Hindawi
Advances in Mathematical Physics
Volume 2021, Article ID 5183088, 12 pages
https://doi.org/10.1155/2021/5183088

Research Article

Hidden Markov Model-Based Video Recognition for Sports

Zhiyuan Wang,[1] Chongyuan Bi,[1] Songhui You,[1] and Junjie Yao[2]

[1]*International College of Football, Tongji University, Shanghai 200092, China*
[2]*School of Computer Science and Technology, East China Normal University, Shanghai 200062, China*

Correspondence should be addressed to Chongyuan Bi; 1932050@tongji.edu.cn and Songhui You; songhuiyou@tongji.edu.cn

Received 5 November 2021; Revised 25 November 2021; Accepted 27 November 2021; Published 20 December 2021

Academic Editor: Miaochao Chen

In this paper, we conduct an in-depth study and analysis of sports video recognition by improved hidden Markov model. The feature module is a complex gesture recognition module based on hidden Markov model gesture features, which applies the hidden Markov model features to gesture recognition and performs the recognition of complex gestures made by combining simple gestures based on simple gesture recognition. The combination of the two modules forms the overall technology of this paper, which can be applied to many scenarios, including some special scenarios with high-security levels that require real-time feedback and some public indoor scenarios, which can achieve different prevention and services for different age groups. With the increase of the depth of the feature extraction network, the experimental effect is enhanced; however, the two-dimensional convolutional neural network loses temporal information when extracting features, so the three-dimensional convolutional network is used in this paper to extract features from the video in time and space. Multiple binary classifications of the extracted features are performed to achieve the goal of multilabel classification. A multistream residual neural network is used to extract features from video data of three modalities, and the extracted feature vectors are fed into the attention mechanism network, then, the more critical information for video recognition is selected from a large amount of spatiotemporal information, further learning the temporal dependencies existing between consecutive video frames, and finally fusing the multistream network outputs to obtain the final prediction category. By training and optimizing the model in an end-to-end manner, recognition accuracies of 92.7% and 64.4% are achieved on the dataset, respectively.

1. Introduction

With the rapid development of computers, networks, and multimedia, and other related technologies, multimedia data has shown an exponential growth trend. A video is a common form of multimedia data, and it is one of the important components of multimedia data, which is closely related to our daily life. Video contains the richest data information, with a more complex structure and a large amount of data. Faced with such a huge video data, automatic video description can better manage and utilize these rich video resources, which can help users to improve the indexing speed as well as the search quality of online videos, so that they can play a greater role. For people with impaired vision, the automatic description of videos and combined with text-to-speech conversion technology converts the text within the computer into continuous natural language for communication [1]. It can help them to understand the content in the video better, thus, making life easier for the visually impaired. In the field of automatic video description research, automatic human action-based video analysis and understanding has gradually become a popular research problem in computer vision and pattern recognition in recent years [2]. Faced with such a huge amount of video data, automatic video description can better manage and utilize these rich video resources and can help users improve the indexing speed and search quality of online videos, so that they can play a greater role. It has a wide application prospect in the fields of intelligent life assistance, advanced human-computer interaction, and content-based video retrieval and is closely followed by researchers at home and abroad. In the face of the current problems in sports video analysis research, such as low-level video features cannot accurately reflect high-level human semantic concepts, high

time complexity and low recognition accuracy of action recognition algorithms in traditional RGB videos, and the use of single features cannot meet the massive growth of existing video data and its recognition of complex actions, the research on automatic description of competitive sports, with free gymnastics as a typical representative, has important heretical research significance and extensive practical application value [3]. In terms of theoretical research, automatic description of free gymnastics is a cross-cutting topic that integrates machine learning, pattern recognition, video analysis, computer vision, and cognitive science, which provides a good research object for these fields and can promote the development of related disciplines.

At the same time, information overload also leads to high-quality information drowning in a sea of information, for example, in a mailbox full of emails, when there is a large amount of historical information to be mined, it will cause difficulties to distinguish which emails are useful and which are useless. Information overload is also a waste of information resources. Although the storage of information has been converted from the paper era to the disk era, the duplication of information overload generates a large amount of worthless information, which still results in a large waste of resources [4]. For the problem of automatic description of free gymnastic videos, a difficult point in visual research, the study of automatic description based on free gymnastic videos has a wide range of application prospects and potential economic value in terms of practical applications. In addition to video retrieval and bringing convenience to visually impaired people as mentioned above, potential application areas include assisted training of sports, human-computer interaction, and program promotion. The movement of the human body in sports videos is very complex and skillful, and the analysis of sports videos is more challenging compared with daily sports [5]. The analysis of sports video can not only bring more viewing effect to sports games but also help coaches to analyze the games and assist athletes' training. Through the study of automatic understanding of free gymnastics, action data analysis is carried out while improving the accuracy of sports action recognition, so that the regular characteristics of the development of gymnastics technological innovation can be explored and the training function can be assisted. For example, with relevant competitors as the main research object, the gap between the difficulty, choreography, and quality of the set movements between the award-winning and ordinary competitors is analyzed, and the trend of the development and innovation of free gymnastics is studied to adjust the training countermeasures, to improve the athletes' skill level.

Further study of the specific correspondence between the change of the action and the fluctuation of the wireless signal can realize the behavior recognition. With the development of the human behavior recognition field, the research task has intensified from the initial restricted conditions where only simple single actions can be recognized to today's complex group behavior recognition in natural scenes, which poses a serious challenge both in terms of information acquisition equipment or algorithm capability. However, accurate recognition and analysis of human behavior in real scenes is still a challenging problem due to the interference of moving perspectives, lighting changes, cluttered scenes, etc. Therefore, how to effectively improve the accuracy of recognizing human behavior in videos is a hot research problem among scholars nowadays. As an important component of behavior recognition, the extraction result largely affects the accuracy and real-time of behavior recognition. Feature extraction is a classical problem in the field of computer vision and machine learning, unlike feature extraction in image space, the feature representation of human action in the video not only describes how a person looks in image space but also must extract the human appearance as well as posture changes, extending the feature extraction problem from two-dimensional space to three-dimensional space, which greatly increases the complexity of behavior mode expression and subsequent recognition tasks, while at the same time broadening the ideas for vision researchers in terms of solution ideas and techniques.

2. Current Status of Research

Motion and behavior analysis has a long history, and its research value is attractive to a variety of disciplines including psychology, biology, and computer science [6]. From an engineering perspective, the field of behavior recognition has expanded to a wide range of high-impact social applications, not only in the areas of intelligent video surveillance, video retrieval, and human-computer interaction as mentioned previously but also video-based behavior recognition has also contributed greatly to retail analysis, user interface design, robot learning, medical diagnosis, and improving the quality of life of the elderly, and thus a growing number of scholars are devoting themselves to the research in the field of behavior recognition. Nowadays, the mainstream research methods for behavior recognition are roughly divided into traditional machine learning methods and deep learning methods, and either of these two types of methods cannot be separated from the extraction of human behavior features from videos to characterize human behavior, so this section divides the features used for human behavior recognition into four categories, which are appearance and shape features, motion features, spatiotemporal features, and fusion features of multiple features, from the perspective of the kinds of features used by various algorithms to introduce the current research development in the field of human behavior recognition at home and abroad. Many convolutional neural network-based methods have been proposed for image recognition, which extends to behavior recognition in video [7]. A convolutional neural network with a deep structure is trained on a large dataset Sports-1 M, but its model is trained on a single image and cannot capture motion information between consecutive video frames [8]. Although any video frame can be represented by an image, a specific spatiotemporal feature representation remains crucial to merge motion patterns that cannot be captured by a model based on motion appearance alone. Thus, the key to this task is how to use deep neural networks for spatiotemporal feature extraction in a rational manner.

