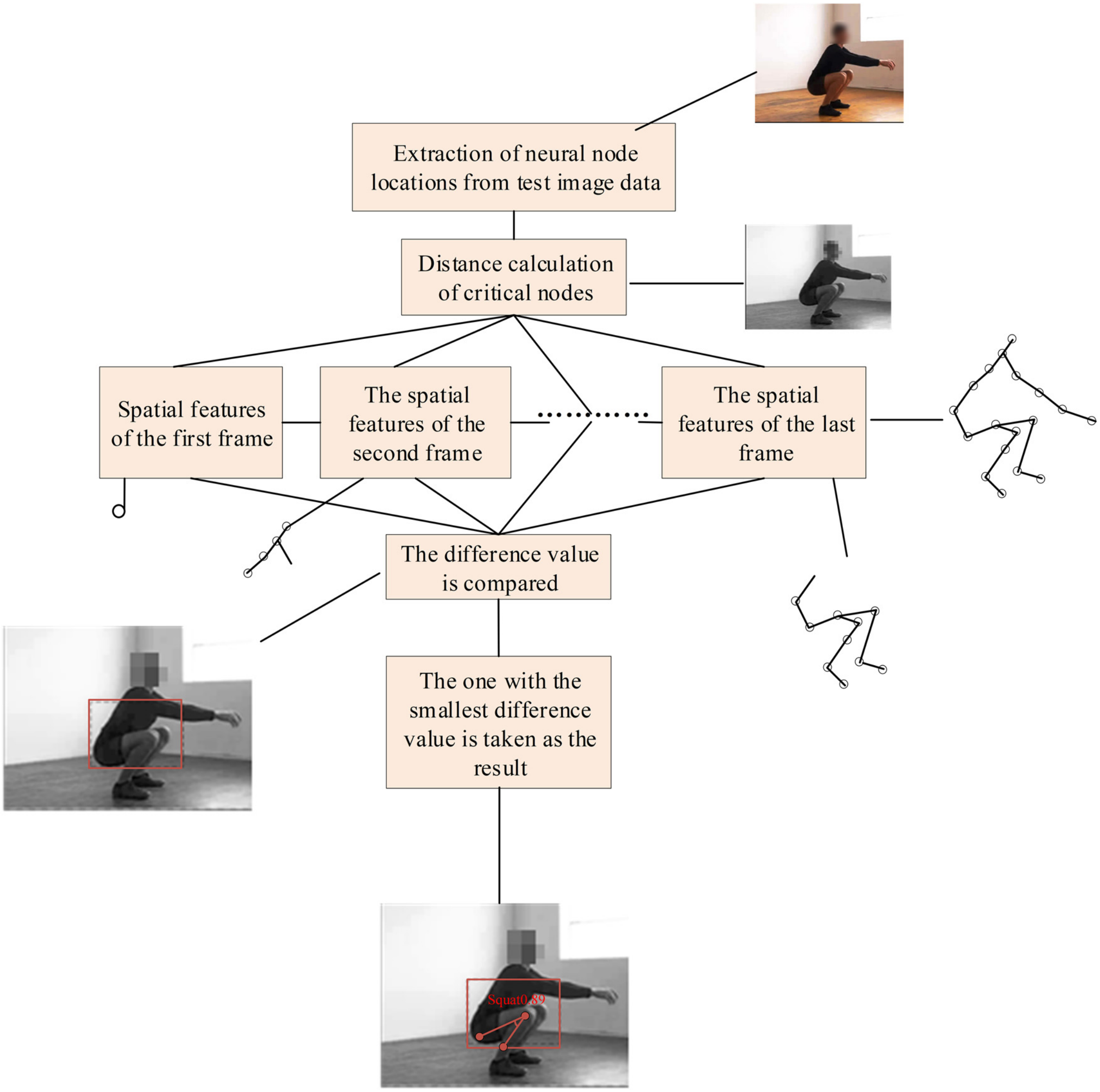
Extraction of neural node locations from test image data
Distance calculation of critical nodes
Spatial features of the first frame
The spatial features of the second frame
The spatial features of the last frame
The difference value is compared
The one with the smallest difference value is taken as the result
Squat0.89

Figure 11: Workflow of residual neural network algorithm.

