Hindawi
Mathematical Problems in Engineering
Volume 2022, Article ID 5257165, 10 pages
https://doi.org/10.1155/2022/5257165

Research Article

Research on Dance Movement Recognition Based on Multi-Source Information

Yunchen Wang

Anhui Normal University, Wuhu 241000, China

Correspondence should be addressed to Yunchen Wang; wyc@ahnu.edu.cn

Received 23 January 2022; Revised 28 February 2022; Accepted 5 March 2022; Published 23 April 2022

Academic Editor: Naeem Jan

A huge number of scientific research institutions and scholars are now researching this topic in depth, with promising results. Meanwhile, research development in dance visual frequency movement detection is rather modest due to the high complexity of dance movement and the challenges of human body self-shielding in dance performance. Aiming at the problem of the combination of motion recognition and dance video, the feature extraction, representation, and motion recognition methods based on dance video are emphatically studied. This paper studies an effective feature extraction method according to the characteristics of dance movements. Firstly, each dance movement video in the data set is separated into equal sections, and the edge characteristics of all video pictures in each segment are gathered into one image, from which the direction gradient histogram features are extracted. Secondly, a group of directional gradient histogram feature vectors is used to represent the local appearance information and shape features of the video dance moves. In view of the existing problem of heterogeneous feature fusion, this paper chooses the multi-core learning method to fuse the three kinds of features for dance movement recognition. Finally, the effectiveness of the proposed dance movement detection algorithm is tested using the Dance DB data set from the University of Cyprus and the Folk Dance data set from my laboratory. Experimental results show that the proposed algorithm can maintain a certain recognition rate for relatively complex dance movements and can still ensure a certain accuracy when the background and target are easily confused. This also confirms the efficacy of the movement recognition system used in this paper for recognizing dance movements.

1. Introduction

Human pose estimation is a key technology in the field of human motion recognition. Its principle is to recognize human pose by extracting features from images. This technique can be used for intelligent dance training, which can obtain a dancer's posture skeleton by extracting the dancer's image features. So as to identify the dancer's dance movements, evaluate and correct the dancer's posture. Early human pose estimation mainly focused on human contour features or component models. Su Yanchao et al., for example, used a boosting classifier to extract edge field characteristics and created a human pose estimation system based on component detection. For human pose estimation, Han Guijin et al. suggested an appearance model incorporating histogram of oriented gradients (HOG) and color features.

In the discipline of computer vision, motion recognition research is a difficult topic to tackle. Its goal is to examine the video footage and recognized human motion using image processing and classification recognition technology. Motion recognition has become a very popular study direction in recent years due to its high research value, and it has drawn a significant number of scientific research institutes and scholars to engage in scientific research in this area. Motion recognition technology can be used in a variety of video scenarios. It is now widely employed in a variety of sectors, including intelligent monitoring, virtual reality, and intelligent human-computer interaction, among others.

1.1. Intelligent Monitoring. Video surveillance provides important security for our lives by recording everything that happens within the range of surveillance. Therefore, at

present, video surveillance is widely used in stations, banks, shopping malls, airports, public places, and other occasions with large traffic and high-security requirements. But, at present, most video monitoring systems rely too much on manual operation, which is usually abnormal. Usually, after the occurrence of abnormal or emergency situations, relevant monitoring videos are found through large-scale screening and analysis. Manual operation will reduce the monitoring efficiency and accuracy. Therefore, in case of emergencies such as robbery and fight, it is impossible to make timely and effective emergency response. The intelligent monitoring system based on motion recognition technology can effectively avoid the disadvantages mentioned above by using motion-recognition-related technology. In the face of emergencies, abnormal events can be identified through real-time analysis of human behavior in surveillance videos. In this way, timely warnings can be issued to improve the response speed and efficiency of relevant personnel and reduce the harm of emergencies to the maximum extent, so as to provide security for public places.