SHARMA used a deep learning-based keyframe detection method for sports videos to extract key poses of athletes' training through the analysis of weightlifting videos [9]. Xing et al. established a deep learning optimized ant colony algorithm model suitable for technical and tactical decision making in badminton to solve practical technical and tactical decision optimization problems [10]. Vijayaprabakaran proposed a video target tracking method for tennis based on an improved mean shift algorithm [11]. Zhang and Zhong proposed a DSP and FPGA-based embedded basketball sports video target tracking algorithm for the sports target tracking problem in basketball games [12]. Song et al. proposed a new method of using neural networks to predict the match-winner, and the proposed new prediction method was to be reliable through testing [13]. Kusakunniran proposed a new sports training sports video analysis system with very high accuracy and comprehensiveness of video and image information analysis and high accuracy of keyframe extraction as well as recall rate [14]. Sharif et al. extracted rich visual features of players in soccer game videos by building a full convolutional twin neural network and trained the network on many datasets containing similarity objects to improve the algorithm's ability to discriminate players of the same team [15]. Three variations in the role of various hand gestures on RSS were identified, and further integrated RSS magnitude variations were to accomplish the goal of detecting different hand gestures. The systems accomplished gesture recognition and did not wear any form of the sensing device. However, these systems generally have low recognition accuracy and are difficult to apply practically in reality. The reason for this is that the characteristics of RSS lead to a low accuracy rate. The attenuation caused by the multipath effect in indoor environments breaks down the decreasing nature of RSS over distance, thus, affecting the accuracy of the measurement. The multipath effect also causes fluctuations in RSS, which is another cause of measurement error. The recommended data sizes are 10, 20, 30, 40, and 50, respectively, and the similarity calculation methods adopt cosine similarity and cooccurrence similarity.

To address the effects of interference information such as scene switching and observation point changes in video images, as well as to better utilize the spatiotemporal saliency images generated in this paper, based on the spatiotemporal information and spatiotemporal saliency information present in video images, a multistream residual neural network is proposed based on the research of two-stream dual-stream convolutional neural network and is validated on the standard UCF101 dataset and HMDB51. The comparison is validated on the standard UCF101 dataset and HMDB51 dataset. In this paper, the proposed multistream residual neural network can effectively capture the spatiotemporal saliency information of foreground person targets in videos and achieve better recognition results on all the above datasets. The automatic description process of free-form gymnastics and the research principle is described, the process framework is specified, and then the current problem areas are presented. Then, the targeted solution is proposed, and the whole process of introducing the attention mechanism into the network structure is described. This is immediately followed by a description and analysis of the self-built dataset used for the experiments in this chapter. Finally, the experimental results of the improved algorithm with different network framework feature extraction methods are compared to verify the feasibility of the network improvement.

3. Hidden Markov Model for Sports Video Recognition Analysis

3.1. Improved Hidden Markov Models. The hidden Markov model can be applied based on two key settings, the first being the flush Markov assumption, which means that the Markov chain of hidden states is assumed to depend only on the state at whatever moment it is at the previous moment.

$$p(i_t|i_{t-1}, o_{t-1}, \cdots, i_t, o_t) = p\left(i_t^2|i_{t-1}^2\right). \tag{1}$$

It denotes the value of the state sequence at time t, and it denotes the value of the observation sequence at time t. This assumption says that the current moment state sequence depends only on the previous moment state sequence. The second is the observation independence assumption, which means that the observed state out at any moment is assumed to depend only on the state of the Markov chain at this moment and has no relation with other states. Each row represents the true attribution category of the data, and the total number represents the number of data instances of that category. Divided into horizontal and vertical optical flow diagrams, the optical flow feature is a set of dense optical flow, which is a set of displacement vector fields between adjacent frames, which can be used to extract motion information and play an important role in video recognition.

$$p(i_t|i_T, o_{T-1}, \cdots, i_t, o_t) = p\left(i_t^3|i_{t-1}^2\right). \tag{2}$$

Meanwhile, there are two models in the hidden Markov model, the generative model, and the discriminant model. The generative model is the model implemented by the generative method, which learns the joint distribution $P(X, Y)$ from the data and then derives the conditional probability distribution $P(X|Y)$ and treats it as a model for prediction, which can be expressed by the following equation.

$$P(X|Y) = \frac{P(X)}{P(X, Y)}. \tag{3}$$

Deterministic generative models have relatively stable transition patterns and can determine the next state with certainty based on the current state, such as a change in a traffic light [16]. Nondeterministic generative models have relatively diverse changes and cannot determine the next state with certainty, such as the change of weather. And discriminative models refer to the models obtained by discriminative methods. The three basic problems of hidden Markov models are the three problems that must be solved to put them into practical application. They are the

probability calculation problem, the learning problem, and the prediction problem. The diagram of its model module is shown in Figure 1.

The disadvantages of this method are poor accuracy and the requirement that the content in the set of items can be easily abstracted into meaningful features the process of abstracting items into features based on their content is usually complex and the process often requires knowledge of the domain associated with the items. The accuracy of the algorithm is heavily influenced by the features abstracted, for example, content such as music, pictures, and books, where it is difficult to extract their underlying features. However, the algorithm makes it difficult to mine the information about one's preferences latent in the users' historical choice data, since the users' historical data are independent of each other. In addition to the abovementioned video retrieval and convenience to the visually impaired, potential application areas also include sports assisted training, human-computer interaction, and project promotion. Finally, through testing on the KTH behavior data set, the confusion matrix and accuracy are used as performance indicators to verify the feasibility of the algorithm.

Based on data annotation, this chapter views the problem of automatic description of free-form gymnastics videos as the problem of automatic generation of video subtitles, which is a frontier problem of research in computer vision [17]. The task of automatic video caption generation is to automatically generate natural language descriptions of video content, and some approaches have been proposed to solve the problem of automatic video caption generation. For example, a video description model is implemented for video frame sequences input and text sequences output and can satisfy that the input video frames and the output text are of variable length. However, for the free gymnastics video, each decomposition action is mainly determined by the athlete's flipping direction, rotation degree, body posture, etc., and in a video sequence, these key actions only appear in some video frames, and the rest of the video frames are transition frames or contain some less important transition actions. In this section, the video frames containing the key movements of free gymnastics are defined as keyframes, and to improve the accuracy of the video description, an effective method is to extract the keyframes with high discriminative power in the video.