The formula for calculating the difference value is

$$D_X = \frac{1}{i = 1 \cdots n} \min \left\{ \frac{\sum_{j=1}^{K} (t - m_{i,j}^{K})}{k(k-1)/2} \right\}. \tag{10}$$

Here, $m_{i,j}^{K}$ represents the behavior information of the Kth frame data in row J and column I of the matrix [24]. The spatial vector graph of the identified image data neural node limbs can be used to calculate the difference values with the behavior template information in the database, and the calculated various difference values can be compared. If the difference value is smaller, it indicates that the behavior recognition pattern is more similar to the database template pattern. If the difference value between a certain frame image and the template image is the smallest among all the differences, then the recognized behavioral characteristics of that frame image are the result of test recognition.

3.6 Pseudo-code

To enable the reader to understand the proposed algorithm more clearly, this article provides the pseudo-code of the algorithm:

Algorithm: Enhanced two-stream residual convolutional neural network behavior recognition

Require: Training dataset $D = \{(x_i^s,\ x_i^t,\ y_i)\}$, where x_i^s is the spatial stream input, x_i^t is the temporal stream input, and y_i is the label. The following parameters are also used: learning rate η, number of epochs E, batch size B, weight initialization scheme W_{init}, loss function $L(\cdot)$, regularization parameter λ, and momentum μ.
Ensure: Trained model parameters θ.

```
Initialize network parameters:
  $\theta \leftarrow W_{\text{init}}()$//Initialize weights according to chosen scheme.
  $t \leftarrow 0$//Initialize time step counter.
repeat (for each epoch):
  Shuffle dataset $D$.
  for batch $B$ in $D$ do
    Compute forward pass for spatial stream:
    $h_s^l \leftarrow f_s^{l(x_i^s;\theta_s^l)} \forall l \in \{1, \ldots, L\}$//forward pass for each layer l.
    Compute forward pass for temporal stream:
    $h_t^l \leftarrow f_t^{l(x_i^t;\theta_t^l)} \forall l \in \{1, \ldots, L\}$.
    Concatenate spatial and temporal features:
    $h_f \leftarrow [h_s^L; h_t^L]$//Concatenate outputs from last layers.
    Compute prediction $\hat{y}$ using classifier:
    $\hat{y} \leftarrow g(h_f; \theta_g)$.
    Calculate loss:
    $L_{\text{total}} \leftarrow L(\hat{y},\ y_i) + \lambda*||\theta||_2^2$// Total loss including regularization term.
    Backpropagation:
    $\delta \leftarrow \nabla L_{\text{total}}/\partial\hat{y}$//calculate gradient of loss w.r.t. prediction.
    for l in reverse($L$) do
      $\delta \leftarrow \nabla L_{\text{total}}/\partial\theta^l$// backpropagate gradients through layers.
      $\theta^l \leftarrow \theta^l - \eta * \delta + \mu * v^l$// update parameters with momentum.
      $v^l \leftarrow \delta$ // update velocity for next iteration.
    end for
  end for
  $t \leftarrow t + 1$// Increment time step.
until deadline or $t = E$// Stop after reaching deadline or a number of epochs.
// Evaluation phase (after training):
for new input $(x^s,\ x^t)$ do
  $h_s^l \leftarrow f_s^{l(x^s;\theta_s^l)} \forall l \in \{1, \ldots, L\}$.
  $h_t^l \leftarrow f_t^{l(x^t;\theta_t^l)} \forall l \in \{1, \ldots, L\}$.
  $h_f \leftarrow [h_s^L; h_t^L]$.
  $\hat{y} \leftarrow g(h_f; \theta_g)$.
  return $\hat{y}$// Predicted label.
```

