Hindawi
Mathematical Problems in Engineering
Volume 2021, Article ID 7963616, 8 pages
https://doi.org/10.1155/2021/7963616

*Research Article*

# Human Motion Recognition Based on Multimodal Characteristics of Learning Quality in Football Scene

**Yuzhou Gao and Guoquan Ma**

*Department of Physical Education, Lanzhou University of Technology, Lanzhou 730050, Gansu, China*

Correspondence should be addressed to Guoquan Ma; mgquan@lut.edu.cn

Received 18 June 2021; Revised 1 August 2021; Accepted 11 August 2021; Published 31 August 2021

Academic Editor: Sang-Bing Tsai



The task of human motion recognition based on video is widely concerned, and its research results have been widely used in intelligent human-computer interaction, virtual reality, intelligent monitoring, security, multimedia content analysis, etc. The purpose of this study is to explore the human action recognition in the football scene combined with learning quality related multimodal features. The method used in this study is to select BN-Inception as the underlying feature extraction network and use uncontrolled environment and real world to capture datasets UCFl01 and HMDB51, and pretraining is carried out on the ImageNet dataset. The spatial depth convolution network takes image frame as input, and the temporal depth convolution network takes stacked optical flow as input to carry out human action multimodal identification. In the results of multimodal feature fusion, the accuracy of UCFl01 dataset is generally high, all of which are over 80%, and the highest is 95.2%, while the accuracy of HMDB51 dataset is about 70%, and the lowest is only 56.3%. It can be concluded that the method of this study has higher accuracy and better effect in multimodal feature acquisition, and the accuracy of single-mode feature recognition is significantly lower than that of multimodal feature recognition. It provides an effective method for the multimodal feature of human motion recognition in the scene of football or sports.

## 1. Introduction

In addition to the detection, recognition, and tracking of moving objects, action analysis and understanding are also included in the action recognition of people in the football scene. This action analysis realizes the interaction of a person and another and that of an object and a person in the football scene. Football video is not only a simple popular entertainment information but also an important media and method for sports professionals and football fans to analyze and understand the game. The application of deep learning in the field of human motion cognition is a hot topic in the academic field.

The main purpose of action recognition is to analyze people's behavior types in video, including computer vision, image processing, pattern recognition, feature engineering, and other fields of knowledge and technology. In the process of recognition, we should not only recognize and track the actions from the image sequence but also analyze the state and trend of the movement so as to infer the specific action categories of people. Compared with image, video has more information, and is actively exploring the method of applying deep learning to video image.

In human action recognition, Liu et al. thought that human behavior recognition is an active research field in the field of computer vision and machine learning. They proposed a large number of algorithms, most of which are designed for the subsets of four learning problems. In these problems, the comparison between algorithms may be further limited by the variance within the dataset, experimental configuration, and other factors. As far as they know, there was no dataset that allows parallel analysis of four learning problems. Their research introduces a new multimode, multiview, interactive dataset to evaluate human behavior recognition methods in four scenarios. The dataset consists of 1760 action samples from 22 action categories,

including 9-person interaction and 13-person interaction, and the accuracy of this method is not high [1]. Sivarathinabala et al. believed that multimodal biometrics improve security by protecting the system from spoofing attacks. The system uses face and gait biometrics for authentication and recognition. These videos are taken from two surveillance cameras, which are located in the frontoparallel and frontonormal views, respectively, as the input of the system. The gait system uses video from the frontoparallel view and uses the modeless method to extract the temporal and spatial motion summary of gait cycle. They compared their gait characteristics by calculating the Euclidean distance between them. The face system uses the video from the frontonormal view and uses the appearance based method to extract features from the user's face. They compared the facial features by calculating the chi-squared dissimilarity between them. A threshold is kept and compared with the score to verify this person. The minimum distance classifier recognizes others by fusing multimodal features; this method is not effective in feature recognition [2]. Sun et al. described their work in the field of static and continuous facial expression recognition. They evaluated the recognition effect of gray depth feature and color depth feature and discussed the fusion of multimodal texture features. For continuous facial expression recognition, they designed two spatiotemporal dense scale invariant feature transform features and combined them with multimodal features to recognize expressions from image sequences. For static facial expression recognition based on video frames, they extracted dense sift and some deep convolution neural network features, including their CNN structure. Aiming at the two kinds of features on the datasets of static wild expression and dynamic wild expression and proposing a fusion network, which fused all the features extracted in the decision-making layer, the failure rate of this method is high [3].