$$p(y_1, \cdots, y_m | x_1, \cdots, x_n) = \mathrm{Inf}\left\{ \prod_{t=1}^{m} p(y_t | h_{n+1}, y_n) \right\}, \tag{4}$$

$$\theta^* = \arg\min x_\theta \sum_{t=1}^{m} p(y_t | h_{n+1}, y_n). \tag{5}$$

A convolutional neural network is a multilayer feedforward neural network that can handle two-dimensional data very well. It is characterized by the fact that the neurons of adjacent layers of the network are connected, and the features of each layer are passed through a shared weight incentive of the convolutional kernel, emulating to some extent biological neural networks. Also, the learning capability of the model can be controlled by varying the depth and width of the network. Compared to other feedforward neural networks, convolutional neural networks have no connections made to neurons in the same layer, effectively reducing the learning complexity with much smaller number of parameters and easier to train. When we need to recognize high-dimensional images, fully connected layer concatenation is not applicable because it adds many parameters. Instead, local connectivity facilitates the learning of filters that capture only the important image features without having to learn and update the global weights. Training a CNN requires millions of parameters and must deal with a large amount of data, which takes up a lot of time. However, CNNs still learn much faster than ordinary feedforward neural networks. It also performs much better compared to standard feedforward neural networks with similar parameters. Combining these features concludes that the end-to-end hierarchical learning architecture of convolutional neural networks is well suited for feature learning and understanding images.

$$I(x, y) = I_x(x, y)^2 - I_y(x, y)^2, \tag{6}$$

$$G(x, y) = \sqrt{G_x(x, y)^2 - G_y(x, y)^2}, \tag{7}$$

$$\alpha(x, y) = \tan \frac{G_x(x, y)}{G_x(x, y) + G_y(x, y)}. \tag{8}$$

Similarity-based collaborative filtering algorithms can be divided into two categories: "user-based" and "item-based," where user-based recommendations are based on analyzing the similarity of user groups and recommending items liked by similar users. The most important element is the calculation of similarity. Most of the current methods are based on vectors, and the smaller the distance between vectors, the greater the similarity. In calculating the similarity, if the similarity between users is calculated, the users' preferences for all items are taken as vectors; if the similarity between items is calculated, the preferences of all users for an item are taken as vectors.

$$w_{i,j} = \frac{|N(i) \cup N(j)|}{|N(j)|}. \tag{9}$$

The main task of the similarity-based collaborative filtering recommendation algorithm is to solve the similarity, and after getting the similarity of users or items, the recommendation can be made based on the user association information or item association information, respectively. This type of algorithm first calculates the similarity between a user and other users or an item and other items, then finds K neighbors of the current user or item in order of similarity, then converts the similarity of neighbors into weights to predict the items, and finally outputs the recommendation results.

Therefore, it is not difficult to find that the item-based collaborative filtering recommendation algorithm is a more appropriate choice in this case, as it can make

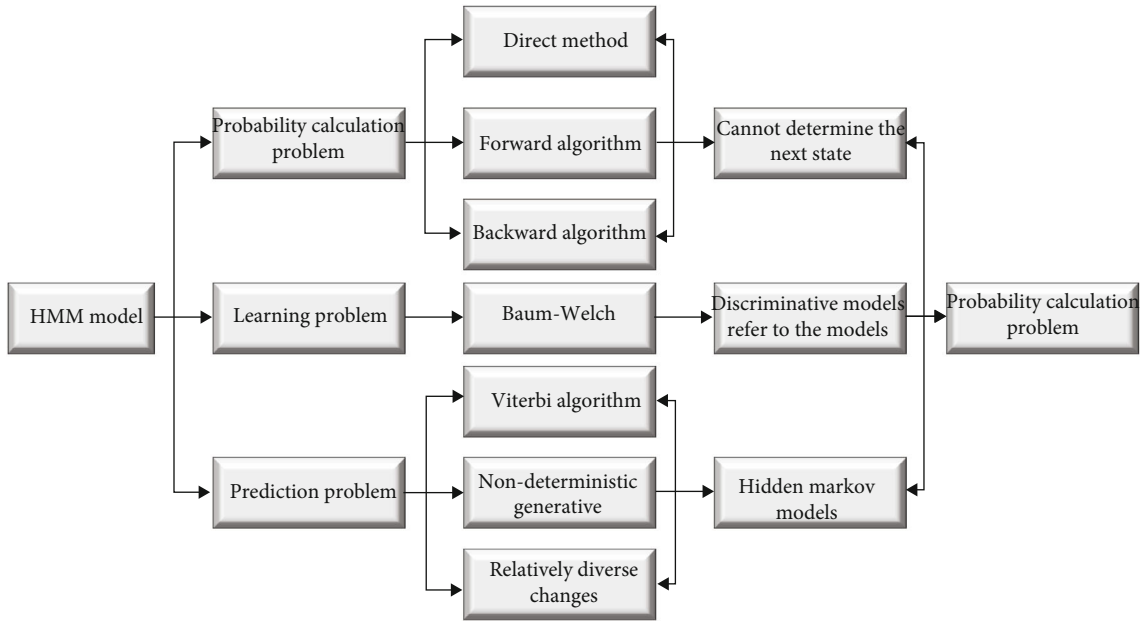

Figure 1: HMM model module diagram.

recommendations like what the user is interested in now and improve the user's stickiness [18]. This reason is difficult for the user to accept because the user is likely to not know the person in the algorithm and is not very interested in other people's interests; however, if it is an item-based recommendation algorithm, it can be explained that the item is like an item previously selected by the user so that the user is more likely to accept the result and accept the recommendation, as shown in Figure 2. To excavate the regular characteristics of the innovation and development of gymnastics technology and realize the auxiliary training function. For example, taking relevant players as the main research object, analyzing the gap between the difficulty, arrangement, and quality of the set of movements between the award-winning players and the ordinary players, studying the development and innovation trends of floor exercise, adjusting training strategies, and improving the skills of the athletes.

In some time-related problems, it encountered that the events occurring at one moment are directly influenced by the events occurring at the previous moment, and the hidden Markov model reaps the best application when dealing with these problems. There is a predefined set of parameters in the HMM model which can give the best explanation for a particular sample. In practical applications, the class to which the sample to be tested is assigned can give the best explanation for this test sample. Hand gesture recognition is a typical problem of this type and can therefore be well utilized. However, the model was less popular and less applied before and has only been widely utilized in recent years. The complex gesture recognition technology in this paper is implemented precisely based on the Fresnel zone model and then using the hidden Markov model.

$$Q(\lambda, \bar{\lambda}) \sum_{I} \ln P(O, I|\lambda) P\left(O, I|\bar{\lambda}\right), \tag{10}$$

$$A = P(i_t). \tag{11}$$

The construction of the free gymnastics decomposition movement dataset is the basic work for the automatic description of free gymnastics. For the construction of this dataset, many high-profile events of professional athletes collected in this paper, including multiple heavyweight events of people such as the Olympic Games, World Championships, and National Games. First, these event videos are preprocessed, and a complete video of a free gymnastics event is completed by the joint participation of several athletes, during which the replay of exciting moments, slow-motion commentary, and judges' scoring and ranking is interspersed. The massive amount of video is cropped on an athlete-by-athlete basis, and only the athlete's free gymnastics sets are retained. The final construction contains 298 videos in the training data and 45 videos in the test data, and all the free gymnastics decomposition moves in the test set are present in the training set. Since the commentaries of sports events are not subtitled, the narrator's voice commentary on the names of the decomposition moves is accompanied by some commentary, highlight replay, etc., and there are too many distracting factors to be implemented by techniques such as speech recognition. The free gymnastics decomposition action description dataset used in this paper can only be manually annotated based on real-time sports narration words. The 298 descriptions corresponding to the 298 videos in the training set are subworded, and word frequency statistics are performed.