The two-stream residual CNN behavior recognition algorithm improves the accuracy and speed of behavior recognition under various lighting conditions by integrating spatial and temporal stream information. The algorithm begins with initializing network parameters and then shuffles the training dataset to ensure data randomness. During each epoch, it performs forward propagation for spatial and temporal streams on each batch of data, computing output features for each layer. The spatial stream focuses on static features of the image, while the temporal stream captures dynamic changes in actions. The features from both streams are then fused to form the final feature representation. Labels are predicted using a classifier, and the total loss, including the regularization term, is calculated. Backpropagation is used to update parameters, with momentum introduced to accelerate convergence. As epochs progress, the network is gradually optimized until the preset number of training cycles is reached. In the evaluation phase, the same forward propagation process is executed on new inputs, and the predicted labels are returned, achieving accurate behavior recognition.

4 Experiments

To verify the feasibility of the proposed algorithm in practical applications, field tests are carried out, the algorithm is deployed in multiple real environments, and behavioral data is collected under different lighting and background conditions. The stability and accuracy of the algorithm in real-world complexity were evaluated by comparing the results of field tests with the performance under laboratory conditions.

The first tester performed 50 walking actions under normal lighting conditions, captured by a high-speed camera, and the stopwatch was used to record the recognition time. Furthermore, the obtained frame images were processed according to the above operations. The processed images were respectively recognized using traditional behavior recognition algorithms and behavior recognition algorithms using dual-stream residual CNNs. The recognition error was marked as ×, and the recognition accuracy was marked as √. Test personnel 2 performed 50 running actions in a normally lit environment; test personnel 3 performed squatting movements 50 times in the same environment; test personnel 4 performed 50 jumps, and the subsequent data processing was the same as the first step. The correct and incorrect numbers were calculated. The test results are shown in Tables 2–4, respectively.

Table 2: Recognition rate of the CNN algorithm

	Walk	Run	Squat	Jump
Correct	45	43	47	45
Error	5	7	3	5
Accuracy	90%	86%	94%	90%
Time (s)	50	45	32	25
Recognition speed (m/s)	1	1.11	1.56	2

Table 3: Recognition rates of traditional behavior recognition methods

	Walk	Run	Squat	Jump
Correct	40	36	40	39
Error	10	14	10	11
Accuracy	80%	72%	80%	78%
Time (s)	60	56	49	56
Recognition speed (m/s)	0.83	0.89	0.98	0.89

The reasons for recognition errors include inaccurate feature extraction due to noise interference, insufficient model generalization ability, insufficient learning of all behavioral features, and inadequate adaptability of algorithms in complex environments.

Using the above table data to compare the behavior recognition algorithm of the dual-stream residual CNN, the ultrasonic sensing recognition algorithm, the GFDP behavior recognition algorithm, and the traditional algorithm all require implementation in a specific behavior recognition task and evaluation. During the evaluation process, the confusion matrix of the model can be obtained, and then the recall rate can be calculated. Figure 12 shows the recall rate of each algorithm.