1.2. Intelligent Human-Computer Interaction. Intelligent human-computer interaction refers to the user does not need to use the mouse, keyboard, and other traditional input devices, but through human body movements or gestures to interact with the computer, in which motion recognition technology provides technical support for human-computer interaction applications. Human-computer interaction is a promising application direction in motion recognition, especially in motion-sensing interactive games. The popularity of motion-sensing games is due to the continuous development of human-computer interaction, which makes such technology into our daily life become a reality. In games, players do not need to operate the mouse and keyboard like traditional games but only need to make corresponding body gestures in front of the camera according to the rules of the game. The game uses the camera to capture the actions of the game players and then uses the human-computer interaction system to identify various actions to perform tasks in the game. In 2010, Microsoft developed a Kinect sensor, which realized the integration of voice, motion, and facial recognition functions; then used RGB and infrared cameras to capture human movements; and then analyzed and recognized their movements, thus realizing intelligent human-computer interaction.

1.3. Video Retrieval. With the rapid development of the Internet and multimedia technology, there are a large number of video data on the Internet. How to retrieve useful information from these videos by manual operation will become extremely difficult and inefficient. The content-based video retrieval mechanism developed in recent years has greatly improved the experience of video retrieval. Its technical principle is to use action recognition technology to learn some specific patterns in video data and generate summaries based on video content through these patterns. At present, content-based video retrieval technology develops rapidly, and content-based video summarization and retrieval have attracted the attention of a large number of researchers around the world. The application of visual frequency retrieval technology has been commercialized cases, such as sports video retrieval, and summary is a very attractive application in video retrieval.

1.4. Action-Assisted Analysis. At present, motion assist analysis is mainly used in sports athlete training and competition, patient rehabilitation, and automobile driver assistance driving. The sports training auxiliary system based on motion recognition technology can analyze the movements of athletes in daily training and sports competitions and compare with the standard movements to correct the wrong movements in time, so as to improve the quality of athletes' training and sports competitions. In the aspect of sports rehabilitation, the patient's rehabilitation status can be judged by comparing the patient's posture and normal movement. The vehicle assistance driving system based on motion recognition can make a real-time analysis of the driver's motion according to the sensor and computer to judge whether the operation conforms to the standard.

In recent years, research on video technology has become one of the hot spots in the academic circle. Among many video technologies, video motion recognition is of great significance to the intelligent application of video, and it has been widely applied in many fields. Video information is normally extracted in two steps: first, the video's relevant visual features are extracted; second, the extracted features are learned and matching description labels are generated. The most important element with this technology is to extract features properly, and one of the most efficient approaches for extracting video features is to use a deep learning algorithm. Traditional extraction approaches based on this method, on the other hand, focus on the spatial domain of video, that is, the extraction of pixel information in the video frame, while ignoring the change in motion state in the time domain of video action. Taking human cognition as an example, the actions of objects are constantly changing. People should not only judge the categories of actions by relying on the static pictures of the actions but also pay attention to the whole process of changes from the beginning to the end of the actions. As one of the important algorithms in artificial intelligence technology, deep learning has almost the same way of extracting traits as human beings.

Dance not only gives people a feeling of beauty but also has the function of fitness and entertainment. Dance training is a complex and comprehensive process, which mainly trains the extension of the trainers' body joints. Therefore, how to ensure the precision and elegance of dancing poses is the difficulty of dance practice [1]. Traditional dance learning relies more on one-way information transmission and lacks a feedback mechanism, so it is difficult to meet learning needs [2]. With the development of technology, through the combination of dance training and movement recognition technology, the data in dance training movements can be captured for standardized matching and comparison so that the trainers can find the gap more intuitively, thus ensuring the accuracy of dance posture training [3–5]. Kinect, as a natural interactive device

with advanced visual technology, realizes real-time body and bone tracking, motion capture, and voice input through Kinect, a motion-sensing peripheral. At the same time, the interactive learning model provided by Kinect, based on the imaging principle of light coding technology, has changed the embarrassment of traditional sensor wearing and made people's movement reflect into the computer virtual world, effectively improving the real experience of dance practice for trainers and meeting the learning pleasure of different ages [6]. Based on this, this paper designs an interactive dance system based on Kinect on the basis of relevant research. Through Kinect, dance movements can be recognized and collected to ensure the accuracy of training and improve the interactivity of dance learning.