3.2. Experimental Design of Sports Video Recognition. When the user walks or does an action in the Fresnel zone, each time the boundary of the Fresnel zone is crossed, the digital waveform received by the receiving device presents the highest or lowest point. And if the user keeps walking along the boundary of a certain Fresnel zone, the digital waveform

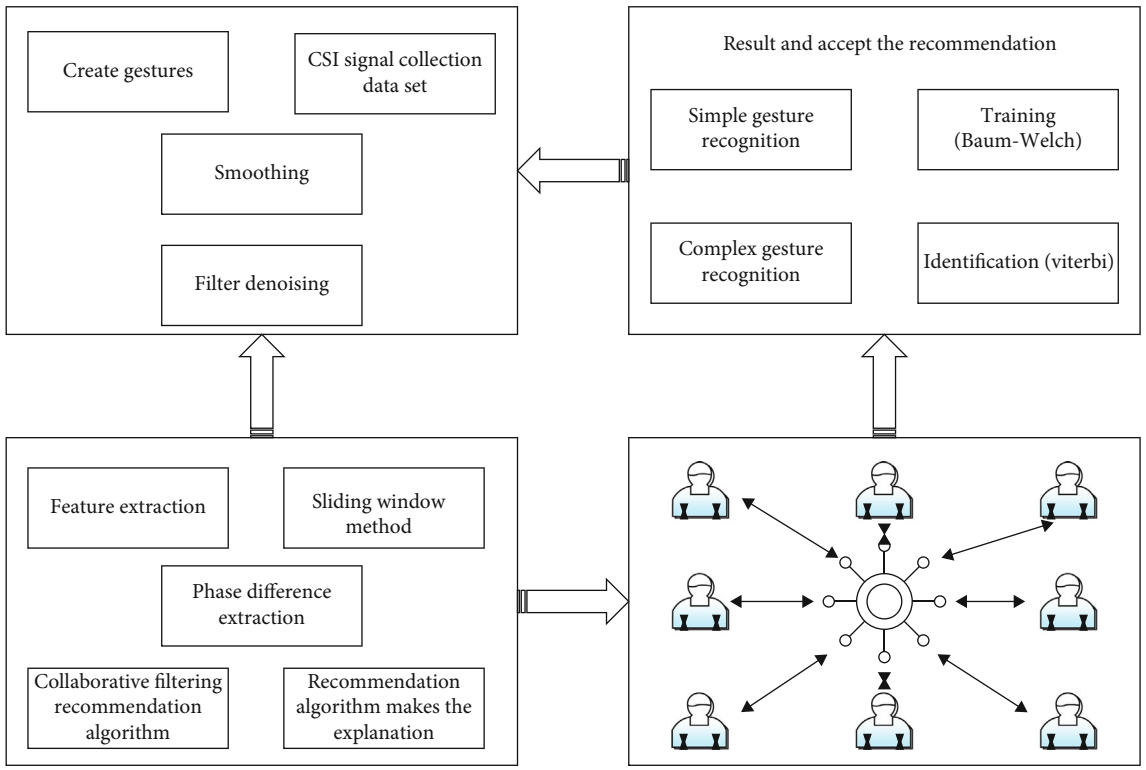

Figure 2: Framework of identification technology.

presented by the signal receiving device does not change. When the user walks continuously through the boundaries of multiple Fresnel zones, the received signal looks like a sinusoidal waveform, and the peaks of this sinusoidal waveform each correspond to the boundaries of odd/even Fresnel zones [19]. The body cuts vertically through the Fresnel zone boundaries and gradually moves away from the center of the Fresnel zone, the distance between two Fresnel zone boundaries decreases, and there is a tendency for the absolute value of the peaks and troughs of the phase difference to decrease within the Fresnel zone. When walking in the opposite direction, the absolute values of the peaks and troughs with phase differences within the Fresnel zone may increase.

The technology consists of two parts, hardware and software, the left frame is the hardware part, and the right frame is the software part. In the hardware module, the sender is a daily used laptop and the receiver is a desktop computer with a wireless card that is in a wireless network environment with the laptop. The laptop is connected to the desktop computer with the help of a wireless card, and the router is treated as a wireless access point, which is used to forward wireless signals. When experimenting, the transmitting device keeps sending out wireless signals, and the experimental volunteers do the designed actions between the wireless access point and the receiving device leading to constant fluctuations in the wireless signals, and further study the specific correspondence between the changes in the actions and the fluctuations in the wireless signals will enable the behavior recognition [20]. Regardless of the information acquisition equipment or algorithm capabilities, severe challenges are presented. However, human behavior in real scenes is disturbed by factors such as moving perspectives, changes in lighting, and cluttered scenes. Accurate identification and analysis of human behavior is still a challenging issue. The software module consists of four parts: establishing action data set, data preprocessing, feature extraction, and age group recognition. After the receiver device acquires the gait data, the next step is to collect the CSI information from the behavior data and build the action dataset. Then, the preprocessing work completed, including filtering denoising and smoothing. Then, the phase difference features are extracted using the sliding window method. After that, the dataset is divided into two parts using the leave-out method as training set and test set and fed into the SVM classification algorithm for training. After testing the user's gait for recognition and then identifying the age group of the user as shown in Figure 3.

After the CSI is selected as a suitable base signal and an appropriately sized action data set is built, it needs to be preprocessed to improve the data reliability. The frequencies of human behavioral activity signals are concentrated between 0-5 Hz and are easily disturbed by other frequency information. During the experiment, the obtained data will have irrelevant values, missing values, duplicate values, outliers, etc. The purpose of preprocessing is to denoise and remove the outliers. This paper uses a Butterworth filter for filtering and denoising, which requires only two parameters to characterize, the order of the filter and the cutoff frequency.

$$H_{\alpha}(j_{\Omega})^{2} = \frac{\Omega_{c} - \Omega}{\Omega_{c}\Omega}. \tag{12}$$