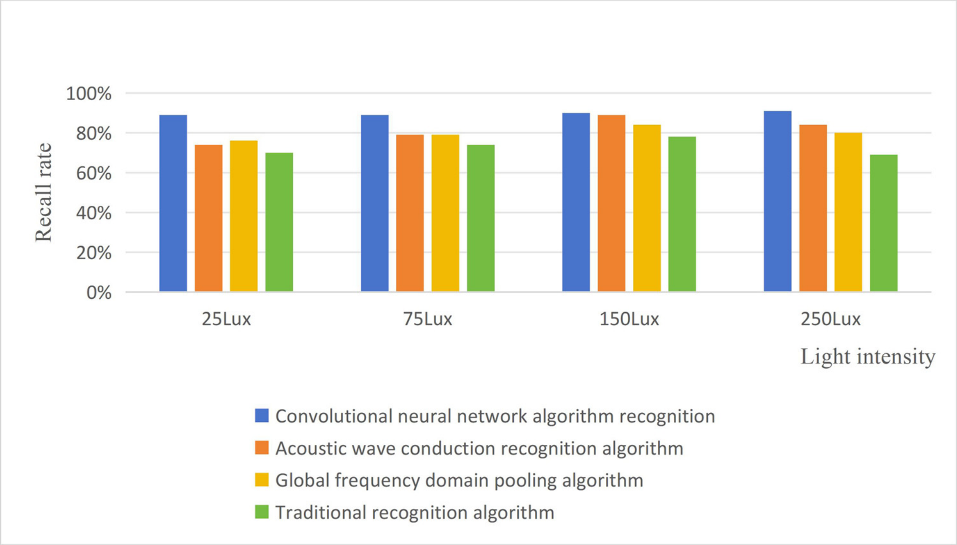

Figure 12: Recall rate of recognition under each algorithm.

From the comparison of recall rates under each algorithm in Figure 12, we can see that blue is the behavior recognition algorithm of the dual-stream residual CNN, orange is the ultrasonic recognition algorithm, yellow is the behavior recognition algorithm based on GFDP, and green is traditional recognition algorithm. Under any light intensity, the recall rate of the other three algorithms is higher than the recall rate of the conventional recognition algorithm. The recall rate is used to measure the model's ability to identify positive examples; that is, the model can correctly predict all true positive examples. Ability, the low recall rate of the algorithm, indicates the low reliability of this method. The behavior recognition algorithm of the dual-stream residual CNN changes with the light intensity, and the recall rate is stable and high at a certain level, indicating that the recall rate of the algorithm is not affected by the intensity of light. The recall rate of the behavior recognition algorithm of the dual-stream residual CNN reaches a maximum of more than 85%, while the recall rate of the behavior recognition algorithm and ultrasonic recognition algorithm based on GFDP is around 80%, and the recall rate of the traditional recognition algorithm is around 70%.

Tester 1 walked 50 times in strong light (250 lux), normal light (150 lux), weak light (75 lux), and dark environment (25 lux). The correct recognition was marked as √, and the recognition error was marked as ×. The number of normal actions was counted. Number 2 performed 50 running actions in four different lighting environments; number 3 squatted 50 times in four different lighting environments; and number 4 jumped 50 times in four different lighting environments. The collected data was recognized and processed using traditional recognition algorithms and behavior recognition algorithms using dual-stream residual CNNs. The correct recognition was marked as √, and the recognition errors were marked as ×. The number of normal cases was counted and listed in a table.

Table 4: Recognition accuracy of different behavior types under different lighting conditions

Behavior	Walk				Run				Squat				Jump			
Illumination intensity	250 Lux	150 Lux	75 Lux	25 Lux	250 Lux	150 Lux	75 Lux	25 Lux	250 Lux	150 Lux	75 Lux	25 Lux	250 Lux	150 Lux	75 Lux	25 Lux
Traditional recognition algorithm	21	41	32	20	18	39	32	21	30	45	39	29	32	45	38	32
Accuracy	42%	82%	64%	40%	36%	78%	64%	42%	60%	90%	78%	58%	64%	90%	76%	64%
Two-stream residual CNN	38	48	42	38	40	47	42	38	46	49	42	40	38	46	37	35
Accuracy	76%	96%	84%	76%	80%	94%	84%	76%	92%	98%	84%	80%	76%	92%	74%	70%

According to Tables 1–3, it can be concluded that the average accuracy of the behavior recognition algorithm of the dual-stream residual CNN for various behaviors in various environments was over 80%, while the average accuracy of traditional recognition methods was less than 75%. This indicates that the behavior recognition algorithm of the neural network has greatly improved recognition accuracy. The traditional recognition algorithm took an average of 55.25 s to recognize and process 50 data points, while the behavior recognition algorithm of the neural network took an average of 38 s, indicating that the behavior recognition algorithm of the neural network can indeed improve the recognition speed of the algorithm.