The deep application of computer-aided instruction not only enriched the teaching content but also made up for the disadvantages of traditional teaching methods. The effective combination of motion capture technology and dance teaching is a manifestation form of computer-aided instruction, which realizes the perfect combination of 3D virtual world and reality technology. In recent years, researchers have made extensive research on the application of motion capture. Some combined motion capture technology with a Chinese puppet show and put forward a set of digital puppet show technical schemes. Some researchers employed a motion capture device to analyze golfers' hip and torso movements during swings and quantify the association between the torso and hip movement and swing, providing theoretical justification for golf scientific instruction. Some proposed animation production methods and processes based on motion capture technology, animation synthesis, and the elimination of slippage in the animation production process. Some proposed animation production methods and processes based on motion capture technology, animation synthesis, and the elimination of slippage in the animation production process. This shows the wide application of motion capture technology [7, 8].

As a form of artistic performance, dance is a kind of body movement accompanied by rhythm, and more and more people are beginning to devote themselves to dance learning. However, it is not easy to learn dance, such as the individual differences of learners, the mastery of rhythm, coach and student psychology, and other factors that will affect the learning progress. To improve this situation, the author puts forward using motion capture technology to assist in the dance of teaching and research; to capture the dance performers of three-dimensional motion data into digital abstract movement, establishing a database of 3D motion of the human body; to generate animation teaching; to change the disadvantages of traditional dance teaching; and improve the teaching quality of education that has important significance [9].

It can be shown that the research findings of movement recognition technology based on dance video can be used for educating, protecting, and excavating creative and cultural legacy, as well as for dance experts to evaluate dance video. Furthermore, the research on emotion detection methods based on dance films will serve as a point of reference for existing research on human motion recognition in a variety of realistic and complicated contexts, enriching the motion recognition technology application sector. Therefore, after analyzing the related problems of movement recognition technology and dance movement recognition, this paper mainly studies movement recognition based on dance video [10].

This paper is organized such that Section 2 defines the research status of motion recognition. Section 3 proposes the dance movement recognition based on multi-feature fusion. Section 4 consists of the analysis of experimental results and data. Finally, the paper ends with a conclusion in Section 5.

2. Research Status of Motion Recognition

In recent years, a considerable number of domestic and international research organizations, as well as related researchers, have devoted themselves to video-based motion recognition research and have made significant contributions to the field's progress. At present, many related projects have been started in foreign countries. As early as 1997, the Defense Advanced Research Project (DARP) from the United States carried out the visual surveillance project with the efficient participation of Carnegie Mellon University and many other institutions, which was mainly engaged in the research of automatic analysis of war and civilian scene surveillance videos. In the same year, Kyoto University in Japan participated in the Cooperative Distributed Vision (CDV) project, which mainly studied the realization of dynamic scene understanding, target tracking, and motion perception through the use of motion sensors. It has been successfully applied in real-time monitoring, teleconference, and intelligent navigation.

The "scene understanding" project of the University of Reading in the UK has conducted relevant research on vehicle and pedestrian tracking and their interaction. International Business Machines (IBM) and Microsoft and other well-known foreign enterprises have successfully developed a large number of applications and commercialization of facial recognition, gesture recognition, attitude estimation, and other technologies. For example, in 2010, Microsoft developed the Kinect sensor, which realized the combination of action and facial and voice recognition functions and realized the analysis and recognition of human movement and posture so as to achieve human-computer intelligent interaction. In 2011, Advanced Micro Devices (AMD) developed an application based on APU, which realized the recognition of human movements by connecting the sensor device. The core technology of this application is to use APU to analyze and calculate in real-time and control the computer in front of it [11]. In 2013, IBM, together with STMicroelectronics and other companies, developed a system that can control smart homes using human gestures.

Domestic research on motion recognition started relatively late. With the application of motion recognition becoming more and more extensive, many domestic universities and scientific research institutions have carried out relevant research on motion recognition. The system can analyze athletes' movements in training and simulate

relevant movements to help correct mistakes and improve their performance [12]. The Institute of Automation of the Chinese Academy of Sciences has carried out relevant research on intelligent surveillance video analysis and other technologies, mainly through the analysis of video image sequence to identify and track the target in the dynamic scene, on this basis to analyze the target behavior to facilitate immediate response in the case of emergencies. The Digital Image Processing laboratory of Shanghai Jiao Tong University has done in-depth research on intelligent monitoring, gait recognition, and motion recognition. The Institute of Artificial Intelligence of Zhejiang University has done a lot of research in the direction of human animation and redirects the movements of cartoon characters by using tracking and 3D reconstruction techniques. After years of vigorous development in the field of motion recognition, many important scientific achievements have been made. The overall research trend has gradually expanded from simple actions to the recognition and analysis of complex actions and from single person motion analysis to multiperson motion analysis. With the further development of the research on motion recognition in universities and scientific research institutions, motion recognition research is facing greater challenges as well as opportunities. Therefore, we believe that the research on motion recognition also has a good prospect for development.