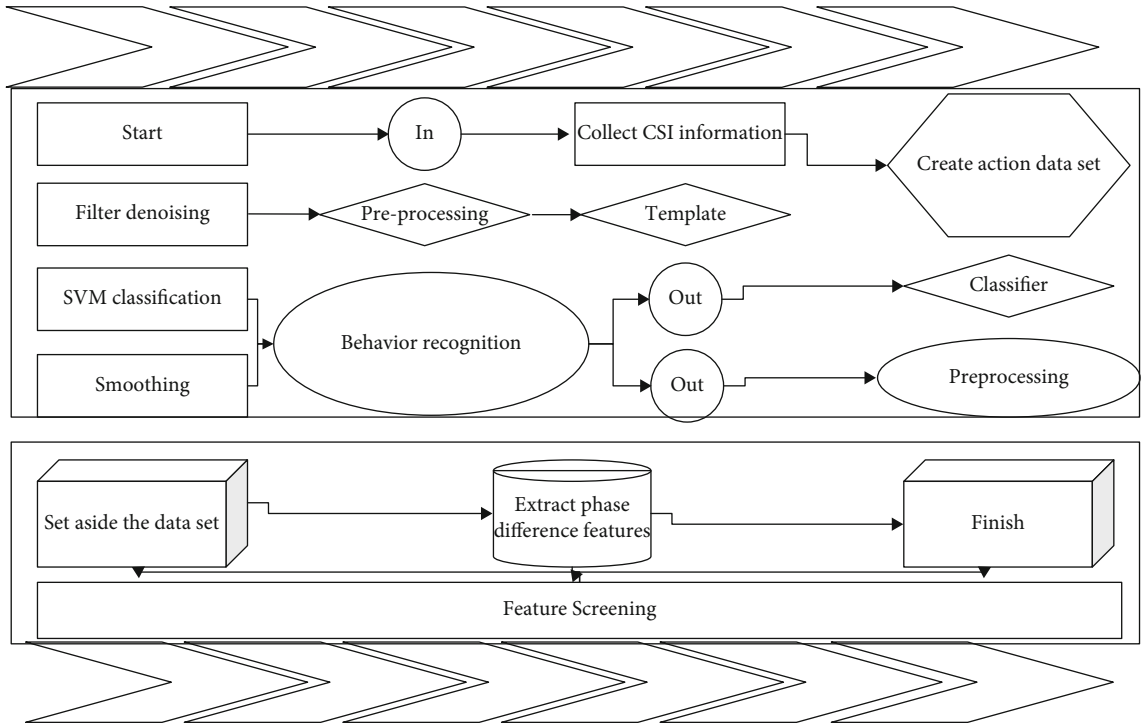

Figure 3: Behavior recognition flow chart.

There are still some sharp burrs on the behavior curve that affect the smoothness, so smoothing is performed. The smoothed behavioral signal will get more realistic features with less possibility of errors during feature extraction. This thesis uses the median filtering method to accomplish the smoothing process. The median filtering method can effectively filter out the points near the middle to achieve the purpose of denoising, and at the same time, it has an excellent filtering effect on the impulse interference, and the filtering can also achieve the protection of the nonnoise signal. In addition, the median filtering method is relatively simple in operation, and the equipment may encounter fewer problems when implementing it. Therefore, this method has been widely utilized in signal processing and other related fields since its emergence.

$$\Delta f = \frac{x}{4(d_n + d_0)}, \tag{13}$$

$$K(x_i, x_j) = x_i^T x_j^T. \tag{14}$$

In this paper, we apply time-domain analysis to extract features, which is essentially a way of analyzing control systems. It means analyzing the system performance according to an expression under the condition that the input is determined. Time-domain features are the characteristics of the signal in the time domain. This method of feature extraction is easy to implement and is not difficult to compute numerically and can be applied directly to the data if the data has already been preprocessed. The most seen time-domain features have a clear physical meaning and can be obtained directly by calculating the signal amplitude over some time according to the formula accordingly, as shown in Figure 4.

Once two subcarriers are selected in the phase difference extraction phase, the longer reflection path will result in a larger phase difference. When the user's position is fixed, the larger the frequency difference between the two subcarriers, the larger the phase difference. If the difference between the two subcarriers is too small, the two waveforms are too close to each other to be differenced. But a large difference between the two subcarriers will lead to phase blurring. Therefore, the choice of subcarriers is particularly important. The changes of the nondeterministic generative model are relatively diverse, and the next state, such as the change of weather, cannot be accurately determined. The discriminant model refers to the model obtained by the discriminant method. The three basic problems of the hidden Markov model are the three problems that must be solved for its practical application.

$$K(x_i, x_j) = e^{-\left(\gamma \left\| x_i - x_j \right\|\right)}, \tag{15}$$

$$W_{mm}^k = \nabla \left\| E^k\left(y_m^k\right) + E^k\left(y_n^k\right) \right\|. \tag{16}$$

The parameters are modified to minimize the generalization error. In this paper, the sample data volume is medium, and the combination of the grid search method and cross-validation method is chosen to complete the determination of parameters. During the experimental process, the parameters were continuously modified by applying the control variables method to find the set of parameters with the highest training and prediction accuracy for LIBSVM [21]. The

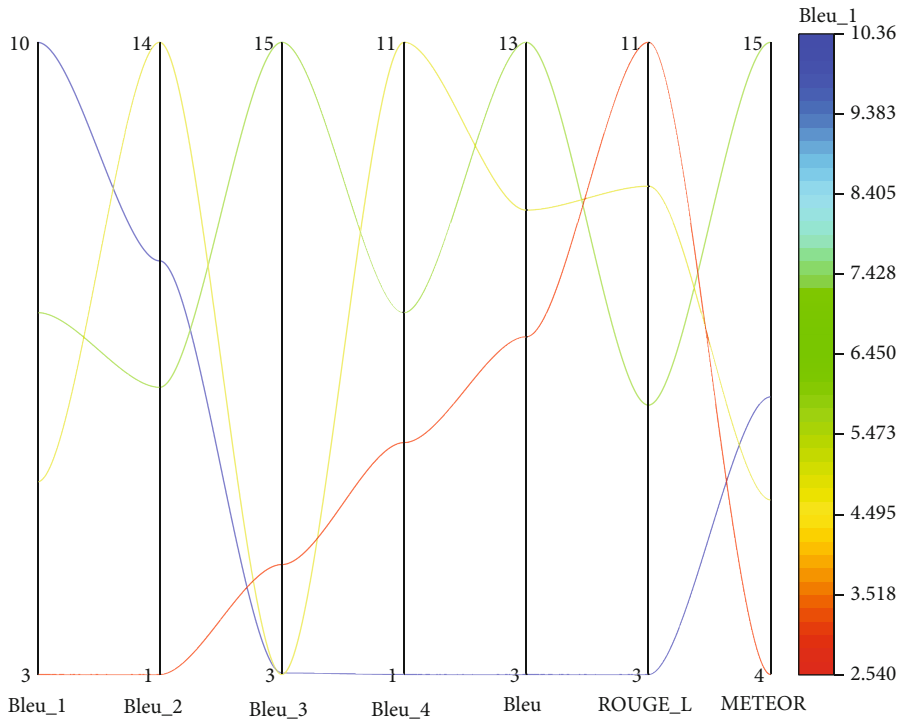

Figure 4: Experimental comparison of feature extraction networks.

SVM model was trained using the dataset prepared for the training process, and the corresponding classifier was also obtained to automatically display the recognition accuracy. The final recognition accuracies for the three different gait patterns were obtained after several tests. Also, the recognition time at each recognition is recorded and used for the final comparison of the complexity of the algorithm.

4. Analysis of Results

4.1. Hidden Markov Model Performance Results. To verify the performance of the recommendation algorithm proposed in this paper, it is necessary to compare the algorithm with other classical personalized recommendation algorithms, including item-based, user-based collaborative filtering recommendation algorithm, and content-based recommendation algorithm, in cross-sectional experiments. For the item-based collaborative filtering recommendation algorithm, recommendations with 10, 20, 30, 40, and 50 recommendation data are performed, where the similarity calculation methods are cosine similarity and same-present similarity, and the experimental results are shown in Figure 5.