The above experiments were based on self-built scene recognition and data collection indoors. The experiments were accidental and unconvincing. In one experiment, roadside cameras were used to capture pedestrians on the road, 10 different data were collected, and behavior recognition was performed through a dual-stream residual CNN. The algorithm for identification is given in Figure 13.

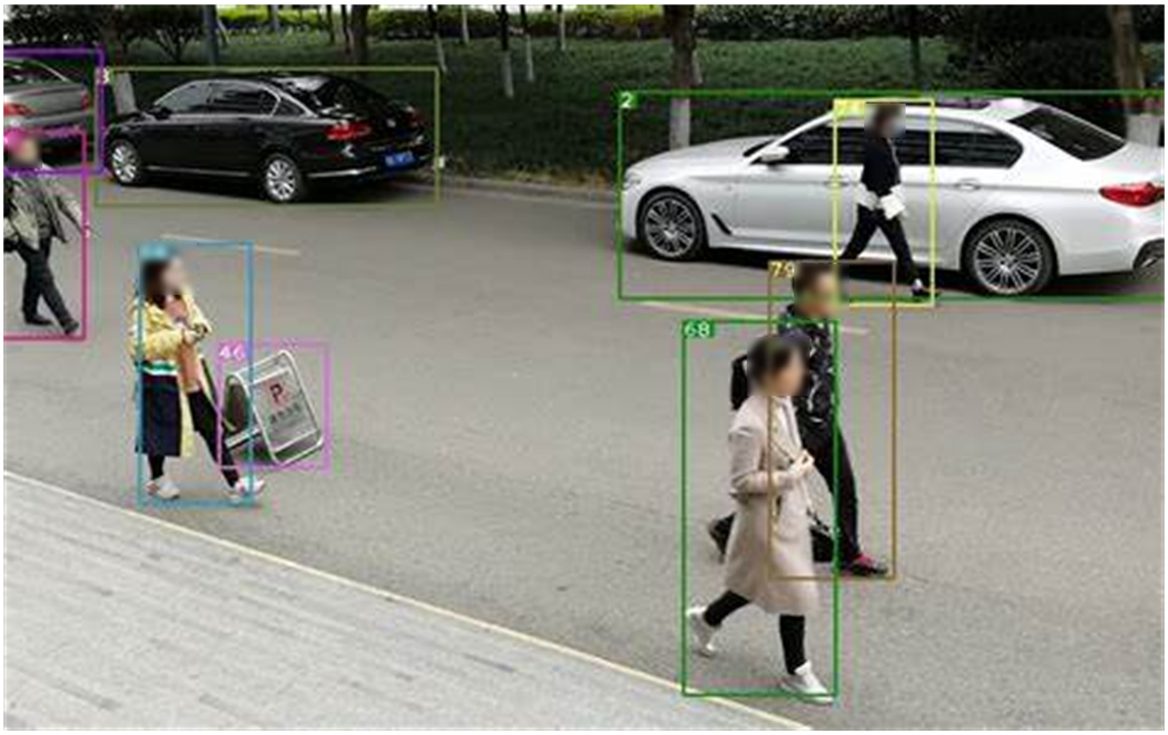

Figure 13: Behavior recognition of outdoor passers-by. Source: Human-to-Human-or-Object Interaction dataset.

There are nine pieces of correct data identified using this method, and multiple human behaviors can be identified in one picture, with an accuracy rate of 90%.

Comparison with recent works:

The behavior recognition algorithm in this study is compared with traditional methods and other recently proposed algorithms [25,26]. The experimental results are shown in Figure 14.