This study first introduces the characteristics of football scene and the analysis of learning quality and, at the same time, summarizes the method and classification of human recognition in detail. In this study, convolution feature extraction, convolution gating recursion unit, and spatial attention module are mainly used in the human feature recognition algorithm. In the experimental part of this study, the key datasets and some parameter settings are listed. Combined with the results, multimodal feature fusion analysis, multimodal fusion identification technology trend analysis, and single-mode and multimodal performance comparison analysis are carried out. It can be concluded that the method of this study has higher accuracy and better effect in multimodal feature acquisition.

## 2. Multimodal Feature Learning Quality and Human Motion Recognition

*2.1. Characteristics of Football Scene.* The football match usually is 90 minutes long. In the high-level football match, the athlete runs a distance of 9000–13000 meters, runs a short distance of more than 2500 meters, and completes hundreds of technical actions, with more than 180 beats/minute in 32 minutes, more than 300 liters of oxygen, and 1500–2000 kcal of calories. The competition is very difficult. There are 11 athletes in each competition group and 8000 people in the venue. The multisquare meter modern football game is a battle around the rights of time and the universe. Good fighters in the field are often fleeting. All players must adjust their technical actions and offensive or defensive strategies according to the changing situation on the field. It is a representative open sport in football match; however, there are also closed scenes such as free throws in football, which are the coexistence of open scenes and closed scenes [4, 5].

In a word, the venue of the football match is very large, and there are many participants as well as players' consumption. The open and closed sports scenes coexist. Due to the coexistence of these factors, the field of football matches changes rapidly, and the sports situation is complex and diverse. The player's tactical choice cannot be carried out without obtaining the information related to the running method, the position of the ball, the technical characteristics, and the physical condition of the partner and the opponent [6].

*2.2. "Microprocess" Analysis of Learning Quality in Football Scene.* According to the analysis of football characteristics, when students deal with specific football scenes, they will experience the following processes:

(1) Get scene information: my position, the people around me (teammates, mudguards), the position of the ball, the people who control the ball, the referee in the game, the audience, the venue environment, and other objective information [7]

(2) Analyze scene information: quickly summarize, integrate, and analyze all objective factors

(3) Make a decision: according to the analysis results, make decisions that bring benefits to the team

(4) Effective action: according to their own decisions, quickly perform the corresponding actions [8]

(5) Effect evaluation and feedback: please evaluate your actions and use your brain to respond quickly

In football, the processing time is very short. If the processing time of each "microprocess" is long, the performance level of football will be improved. Moreover, such a "microprocess" is constantly circulating between the tourmaline and flint before the end of the whole activity. It can be seen from this that if students cannot take their own ideas as the leading consideration, it is hard to say that they cannot make mistakes in that field or solve the problems at that time [9, 10].

*2.3. Action Identification and Classification.* According to different cognitive concepts, action recognition technology can be divided into the previous action recognition technology and action recognition technology based on deep learning. Traditional action recognition technology is keen to extract solid action features from video and train the

classifier for the next classification, but action recognition based on in-depth learning is to design a more reasonable end-to-end neural network and design more effective network training methods. The previous methods of action recognition can be divided into two types: recognition methods based on low-level action features and those based on high-level meaning information [11].

*2.3.1. Behavior Recognition Based on Low-Level Features.* The typical action recognition method based on low-level features firstly extracts the spatiotemporal interested points from the video, extracts the behavior characteristics of the area around the interested points, and then uses the model such as visual word package to model the extracted behavior characteristics. Finally, classifiers (support vector machine, hidden Markov model, etc.) are used to classify actions. Here are three basic aspects of action awareness. Three basic aspects are action feature extraction, action space description, and action classification [12].

The extraction of behavior features in video usually depends on the detection operator of interest points. The detection process of interest in time and space uses constructed detection operators to detect spatiotemporal response points in different positions of 3D spatiotemporal data in video. In this paper, a method of detecting spatiotemporal feature points based on 3D nested head is proposed. In this method, the spatial-temporal location of feature points is selected by calculating the nested matrix of multiscale 3D video data. This method greatly reduces the complexity of feature point detection time [13, 14].

After detecting the spatiotemporal feature points, the feature points must be executed near the feature points. In order to record three data, two histograms are counted in each time-space grid. If the action feature is extracted, in order to get a more standardized description of the whole video, it needs to be modeled. Tracking the trajectory of human movement, each trajectory has a clear physical meaning. Trajectories contain clear structural information, which can strengthen the relationship between points of interest in space and time. In recent years, the orbit tracking technology based on optical flow method has achieved excellent experimental results in the field of action recognition. This method can extract the motion characteristics along the optical flow orbit. When the camera is still, the optical flow represents the motion state of the human body. Compared with the method based on spatiotemporal interest, the method based on optical flow can eliminate background interference and obtain more motion information.