As observed in Figure 5, the improved artificial bee colony algorithm can converge to the optimal solution faster and obtain better fitness values compared to the traditional bee colony in terms of optimizing the support vector machine parameters. The improved algorithm converges to the optimal value faster after 30 iterations because the improved swarm search strategy replaces the location of the already searched optimal solution with the location of the randomly generated solution when the bees search for new values later in the algorithm, enhancing the speed of convergence and mining capability of the algorithm for finding the optimal solution. To validate the performance of the optimized classifier, the confusion matrix is used as a performance evaluation metric to validate the algorithm. The confusion matrix also called the error matrix, and it is also the most basic and intuitive calculation in measuring the accuracy of the classification model. Each column of the confusion matrix represents the prediction type, and its total represents the number of data predicted to be in that category; each row represents the true attribution category of the data, and its total represents the number of data instances in that category. The optical flow maps are divided into horizontal and vertical directions. The optical flow feature is a set of dense optical flow, a set of displacement vector fields between adjacent frames, which can be applied to extract motion information and play an important role in video recognition. The network framework designed in this paper forms optical flow images imported in the input optical flow information encompasses the motion information of each static video frame image, which improves the correlation of spatiotemporal features on pixel points and the robustness of processing video frame sampling.

For the training of the spatial streaming neural network, RGB images are used as the input to the spatial stream, and the original video frames provide the basic appearance features of the video. The input to the spatial stream consists of multiple RGB images, and the above RGB images are obtained from the extracted video frames, randomly sampled according to the same time interval. Like the temporal segmentation network training strategy, ten video images are randomly selected from the videos to represent the

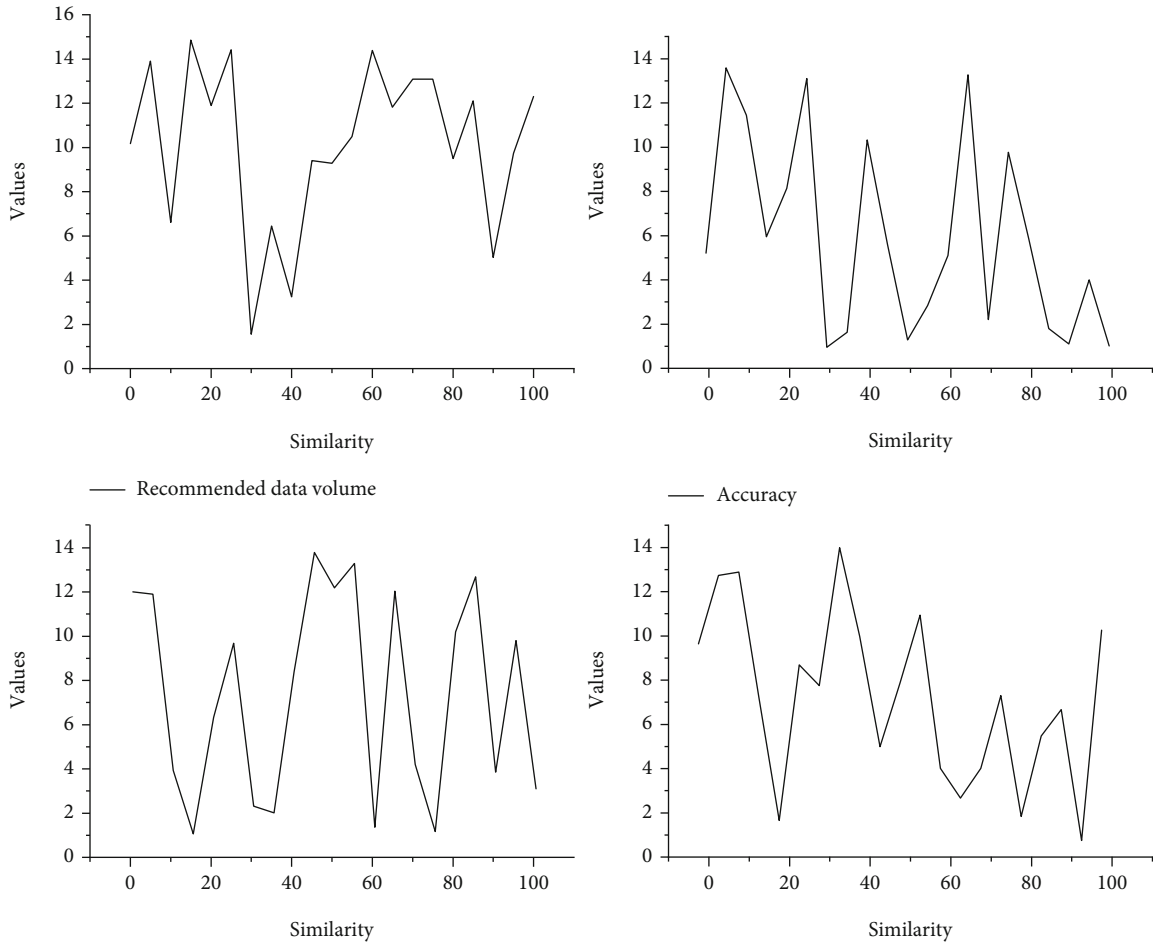

Figure 5: Results of hidden Markov model algorithm.

corresponding videos. Each of the ten video images is input to each CNN, and the loss values are calculated, which are then used as the final loss for backpropagation. The output vector after the feature extractor is fed into the model of the attention mechanism network, and finally, the hidden state vector of the attention mechanism is subsequently predicted by a softmax classifier for classification as shown in Figure 6. To a certain extent, it imitates the biological neural network. At the same time, the learning ability of the model can also be controlled by changing the depth and width of the network.

By analyzing the experimental results in the above figure, it can be found that on the UCF101 dataset, the recognition accuracy of the model is significantly higher when the STS-ALSTM network model takes two modal video images as input data than the input of a single modal video image. For example, with RGB images and optical flow images as input, the recognition accuracy is 7.0% and 5.8% higher than that of a single RGB image and a single optical flow image, respectively; with RGB images and spatiotemporal saliency images as input, the recognition accuracy is 2.2% higher than that of a single RGB image; and with optical flow images and spatiotemporal saliency images as input, the recognition accuracy is 2.6% higher than that of a single optical flow image. On the HMDB51 dataset, the recognition accuracy of the network model with two modal data inputs also outperformed that with single modal data input, further confirming that it can provide the network model with more

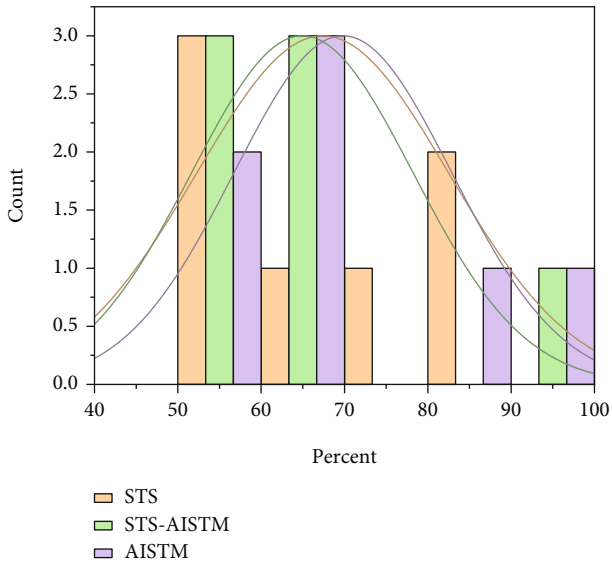

Figure 6: Behavior recognition comparison results.