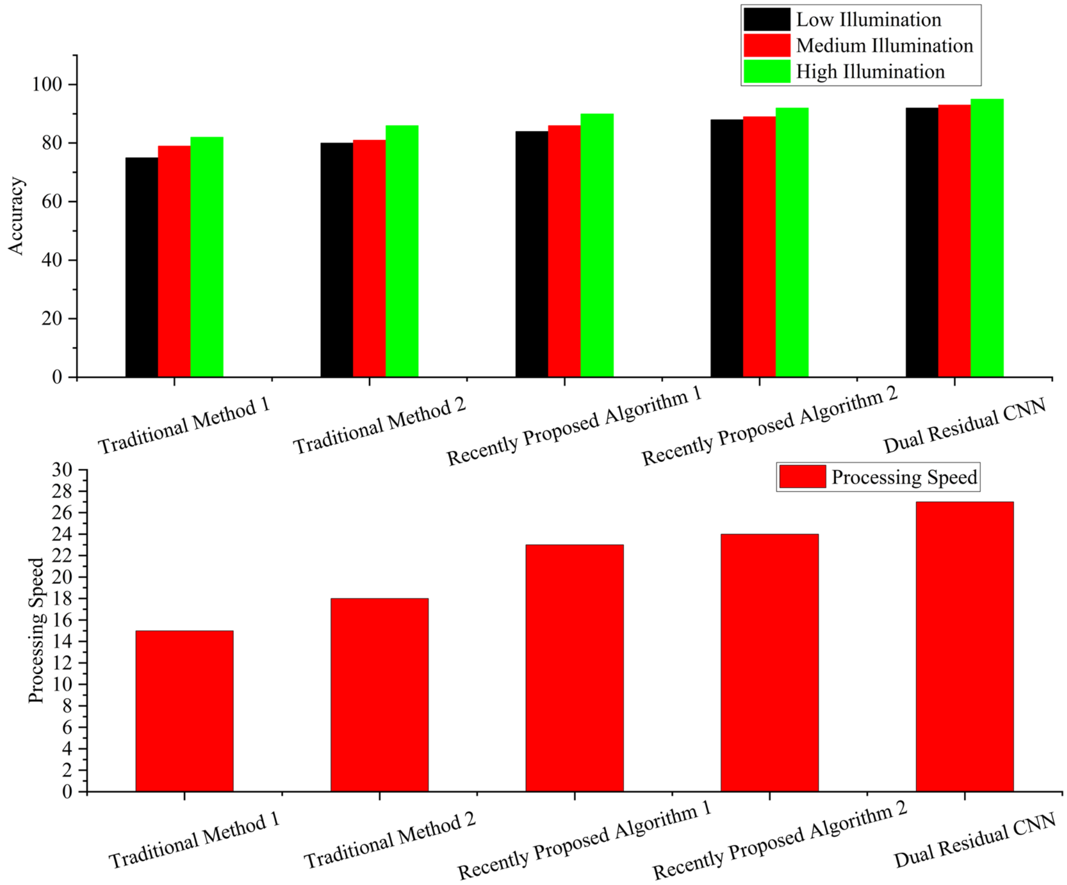

Figure 14: Comparison of experimental results.

By comparing the experimental results, it is found that the dual residual CNN has higher recognition accuracy and faster processing speed under different illumination conditions.

Results: The proposed algorithm has significant advantages in improving the accuracy of behavior recognition, especially under low-light conditions. In addition, the algorithm also shows improvement in processing speed, with the average recognition time reduced by 17.25 s compared with the traditional algorithm. Nevertheless, it is also found in the experiment that the algorithm has certain limitations when dealing with complex backgrounds or similar behavior patterns, which may lead to confusion in recognition. Future work will focus on further optimizing the algorithm to improve its generalization ability and adaptability to meet the needs of a wider range of applications.