*2.3.2. Behavior Recognition Based on High-Level Semantic Information.* The action recognition model based on high-level meaning information uses a series of basic action attributes and meaning information and uses video to show people's actions. In this method, spatial local information or semantic information is usually used to build action charts, define a series of behavior attributes, and map human behavior to attribute space for classification. The method of action recognition based on high-level meaning information is the performance of human action recognition. This method has strong resistance to the changes of okra and light, is suitable for recording complex human actions, and has better applicability in various monitoring environments [15].

*2.3.3. Behavior Recognition Based on Deep Learning.* In the past, we needed to design corresponding action characteristics according to experience and application scheme. This kind of action characteristic has high recognition effect and low generality only in specific environment. The technique based on deep learning has made amazing achievements in the field of computer vision. Deep neural network has achieved better recognition results in the field of action recognition than before and greatly promoted the development of the field of action recognition. After the epoch-making results of neural networks in the field of image recognition and language recognition, researchers extend neural networks to the task of human action recognition based on video. Recognition rate will be higher, but it will be limited by a large amount of video data. Such a technique requires high computing power, and network training also requires certain skills [16, 17].

*2.4. Behavior Detection.* There is no clear line between actions, and the time difference between different action spans is very large. For the research in this field, more complex processing strategies are needed. The mainstream long video motion detection methods can be divided into two categories. The first method is to analyze each frame or several consecutive images, respectively, and use the time domain smoothing technology to integrate the recognition results of the whole video. The action function of 3D convolution neural network is used to train the current neural network and classify each function. After the whole video sequence is predicted, the output of the current neural network is postprocessed to obtain the sample operation category of the video and determine the sample time boundary, respectively. By connecting multiple inputs and outputs to simulate the time relationship, the category labels are allocated to each action video frame, and multiple labels are placed closely in the video sequence. A long-term and short-term memory depth network is proposed. The above method requires additional smoothing and integration techniques to obtain the position of the action [18].

The second method is to use the specified mechanism to generate the segments where the specified action occurs [19] and classify the specified segments into independent classification models. This method is called a multistage method. In the first stage, advanced timing candidate technology is used to obtain candidate video clips with high reproduction rate and determine the positions of these clips [20]. In the second stage, the specified video clip is sent to the classification model to determine the action category. From the perspective of detection effect, the recognition rate of the second method is generally higher than that of the first method [21]. However, in the multilevel method, since the assignment and classification of action segments are treated

as two separate processing levels, the training and collaborative optimization cannot be adjusted in different stages, and the calculation can be repeated between multiple stages [22, 23].

*2.5. Application of Multimodal Biometrics.* Based on the above understanding of the concept of multimodal biometrics, compared with single multimodal biometrics, it shows many unique advantages. With the gradual reduction of cost and maturity of technology, multimodal biometrics will move from government and military to market and civilian. The main application places are as follows:

(1) Criminal investigation: this is the place where multimodal identification technology is the earliest and most widely used. Because the scene of criminal investigation is easy to destroy and due to other reasons, only one part of the identity feature is often collected. At this time, multimodal identification can provide more professional retrieval [24].

(2) Finance: as the financial field involves a lot of property privacy, it is particularly important to ensure identity authentication [25].

(3) Access control: this is a place where single-mode fingerprint identification is widely used, but it is often difficult to carry out fingerprint identification due to other hand-held items, etc., so using multimode identification will have a great experience improvement.

(4) Social security: in order to ensure the stability and development of the society, it is particularly important to receive pension for personal identity certification. However, multimodal fusion recognition ensures high efficiency and safety and strictly restricts the relationship between human and card [26].

*2.6. Human Behavior Recognition Algorithm*

*2.6.1. Convolution Feature Extraction.* Because the layer function of CNN model is not affected by spatial information and location information, this section uses the functional map of convolution layer as image function, so as to retain the spatial structure information of image frame to a certain extent. Specifically, in the case of convolution layer with D channel, the size of each feature graph is $k \times K$, and, through the forward propagation process of CNN model, the video composed of $D$ frame image will produce feature $x = (x1, X2, ..., XR) \in RN$.