Technology related to size measurement, positioning, and azimuth determination of objects in physical space can be directly processed by the computer [13]. A tracker is set in the key parts of the moving object, and the motion capture system is used to record the process of the object's movement. After processing by the computer, three-dimensional space coordinate data can be obtained.

2.1. Statistics-Based Methods. Previous hierarchical action recognition methods based on statistics mainly focus on the extension of the hidden Markov model and the application of dynamic Bayesian networks, such as the hierarchical hidden Markov model and the dynamic Bayesian network method. Shi et al. suggested a hierarchical action recognition approach employing the genetic network to analyze real-time and sequential subactions, based on the fact that subactions might occur simultaneously or continuously. Tang et al. model high-level complex actions with hidden state variables and time interval variables and learn complex actions with discriminant models. A four-layer probabilistic hidden state model was suggested by Yin et al. Spatiotemporal properties are initially recognized in this method, and then atomic actions are clustered using a hierarchical Bayesian model. Meanwhile, without specifying the number of hidden states, the hierarchical probabilistic hidden state model based on LDA is utilized to identify activities. We try to use the local temporal and spatial features of clustering as the representation of atomic action, hierarchical description, and complex action. The statistical method has a strict definition for the temporal order of two-state nodes, so this method can only be used to describe the temporal relationship but cannot model the co-occurrence relationship. Therefore, the statistics-based approach cannot efficiently model the temporal structure of complex actions.

2.2. Grammar-Based Approach. The grammar-based hierarchical approach is based on the premise that actions are represented by strings of symbols, each symbol corresponding to an atomic action, and these atomic actions can be recognized using either the hierarchical or single-layer approaches. The grammar-based technique has the drawback that grammar constraints confine the temporal order of atomic actions, making it impossible to identify simultaneous atomic operations. Joo et al. presented a random context-free attribute grammar with the goal of modelling constraint relations between features while expressing the sequence structure of atomic actions. Generally, grammar-based methods are divided into two layers. The main function of the bottom layer is to identify atomic actions of low-level actions, and the function of the top layer is to use grammar analysis to identify high-level actions. Ivanov et al. proposed a layered approach that uses a hidden Markov model to detect atomic actions at the bottom level and uses random grammar to identify complex actions at the top level. Another limitation of the grammar-based approach is that a set of grammar rules must be provided in order to overcome this limitation.

2.3. Description-Based Methods. Description-based hierarchical methods can explicitly model the spatiotemporal structure of human actions, which differs from the statistical and grammar-related methods described above. The description-based approach represents human actions as the appearance of some subactions, which need to satisfy specific temporal, spatial, and local relations. Description-based hierarchical action recognition usually uses context-free grammar to represent actions. Gupta et al. employed a tree or graph model to evaluate complicated activities and found that describing the causal relationship between atomic events enhanced recognition accuracy. To overcome the identification failures of low-level parts caused by the inadequacies of description-based approaches, Gupta et al. suggested the probability extension method of many identification methods. The bottom-up method was utilized to examine the hierarchical representation of complicated actions, and the clustering algorithm was used to examine atomic actions in subvideo segments and generate discriminative middle-level action phrases based on and or graphs.