5 Conclusions

This article mainly discussed the issue of tardiness recognition speed and insufficient recognition precision of traditional behavior recognition algorithms in daily life and proposed a behavior recognition algorithm based on dual-stream residual CNN to solve these problems. The advantage of the algorithm is that it improves recognition accuracy and speed and adapts to different lighting environments. This article started with the proposal of algorithms, followed by data processing and feature selection, and finally, model training was conducted to obtain the final results. The results showed that the behavior recognition algorithm model of the dual-stream residual CNN was beneficial for improving recognition efficiency and recognition accuracy in various lighting environments. This algorithm can improve recognition performance in various aspects, but there are still some shortcomings. The neural network algorithm ignores appearance information based on the sequence of human neural nodes. If there are two sets of similar behavioral movements in the recognition process, it is easy to cause recognition confusion, and it cannot recognize the two similar behavioral features well. The behavior recognition algorithm of dual-stream residual CNNs has broad application prospects in the future. Due to the improvement of performance in various aspects, the application field is also constantly expanding. In the fields of intelligent driving and unmanned vehicles, this algorithm can be used to recognize the behavior of drivers and passengers, such as making phone calls, fatigue driving, and traffic violations, to improve driving safety and riding experience. It can also be applied in health monitoring and auxiliary diagnosis. By identifying user head and hand movements, this algorithm can achieve gesture recognition and head tracking in virtual reality and augmented reality applications, providing users with a more natural and intuitive interactive experience. The widespread application of dual-stream residual CNNs for behavior recognition provides important support for the development of intelligent systems and the improvement of human life. In future work, model optimization, data set expansion, and hardware acceleration will be explored to enhance stability and reliability in long-term monitoring missions.

Funding information: This research was supported by the Liaoning Provincial Department of Education's Basic Research Project for Universities [Grant No. JYTMS20230711] and Liaoning Province Science and Technology Plan Joint Program (Fund) Project [Grant No. 2023JH2/101700009].

Author contribution: DWZ designed and performed research, analyzed data, and wrote the paper.

Conflict of interest: The authors declare no conflict of interest.

Data availability statement: The datasets analyzed during the current study are available from the corresponding author on reasonable request.

References

[1] Liu L. Design of a student classroom behavior recognition system based on convolutional neural networks. Mod Electron Technol. 2024;47(6):142–6. doi: 10.16652/j.issn.1004-373x.2024.06.023.

[2] Jia Z, Zhang H, Zhang C, Yan M, Chu J, Yan Z. Behavior recognition algorithm based on global frequency domain pooling. Computer Res. 2024;41(9):1–7. doi: 10.19734/j.issn.1001-3695.2023.11.0596.

[3] Yang Y, Zhang X. Behavior recognition method based on ultrasonic signals. Computer Age. 2023;(12):162–6. doi: 10.16644/j.cnki.cn33-1094/tp.2023.12.035.

[4] Wang C, Wang Y, Shang S, Zhang N. Chicken basic behavior recognition method based on YOLOv5x research; 2024. p. 1–6. http://kns.cnki.net/kcms/detail/37.1433.th.20240304.1914.010.html.

[5] Chen Y, Huang C, Qin X, Peng J, Lei H, Zhou L. Image segmentation algorithm based on non local information and subspace fuzzy C-ordered mean clustering. J Computer Aided Des Graph. 2024;1–13. http://kns.cnki.net/kcms/detail/11.2925.tp.20240204.1553.045.html.

[6] Liu P, Li L, Zhang Z, Zhu X, Cheng D. Based on intensive residual connected u-shaped network noise image resolution reconstruction algorithm. Beijing, China: Tongfangzhiwang (Beijing) Technology Co., Ltd; 2024. p. 1–10. doi: 10.13272/j.issn.1671-251x.2023080098.

[7] Du C, He Y, Deng H, Chang S, Wang Y. The two-dimensional convolutional neural network framework is based on video plant recognition in the dark. J Electron Meas Instrum. 2023;37(8):21–9. doi: 10.13382/j.jemi.B2306625.

[8] Yang W, Li Z, Li F, Lv S, Fan J. Research on XRF soil heavy metal exceedance analysis method based on CARS and 1D-CNN combined. Spectrosc Spectr Anal. 2024;44(3):670–4.