In the model, BN-Inception is selected as CNN model of feature extraction. BN-Inception model is an upgraded version of GoogLeNet model, in which batch normalization is added. In-depth training, in order to simulate complex problems, it usually increases the depth of the network and increases the number of training parameters, so as to improve the difficulty of training. Because of the strong combination between the two layers of the depth network, if the parameters of the upper layer are changed, the input data distribution of the next layer will change. This phenomenon is called internal covariate shift. It is suggested to normalize the input data in batches to keep the same distribution in the model training process.

In the test phase, because there is only one or more samples in the input, if there is no small batch of data, the average and variance cannot be calculated. At this time, the average value and dispersion value of each small batch in the training stage can be recorded, and the corresponding mathematical expectation value can be calculated as the average value and dispersion value of BN layer in the test stage [27, 28].

The average and scattered calculation formula and transformation and reconstruction formula in the test stage are as follows:

$$\begin{aligned} E[x] &\leftarrow E_B(\mu_B), \\ \mathrm{Var}[x] &\leftarrow \frac{m}{m-1} E_B\left(\sigma_B^2\right), \\ y &= \frac{\gamma}{\sqrt{\mathrm{Var}[x] + \varepsilon}} \cdot x + \left(\beta - \frac{\gamma E[x]}{\sqrt{\mathrm{Var}[x] + \varepsilon}}\right). \end{aligned} \tag{1}$$

*2.6.2. Convolutional Gated Recursive Unit.* Video consists of a series of image frames which can represent the timing information of the action process, so RNN model is needed to model the action information. Note that the generation of heat map is related to the hidden state $h_{t-1}$ of RNN. However, the traditional RNN can only deal with one-dimensional feature vector and cannot encode three-dimensional feature map including structure information, so it cannot make full use of the spatial and temporal information of image frame. In order to keep the spatial information changing with time, this section uses convolution operation to replace the vector operation of GRU. The formula of collapse GRU (collapse GRU, convgru) is as follows:

$$\begin{aligned} Z_t &= \sigma\left(W_Z * \widetilde{X}_t + U_z * H_{t-1} + b_z\right), \\ R_t &= \sigma\left(W_r * \widetilde{X}_t + U_r * H_{t-1} + b_r\right), \\ \widetilde{H}_t &= \tanh\left(W_h * \widetilde{X}_t + U_h * \left(R_t \Theta H_{t-1}\right) + b_n\right), \\ \widetilde{H}_t &= Z_t \Theta H_{t-1} + (1 - Z_t) \Theta \widetilde{H}_t. \end{aligned} \tag{2}$$

$^*$ represents convolution operation, and $\Theta$ represents Hadamard product. $W$ and $U$ are the convolution kernels of two-dimensional convolution operation, and $b$ is the offset term. These are the model parameters that need to be learned in the training phase and shared with time. The two control gates $Z$ and $R$, the candidate hidden state, and the hidden state are all three-dimensional. $X$ is the characteristic image of the $t$-th frame image processed by the spatial attention module. In addition, zero filling must be performed before convolution processing so that the hidden state of Gong does not change the dimension during convolution.

*2.6.3. Spatial Attention Module.* In order to generate the spatial attention heat map corresponding to the current input characteristic map, a small network with two

convolutions is designed, and the spatial attention heat map is calculated by using the softmax layer:

$$S_t = U_s * \tanh\left(W_{xs} * X_t + W_{hs} * H_{t-1} + b_s\right) + b_{us}. \quad (3)$$

$S_t$ is the characteristic image of frame $t$, $H_{t-1}$ is the last hidden state, and $U_s$, $W_{xs}$, and $W_{hs}$ are folded cores, with the size of $1 \times 1$. Then, the 2D soft maximization layer is used in $S_t$ to obtain the space attention heat map corresponding to the current input $X_t$. The formula of 2-dimensional element is as follows:

$$A_t^{ij} = \frac{\exp\left(S_t^{ij}\right)}{\sum_k \sum_t \exp\left(S_t^{kl}\right)}. \quad (4)$$

At this time, $A_t^{ij}$ is the spatial attention weight on position $(i, j)$ of the input characteristic graph $X_t$, the size is between 0 and 1, and the weight sum of all positions is $l$. At this time, pay attention to the heat map corresponding to the space input of the current input $X_t$. $X_t$ can be obtained by multiplying the heat map and each feature map on the corresponding elements.

$$\tilde{X}_t = A_t \Theta X_t. \quad (5)$$

## 3. Human Action Recognition Experiment with Multimodal Features

*3.1. Dataset.* In the experiment, two valuable datasets UCFl01 and HMDB51 were used. These datasets are obtained in an uncontrolled environment and in the real world. HMDB51 is a medium-sized dataset, and UCFl01 is a large-scale dataset.