3. Dance Movement Recognition Based on Multi-Feature Fusion

Feature extraction is usually the initial step in motion recognition research. Feature extraction is the process of extracting feature information from the active data set to define the target action in the video, which is an important stage in motion recognition research. It can be shown that the retrieved characteristics have a significant impact on the accuracy of the action recognition results as well as the

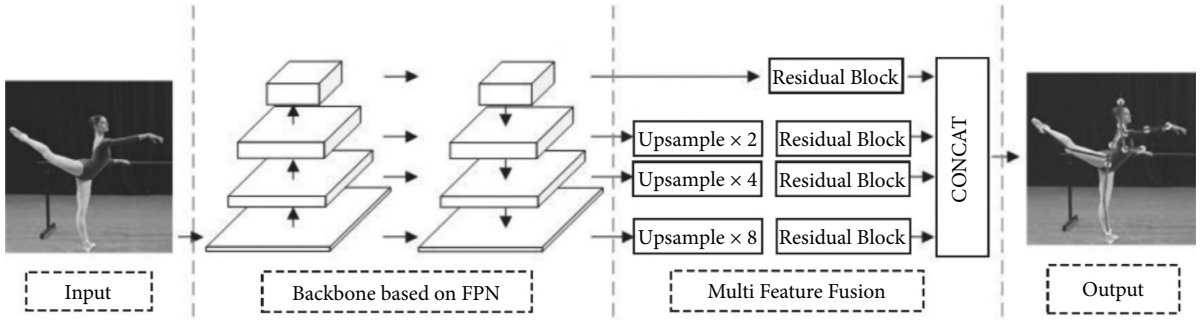

Figure 1: Schematic diagram of dance movement recognition algorithm based on multi-feature fusion.

method's robustness. The direction gradient histogram feature, optical flow direction histogram feature, and audio feature retrieved from dance films are used to depict dance movements in this article after completely evaluating the peculiarities of dance motions. The optical flow direction histogram features are used to describe the movement information of dance movements, while the direction gradient histogram features are used to characterize the local appearance and shape properties of dance movements. Additionally, dance movement recognition research should consider the impact of music on dancing. Dancers do a dance with music playing in the background, and the music style is tied to the genre of dance. Audio features, on the other hand, include a lot of information, making them a significant auxiliary element that can help lessen the impact of self-occlusion on dance motions. In this paper, the audio features are extracted from the corresponding audio files of dance movement video and combined with the above two features for dance movement recognition. The feature extraction process in this paper is mainly composed of three parts: the first part, which is used in this paper, from the accumulation of gradient direction histogram feature extracting edge character method; the second part mainly from the dance optical flow direction histogram feature extracting data concentration, the third part from the dance moves in the video to extract the corresponding audio stream file and then to extract audio from the audio stream file signature features.

Because of the correlation between music and dance, we can regard music as the soul of dance, and every dance is performed with the accompaniment of music. A dance video also contains audio information. We should fully explore the relationship between this audio information and dance movements for the research of dance movement recognition. Therefore, the influence of music in the video on dance movements should be considered when analyzing dance videos. Audio features contain a large amount of information, which is an important auxiliary feature. Moreover, the computational amount and storage space of audio features are much smaller than traditional visual features, which is an advantage for us to engage in dance movement recognition. At the same time, in this paper, after the research and analysis of the audio-related features, the audio signature feature is used as the audio feature in this research method.

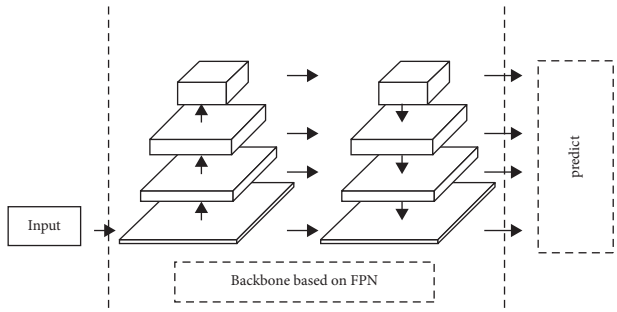

Figure 2: Schematic diagram of backbone network based on FPN.

The new algorithm uses a Feature Pyramid Network (FPN) [14] for feature extraction and then deepens the extraction for features of different scales. Finally, each feature is sampled up to the size of the original image for feature fusion as shown in Figure 1.

3.1. Backbone Network Based on Feature Pyramid. The backbone network in this paper is similar to ResNet [15], both of which contain five feet of feature extraction stages. The convolution features at different scales were represented as C1, C2, C3, C4, and C5 stages. Each stage was composed of residual blocks, which finally extracted features and generated a thermal diagram of key points of human body posture. The resolution of the shallow layer feature map of the model is large. With the deepening of the network, the feature map shrinks to one-half of the original after every downsampling operation, and the resolution also decreases. Therefore, the shallow features in C1, C2, and C3 have a high spatial resolution. However, it contains insufficient semantic information, while the deep features in C4 and C5 are opposite. In order to have the spatial resolution and semantic information of the feature layer at the same time, the FPN structure is adopted in this paper to further integrate the high- and low-level feature information.