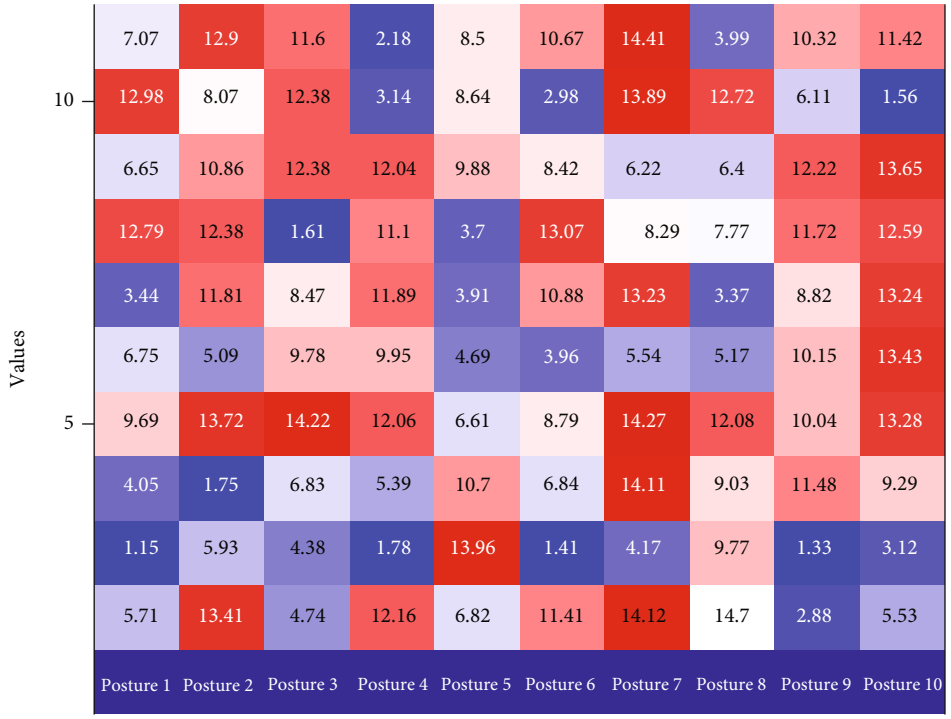

Figure 7: Accuracy of complex gesture recognition.

adequate spatial and temporal information when fed with multiple video image data inputs.

In addition, by using a network model combining a spatiotemporal saliency multistream network with an attention mechanism, it can be found that using optical stream images and spatiotemporal saliency images as input improve the recognition accuracy by 2.6% and 0.8% on the UCF-101 and HMDB-51 datasets, respectively, compared to a single optical stream image input. The addition of temporal saliency streams provides foreground information about the target object in the video image and reduces background interference, improving the video recognition accuracy. The same applies to the network model with RGB images and spatiotemporal saliency images as input, which improves the accuracy by 2.2% and 4.1% over the RGB image input, on the UCF-101 and HMDB-51 datasets. When fusing all streams, the highest recognition accuracies that can be obtained on the UCF-101 and HMDB-51 datasets are 92.7% and 64.4%, respectively.

4.2. Experimental Results of Sports Video Recognition. The confusion matrix for the four basic gesture recognition performance is shown in Figure 7. Each row represents the actual gestures performed by the user, and each column represents the gestures classified for recognition. Each element of the matrix corresponds to the score of the gesture in the row that is classified as a gesture in the column. The average accuracy for classifying the four basic gestures is 92.7%. Comparing the recognition accuracy of the four gestures, it is found that the recognition accuracy is similar. Among them, the recognition accuracy was higher for the front finger and back pull and lower for the left waving and right waving. The results show that gestures that cut the boundary of the Fresnel region vertically are more likely to be recognized. The good or bad effect of simple gesture recognition determines the effect of complex gesture recognition, and this paper tested three complex gestures in different experimental environments when the volunteers stood in different positions under different scenarios. The average accuracy of complex gesture recognition is above 86%, which verifies the robustness of the complex gesture recognition technique proposed in this paper.

The model has two tributary networks, where the spatial flow network extracts feature about spatial information from the input RGB map and the temporal flow network extracts features about motion information from the optical flow map, the final recognition results are obtained by fusing the outputs of the two networks, spatial flow, and temporal flow, and the initial dual-flow network structure is replaced by Reset-50 to improve the feature representation capability of the model. To improve the feature representation capability, the initial two-stream network structure was replaced by ResNet-50, and the effects of pretraining initialization, segment sampling strategy, data enhancement, and dropout on the performance of the two-stream residual network model were investigated, respectively. The average recognition accuracy of the model is experimentally compared on both UCF101 and HMDB51. The experimental results demonstrate that the performance of the dual-stream residual convolutional neural network-based behavior recognition model is significantly improved compared to the initial dual-stream network model.

Figure 8 shows the two-stream fusion confusion matrix of the two-stream residual network model based on the

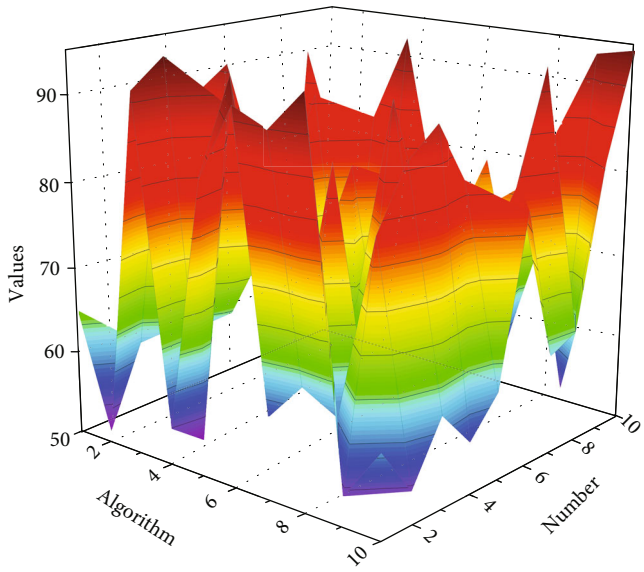

Figure 8: Comparison of the recognition accuracy of some behavioral categories in the confusion matrix.

Res2Net module on the video dataset UCF101 (split1). Figure 8 presents the recognition accuracies of some of the behavior categories in the confusion matrix and compares them with the results of the initial two-stream residual network model. From the results of the graphs, the recognition accuracies for the four behaviors crickets hot, hammering, brushing teeth, and cricket bowing are between 60% and 70%, respectively, which is a significant improvement over the results of the initial two-stream residual network model (50% to 60%).