HMDB51 dataset includes 11 action categories: football shooting, trotting, chasing, fast running, intercepting, passing, heading, goalkeeper catching, midfield serving, player passing, and referee and ball. For each category, the video is divided into 25 groups, each of which contains more than 4 action clips.

*3.1.1. UCFl01.* This is an action recognition dataset containing 50 action categories, which is derived from the video obtained by HMDB51. For all 50 categories, the video is divided into 25 groups. The real scene and simulation scene of football are shown in Figure 1.

In the two datasets, videos in the same group have some common features, such as the same person, similar background, and similar viewpoint. The model of this study is evaluated in the original video.

*3.2. Parameter Setting*

*3.2.1. Convolution Feature Extraction.* With the development of deep learning, there are more and more deep network layers. Many research studies indicate that the performance of deep network will be improved. Because the accuracy and calculation efficiency are very good, BN-Inception is selected. As a basic feature extraction network, perception plays an important role, and pretraining is conducted through ImageNet dataset. Like the dual flow network, the convolution network with deep space receives the image frame as the input, and the convolution network with deep time series receives the stack operation flow as the input.

*3.2.2. Spatial Attention Module.* Using BN-Inception, the function diagram of the last concept module is used as input, and the size of the function diagram is $7 \times 7 \times 1024$. The size of the folded core of convgru is $3 \times 3$. The convolution kernel size for generating thermal map is $1 \times 1$. The number of channels for both convgru and the space service module is 1024.

For sequential attention mechanism, the number of hidden units of bidirectional GRU is 1024. 3 layers of insight are used to generate timed attention. The activation function of the first two layers is tanh, and the last is signal mode, which is used to compress the output to a value of 0-l. Table 1 shows the experimental platform and software versions.

It is a mini in training. The batch size is set to 10, randomly selected from the training package. Each video is first divided into 25 segments; each segment selects one frame of image or stack operation stream as input and expands the input data. There are about 50 video frames in the HMDB51 dataset, and the image is less than 25, which satisfies the final frame image or corresponding optical flow. Penalty coefficient A is set to 10, and a pull-down layer is added before the fully connected layer for classification. The pull-down values of deep convolution network in space and deep convolution network in optical flow are set to 0.8 and 0.7, respectively. If the probability gradient descent method is used as the photoconductor, the initial learning rate is set to $10^{-2}$, and every 20K iterations are reduced to 1/10 of the original. The maximum number of iterations is 100k. The learning efficiency of convolution function extraction network is $10^{-6}$, and the learning efficiency in training is unchanged.

## 4. Human Action Recognition Analysis of Learning Quality-Related Multimodal Features

*4.1. Multimodal Feature Fusion Analysis.* In order to verify the role of the proposed model in multimodal function fusion, four sets of comparative experiments were carried out: (1) without multimodal function fusion: after completing the training of spatial depth network and time series depth network, classification probability fusion was directly used to get the result of action recognition; (2) multimodal function fusion using the corresponding elements to get the average value; (3) maximum using the corresponding elements and multimodal functional fusion of values; and (4) functional fusion using fully connected loops. Table 2 shows the accuracy comparison of multimodal function fusion methods.

Table 2 shows the recognition accuracy of four comparison experiments on two datasets. It can be seen from the results in the table that the multimodal feature fusion


Figure 1: Real scene and simulation scene of football.

Table 1: Experimental platform and software version.

| Operating system | Ubuntu 14.04 LTS 64bit |
|---|---|
| CPU | Intel(R) Core(TM)i7—4790K CPU@4.00 GHz |
| GPU | NVIDIA GeForce GTX TITAN X GPU 12 GB |
| RAM | 32 GB |
| CUDA | 9.0 |
| PyTorch | 0.4.1 |

Table 2: Accuracy comparison of multimodal feature fusion methods.

| Dataset | Spatial depth network | Sequential depth network | Multimodal feature fusion | | | |
|---|---|---|---|---|---|---|
| | | | (1) | (2) | (3) | (4) |
| UCFl01 | 85.8 | 87.1 | 93.4 | 94.5 | 93.9 | 95.2 |
| HMDB51 | 56.3 | 61.8 | 71.3 | 70.9 | 73.4 | 71.7 |

method proposed in this study has a great improvement effect on the task of action identification, and there is no obvious gap between the multimodal feature fusion model and the other three models. Figure 2 shows the accuracy analysis of multimodal fusion.