[9] Zhang Y, Liang P, Xia Z, Li C, Liu J. A method for evaluating bridge technical condition indicators by integrating multi-source data. Bridge Constr. 2024;54(1):75–81. doi: 10.20051/j.issn.1003-4722.2024.01.011.

[10] Ling S. Research on image grayscale processing method based on HIS model. J Changsha Aviat Vocat Tech Coll. 2023;23(1):33–5. doi: 10.13829/j.cnki.issn.1671-9654.2023.01.009.

[11] Yang B, Ding L. A fuzzy image multi threshold block enhancement method based on partial differential equations. J Hubei Univ Arts Sci. 2023;44(11):13–20.

[12] Wang H, Zhou J, Tuo X, Wang M, Wang X, Feng L. Research and application of convolutional class Gaussian shaping filtering algorithm. Nucl Technol. 2024;47(1):91–98.

[13] Wang S, Xu J, Gao Y, et al. A geometric model of human factors parameters for active noise reduction path identification. J Acoust. 2024;49(2):226–37.

[14] Wang M, Wang Y, Chen E, Liu Y, Liu P. A high-speed train tread wear prediction model based on identity mapping multi-layer limit learning machine. J Mech. 2022;54(6):1720–31.

[15] Yang H, Li S, Shu J, Xu C, Ning Y, Ye J. Neural network based on array ultrasonic and feature fusion within reinforced concrete structure crack detection. Beijing, China: Journal of Building Structures; 2024. p. 1–12. doi: 10.14006/j.jzjgxb.2023.0729.

[16] Ma X. Artificial intelligence-driven education evaluation and scoring: Comparative exploration of machine learning algorithms. J Intell Syst. 2024;33(1):20230319.

[17] Al-zubidi AF, Farhan AK, Towfek SM. Predicting DoS and DDoS attacks in network security scenarios using a hybrid deep learning model. J Intell Syst. 2024;33(1):20230195.

[18] Liu W. Action recognition based on 3D bone key features, layered DNN, and immune optimization. Beijing, China: Donghua University. 2021. doi: 10.27012/d.cnki.gdhuu.2021.000156.

[19] Ren K, Zhuang F. A multidimensional scaling localization algorithm for correcting the shortest path distance matrix. J Sens Technol. 2016;29(1):129–35.

[20] Abdulhussain SH, Mahmmod BM, Baker T, Al-Jumeily D. Fast and accurate computation of high-order Tchebichef polynomials. Concurr Comput: Pract Exper. 2022;34(27):e7311.

[21] Abdulhussain SH, Mahmmod BM, Flusser J, AL-Utaibi KA, Sait SM. Fast overlapping block processing algorithm for feature extraction. Symmetry. 2022;14(4):715.

[22] Zheng Y, Chen Y, Bai W, Chen P. Fusion of event data and image frames for vehicle target detection. Calculation. 2024;44(3):931–7. http://kns.cnki.net/kcms/detail/51.1307.tp.20230810.1310.002.

[23] Yang X, Li Z, Zhang Z, Yu J, Chen J, Wang D. Based on fusion filter between the layers and social neural citation network of the recommendation algorithm. Shanghai, China: East China Institute of Computing Technology; 2024. p. 1–10. doi: 10.19678/j.issn.1000-3428.0068532

[24] Bi C. High dynamic dance motion recognition method based on 2D pose estimation. Xi'an Aviat Vocat Tech Coll. 2024;54(1):75–81. doi: 10.20051/j.issn.1003-4722.2024.01.011.

[25] Verma KK, Singh BM, Dixit A. A review of supervised and unsupervised machine learning techniques for suspicious behavior recognition in intelligent surveillance system. Int J Inf Technol. 2022;14(1):397–410.

[26] Xiao W, Liu H, Ma Z, Chen W. Attention-based deep neural network for driver behavior recognition. Future Gener Computer Syst. 2022;132:152–61.