Considering that a support vector machine classifier requires several experiments to determine the appropriate performance parameters when applied to different datasets, an improved artificial bee colony algorithm is used to optimize the parameters of the support vector machine to achieve the appropriate optimal performance parameters for each specific dataset. First, the artificial swarm algorithm is improved by introducing a segmented logistic chaos mapping to generate the initial population and proposing a new search strategy to increase the diversity of the population samples and the algorithm's ability to be mined at a later stage; then, the improved artificial swarm algorithm is optimized for the support vector machine parameters; finally, the algorithm is tested on the KTH behavioral dataset with the confusion matrix and accuracy as performance metrics to verify the feasibility of the algorithm. The current research on human behavior recognition of video has been a challenging problem in the field of computer vision, and how to apply it to real-life is the top priority. The improved algorithm studied in this paper is based on video image human behavior recognition, how to extract efficient and powerful features and high-performance classifier is the core of this paper, through experimental verification the proposed improved algorithm has achieved a good improvement in recognition effect.

5. Conclusion

This paper analyzes and summarizes the research in the field of behavior recognition, recognizing the huge potential for application and development in this field. After fully understanding the implementation process of behavior recognition technology in a wireless environment, the bottleneck and development direction of the current stage of behavior recognition technology are carefully analyzed. And after comparing the two signals, RSS and CSI, CSI is chosen as the base signal, and the method of this paper is proposed. Whether it is a scenario requiring a high-security factor with real-time feedback or a more complex public often with a large flow of people, the technology proposed in this paper can provide the required service and prevention. Complex gesture recognition based on hidden Markov model gesture features, which applies Hidden Markov Model features to gesture recognition, can recognize complex gestures that are a combination of simple gestures. The combination of the two modules forms the overall technique of this paper, which has practical applications in many indoor scenarios. To test the robustness and usability of the algorithm, three contrasting experimental scenarios, classroom, conference room, and laboratory, which have different disturbances and help to verify the effectiveness and usability under different levels of disturbances. The behavior recognition is completed in the three environments, respectively, and the recognition accuracy is compared. The experimental conclusions can show that the multifeatured behavior recognition technology proposed in this paper can effectively recognize the age group of people and complex gestures in indoor scenes, and the recognition accuracy is 91.2% and 86%, respectively, with strong robustness.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the Fundamental Research Funds for the Central Universities (22120200376).

References

[1] Y. Lu and S. An, "Research on sports video detection technology motion 3D reconstruction based on hidden Markov model," *Cluster Computing*, vol. 23, no. 3, pp. 1899–1909, 2020.

[2] A. Nadeem, A. Jalal, and K. Kim, "Automatic human posture estimation for sport activity recognition with robust body parts detection and entropy Markov model," *Multimedia Tools and Applications*, vol. 80, no. 14, pp. 21465–21498, 2021.

[3] M. Uto, Y. Miyazawa, Y. Kato, K. Nakajima, and H. Kuwata, "Time- and learner-dependent hidden Markov model for writing process analysis using keystroke log data," *International Journal of Artificial Intelligence in Education*, vol. 30, no. 2, pp. 271–298, 2020.

[4] O. AlShorman, B. Alshorman, and M. S. Masadeh, "A review of physical human activity recognition chain using sensors," *Indonesian Journal of Electrical Engineering and Informatics (IJEEI)*, vol. 8, no. 3, pp. 560–573, 2020.

[5] A. Mottaghi, M. Soryani, and H. Seifi, "Action recognition in freestyle wrestling using silhouette-skeleton features," *Engineering Science and Technology, an International Journal*, vol. 23, no. 4, pp. 921–930, 2020.

[6] M. Tadayon and G. J. Pottie, "Predicting student performance in an educational game using a hidden Markov model," *IEEE Transactions on Education*, vol. 63, no. 4, pp. 299–304, 2020.

[7] T. Huynh-The, C. H. Hua, N. A. Tu, and D. S. Kim, "Physical activity recognition with statistical-deep fusion model using multiple sensory data for smart health," *IEEE Internet of Things Journal*, vol. 8, no. 3, pp. 1533–1543, 2021.

[8] M. F. Ukrit and P. Nithyakani, "The systematic review on gait analysis: trends and developments," *European Journal of Molecular & Clinical Medicine*, vol. 7, no. 6, pp. 1636–1654, 2020.

[9] A. Sharma and N. Varshney, "Identification and detection of abnormal human activities using deep learning techniques," *European Journal of Molecular & Clinical Medicine*, vol. 7, no. 4, pp. 408–417, 2020.

[10] Y. Xing, B. Tian, C. Lv, and D. Cao, "A two-stage learning framework for driver lane change intention inference," *IFAC-Papers OnLine*, vol. 53, no. 5, pp. 638–643, 2020.

[11] K. Vijayaprabakaran, K. Sathiyamurthy, and M. Ponniamma, "Video-based human activity recognition for elderly using convolutional neural network," *International Journal of Security and Privacy in Pervasive Computing*, vol. 12, no. 1, pp. 36–48, 2020.

[12] G. Zhang and L. Zhong, "Research on volleyball action standardization based on 3D dynamic model," *Alexandria Engineering Journal*, vol. 60, no. 4, pp. 4131–4138, 2021.

[13] Z. Song, S. Ali, and N. Bouguila, "Bayesian inference for infinite asymmetric Gaussian mixture with feature selection," *Soft Computing*, vol. 25, no. 8, pp. 6043–6053, 2021.

[14] W. Kusakunniran, "Review of gait recognition approaches and their challenges on view changes," *IET Biometrics*, vol. 9, no. 6, pp. 238–250, 2020.

[15] M. Sharif, M. A. Khan, F. Zahid, J. H. Shah, and T. Akram, "Human action recognition: a framework of statistical weighted segmentation and rank correlation-based selection," *Pattern Analysis and Applications*, vol. 23, no. 1, pp. 281–294, 2020.

[16] A. Taheri, A. Meghdari, and M. H. Mahoor, "A close look at the imitation performance of children with autism and typically developing children using a robotic system," *International Journal of Social Robotics*, vol. 13, no. 5, pp. 1125–1147, 2021.

[17] M. Altuve, P. Lizarazo, and J. Villamizar, "Human activity recognition using improved complete ensemble EMD with adaptive noise and long short-term memory neural networks," *Biocybernetics and Biomedical Engineering*, vol. 40, no. 3, pp. 901–909, 2020.

[18] Y. Ünlü and Z. Taş, "A bibliography experiment on research within the scope of industry 4.0 application areas in sports," *Journal of Human Sciences*, vol. 17, no. 4, pp. 1149–1176, 2020.

[19] D. R. Beddiar, B. Nini, M. Sabokrou, and A. Hadid, "Vision-based human activity recognition: a survey," *Multimedia Tools and Applications*, vol. 79, no. 41-42, pp. 30509–30555, 2020.

[20] C. Nalmpantis and D. Vrakas, "Machine learning approaches for non-intrusive load monitoring: from qualitative to quantitative comparation," *Artificial Intelligence Review*, vol. 52, no. 1, pp. 217–243, 2019.

[21] R. V. Pawar, R. M. Jalnekar, and J. S. Chitode, "Review of various stages in speaker recognition system, performance measures and recognition toolkits," *Analog Integrated Circuits and Signal Processing*, vol. 94, no. 2, pp. 247–257, 2018.