From Figure 2, it can be seen that the accuracy of UCFl01 dataset is generally high, more than 80%, and the highest value reaches 95.2%; and the accuracy of HMDB51 dataset is about 70%, and the lowest is 56.3%. Compared with these two datasets, this research method can get a conclusion with higher accuracy in multimodal feature acquisition, with good effect.

At the same time, among the three fusion methods, the fully connected pool fusion improves the performance and achieves the highest recognition accuracy of the two datasets. Compared with the model without multimodal function fusion, the accuracy of this recognition has increased by 1.8% and 2.3%. In addition, the fusion method using the maximum value of corresponding elements is the worst of the three.

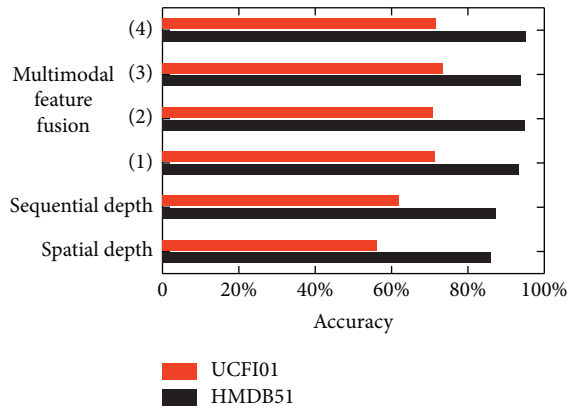


Figure 2: Accuracy of multimodal fusion.

*4.2. Trend Analysis of Multimodal Fusion Identification Technology.* Each biological feature has its limitations, so it is difficult to find a single-mode biological feature to meet all the above requirements. Traditional biometric technologies, such as fingerprint and vein recognition, are difficult to collect in the case of insensitive collector, dirty area, or finger injury, and these features can be imitated for authentication; the face is also vulnerable to changes in expression, light, age, and so forth. Figure 3 shows the performance comparison of common human behavior characteristics.

To sum up, the single-mode biometrics show limitations in acquisition, matching accuracy, easiness to crack, and environmental adaptability. The research of multimodal fusion technology is the inevitable trend of the high-speed development of the information society. Therefore, biometric technology is increasingly developing in the direction of multifeature fusion; one of the important ways is multimodal biometric technology.

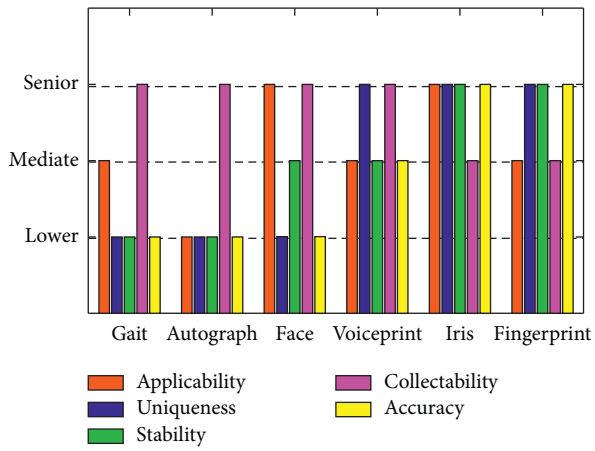


Figure 3: Performance comparison of common human behavior characteristics.

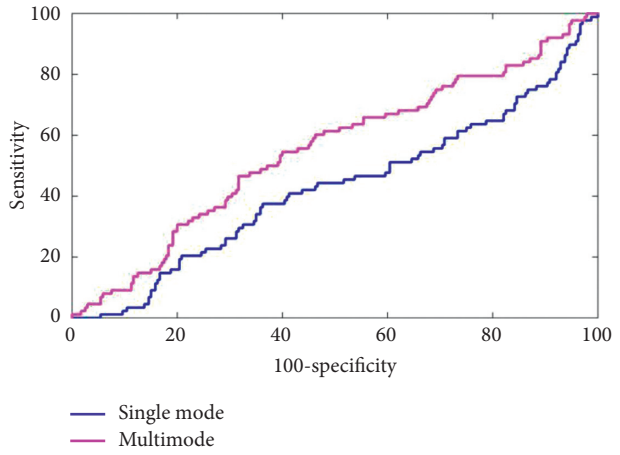


Figure 4: Single-mode and multimode feature recognition rate.

Multimodal biometrics can overcome some shortcomings of single-mode biometrics by fusing more than two single-mode biometrics (or action functions) as data objects. It has higher stability and safety, and the authentication process is more accurate and safe. Multimodal diffusers can also overcome the loss of biometric features such as cut, scratch, dry, or inborn unknown features.

*4.3. Comparative Analysis of Single-Mode and Multimode Performance.* ROC diagram of single-mode and multimode feature recognition rate is shown in Figure 4.

It can be seen from Figure 4 that the relative area of purple curve in ROC curve is large, and the recognition accuracy of single-mode feature is significantly lower than that of multimode feature. In the experiment, fingerprint recognition algorithm uses the recognition algorithm based on feature dimension expansion, and voiceprint recognition uses the classic GMM model with a mixed number of 128. In order to observe the effect of the fusion, the fingerprint and voiceprint recognition algorithm has no targeted optimization measures, so the recognition rate is not high in their single-mode state.

## 5. Conclusion

This research takes human behavior cognition in football as the research background and makes a model evaluation. Aiming at the importance of spatial information and the complementarity of multiple features in video classification, a human action recognition method based on multimodal feature fusion is proposed. Most of the most advanced video motion detection methods are based on 3D convolutional neural network. 3D convolutional neural network has strong generalization ability and rich expressive force, but it has many model parameters, large memory consumption, slow training, and difficult convergence. In the future, we can study more optimized convolution form and more optimized network structure, further reduce the complexity of the model, reduce the amount of GPU memory, and obtain faster and accurate work detection methods.

In this model, two flow convolution neural networks are used to extract static and dynamic action features of RGB and optical buoy, respectively. Attention mechanism is added to current neural network (RNN) and action features are classified. The experimental results of two large-scale challenging action recognition databases show that the behavior classification model based on attention mechanism effectively distinguishes video frames of different importance and achieves more accurate and effective results compared with other most advanced nonattention models.

When multimodal functions are selected, only one function has insufficient expressiveness, which needs to rely on the complementarity of multiple functions to strengthen each other. Therefore, the method of maximizing the complementarity of various functional information is worth exploring. This is also one of the important points of learning on this topic. Function descriptors accept multimodal images as input and use the advantages of CNN and LSTM to extract temporal and spatial functions and global action functions. The results of UCF-50 dataset show that this method is better than most of the existing descriptors.

## Data Availability

No data were used to support this study.

## Conflicts of Interest

The authors declare that they have no conflicts of interest.

## References

[1] A. A. Liu, N. Xu, W. Z. Nie, Y.-T. Su, Y. Wong, and M. Kankanhalli, "Benchmarking a multimodal and multiview and interactive dataset for human action recognition," *IEEE Transactions on Cybernetics*, vol. 47, no. 7, pp. 1–14, 2016.

[2] M. Sivarathinabala, S. Abirami, M. Deivamani, and M. Sudharsan, "A smart security system using multimodal features from videos," *Pattern Recognition and Image Analysis*, vol. 29, no. 1, pp. 89–98, 2019.

[3] B. Sun, L. Li, G. Zhou, and J. He, "Facial expression recognition in the wild based on multimodal texture features,"

*Journal of Electronic Imaging*, vol. 25, no. 6, pp. 061407.1–061407.8, 2016.

[4] C. Ahmad, K. Michael, and N. Tamim, "Multimodal radiomic features for the predicting gleason score of prostate cancer," *Cancers*, vol. 10, no. 8, p. 249, 2018.

[5] M. Merler, K.-N. C. Mac, D. Joshi et al., "Automatic curation of sports highlights using multimodal excitement features," *IEEE Transactions on Multimedia*, vol. 21, no. 5, pp. 1147–1160, 2019.

[6] P. Gunjan, T. N. Nagabhushan, P. Bhanu, and R. Pushkarna, "Early detection of Parkinson's disease through multimodal features using machine learning approaches," *International Journal of Signal and Imaging Systems Engineering*, vol. 11, no. 1, p. 31, 2018.

[7] A. J. Rong, K. C. Fan, B. Golshani et al., "Multimodal imaging features of intraocular foreign bodies," *Seminars in Ophthalmology*, vol. 34, no. 7-8, pp. 1–15, 2019.

[8] Z. Jiang, T. Zhang, X. Liu et al., "Multimodal imaging features of bilateral choroidal ganglioneuroma," *Journal of Ophthalmology*, vol. 2020, no. 3, 8 pages, Article ID 6231269, 2020.

[9] L. A. Dalvin, C. L. Shields, D. A. Ancona-Lezama et al., "Combination of multimodal imaging features predictive of choroidal nevus transformation into melanoma," *British Journal of Ophthalmology*, vol. 103, no. 10, pp. 1441–1447, 2019.

[10] C. Faure, M. Paques, and I. Audo, "Electrophysiological features and multimodal imaging in ritonavir-related maculopathy," *Documenta Ophthalmologica*, vol. 135, no. 3, pp. 241–248, 2017.

[11] B. O. Muhammed and S. I. Shamsuddin, "A multimodal biometric system using global features for identical twins identification," *Journal of Computer Science*, vol. 14, no. 1, pp. 92–107, 2018.

[12] D. Q. Li, J. Golding, C. Glittenberg, and N. Choudhry, "Multimodal imaging features in acute exudative paraneoplastic polymorphous vitelliform maculopathy," *Ophthalmic Surgery, Lasers and Imaging Retina*, vol. 47, no. 12, pp. 1143–1146, 2016.

[13] T. Kim, B. Kang, M. Rho, S. Sezer, and E. G. Im, "A multimodal deep learning method for android malware detection using various features," *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 3, pp. 773–788, 2019.

[14] S. Sayeed, I. Nasir, and T. S. Ong, "An efficient multimodal biometric authentication integrating fingerprint and face features," *American Journal of Applied Sciences*, vol. 13, no. 11, pp. 1221–1227, 2016.

[15] C. Chen, R. Jafari, and N. Kehtarnavaz, "A survey of depth and inertial sensor fusion for human action recognition," *Multimedia Tools and Applications*, vol. 76, no. 3, pp. 4405–4425, 2017.

[16] C. Chen, R. Jafari, and N. Kehtarnavaz, "A real-time human action recognition system using depth and inertial sensor fusion," *IEEE Sensors Journal*, vol. 16, no. 3, pp. 773–781, 2016.

[17] X. Cai, W. Zhou, L. Wu, J. Luo, and H. Li, "Effective active skeleton representation for low latency human action recognition," *IEEE Transactions on Multimedia*, vol. 18, no. 2, pp. 141–154, 2016.

[18] S. Yu, Y. Cheng, S. Su, G. Cai, and S. Li, "Stratified pooling based deep convolutional neural networks for human action recognition," *Multimedia Tools and Applications*, vol. 76, no. 11, pp. 13367–13382, 2017.

[19] H. Song and M. Brandt-Pearce, "A 2-D discrete-time model of physical impairments in wavelength-division multiplexing systems," *Journal of Lightwave Technology*, vol. 30, no. 5, pp. 713–726, 2012.

[20] S. Zhou, L. Chen, and V. Sugumaran, "Hidden two-stream collaborative learning network for action recognition," *Computers, Materials & Continua*, vol. 63, no. 3, pp. 1545–1561, 2020.

[21] A. Tharwat, H. Mahdi, M. Elhoseny, and A. E. Hassanien, "Recognizing human activity in mobile crowdsensing environment using optimized k -NN algorithm," *Expert Systems with Applications*, vol. 107, pp. 32–44, 2018.

[22] J. Zhang, Y. Han, and J. Jiang, "Tucker decomposition-based tensor learning for human action recognition," *Multimedia Systems*, vol. 22, no. 3, pp. 343–353, 2016.

[23] C. Beaudry, R. Péteri, and L. Mascarilla, "An efficient and sparse approach for large scale human action recognition in videos," *Machine Vision and Applications*, vol. 27, no. 4, pp. 529–543, 2016.

[24] M. M. Moussa, E. E. Hemayed, H. A. E. Nemr et al., "Human action recognition utilizing variations in skeleton dimensions," *Arabian Journal for Science & Engineering*, vol. 43, no. 5, pp. 1–14, 2017.

[25] M. M. Laruccia and V. L. Martyniuk, "Racism in football: a narrative path," *Advances in Journalism and Communication*, vol. 04, no. 4, pp. 103–112, 2016.

[26] E. Vargese, M. Galliara, and M. Srivastava, "OSCAR foundation: empowering lives through football," *Emerald Emerging Markets Case Studies*, vol. 6, no. 3, pp. 1–32, 2016.

[27] J. B. Julian, J. Ryan, R. H. Hamilton, and R. A. Epstein, "The occipital place area is causally involved in representing environmental boundaries during navigation," *Current Biology*, vol. 26, no. 8, pp. 1104–1109, 2016.

[28] M. Chapman, "Chicano signifyin': appropriating space and culture in el henry," *Theatre Topics*, vol. 27, no. 1, pp. 61–69, 2017.