The backbone network structure based on FPN is shown in Figure 2. This backbone network can effectively locate simple visible key points, but it is difficult to identify human posture key points in a complex environment, such as occlusion and hidden key points. In this paper, a multi-feature fusion module is designed to locate such complex key points, which usually require more abundant feature information.

3.2. Time-Space-Based Double Convolutional Neural Network. The feature extraction method in the spatial domain is consistent with that in the image information extraction method, and the features in the temporal domain are identified by Optical Flow. The circular neural network shows that the optical flow reflects the changing track of pixels after the motion state of objects in space changes, which is widely used in motion detection. The acquisition method is as follows:

For the spatial coordinate position $O(x, y)$ at time t, the pixel brightness of this point is $I(x, y, t)$ In dt time, the point moves to a new position $(x + dx, y + dy)$ in the next frame. At this point, due to the very short time, the relationship between the pixel brightness of this point is

$$I(x, y, t) = I(x + dx, y + dy, t + dt). \tag{1}$$

Its Taylor expansion is shown in the following equation:

$$\frac{\partial I}{\partial x}\frac{\Delta x}{Vt} + \frac{\partial I}{\partial y}\frac{\Delta y}{Vt} + \frac{\partial I}{\partial x}\frac{\Delta t}{Vt} = 0. \tag{2}$$

At this point, the optical flow equation of this point can be obtained, as shown in the following equation:

$$\frac{\partial I}{\partial x}V_x + \frac{\partial I}{\partial y}V_y + \frac{\partial I}{\partial t} = 0. \tag{3}$$

In the following equation, Vx and Vy are optical flow vectors. From this differential equation, it is necessary to introduce the LK algorithm to solve the optical flow vector.

$$Av = b. \tag{4}$$

For the pixel region with a size of 3×3, there are 9 optical flow tracks in total, which can be expressed as follows in the form of a matrix:

$$A = \begin{bmatrix} I_x(q_1) \\ I_x(q_2) \\ \cdots \\ I_x(q_9) \\ I_y(q_1) \\ I_y(q_2) \\ \cdots \\ I_y(q_9) \end{bmatrix} v = \begin{bmatrix} V_x \\ V_y \end{bmatrix}, b = \begin{bmatrix} -I_t(q_1) \\ -I_t(q_2) \\ \cdots \\ -I_t(q_9) \end{bmatrix}. \tag{5}$$

The variables are shown in the following equation:

$$\begin{aligned} A^T Av &= A^T b, \\ v &= \left(A^T A\right)^{-1} A^T b. \end{aligned} \tag{6}$$

The following equation can be solved by using (6):

$$\begin{bmatrix} V_x \\ V_y \end{bmatrix} = \begin{bmatrix} \sum_t I_x(q_t)^2 & \sum_t I_x(q_t) I_y(q_t) \\ \sum_t I_x(q_t) I_y(q_t) & \sum_t I_y(q_t)^2 \end{bmatrix}^{-1} \bullet \begin{bmatrix} -\sum_t I_x(q_t) I_t(q_t) \\ -\sum_t I_y(q_t) I_t(q_t) \end{bmatrix}. \tag{7}$$

3.3. Background Subtraction. After learning the parameters, the image of the current frame is compared with the background model. Any area with a large difference is considered a foreground object. In recent years, researchers have proposed the Gaussian mixture model, ViBe model, kernel density estimation, and other methods for the study of background reduction, among which the Gaussian mixture model is widely used because it can well model complex backgrounds and has strong adaptability.

A Gaussian mixture model is a probabilistic model in which all data points are generated by combining a finite number of Gaussian distributions with unknown parameters.

Therefore, the Gaussian mixture model-based method can deal with multi-model background distribution problems well. Gaussian mixture model usually describes the video image sequence as a probability distribution function of pixels. The specific process of the Gaussian mixture model is as follows:

3.3.1. Model Building. Here, it is assumed that the value of a pixel point at time T is X_t, then the probability of occurrence of X_t can be obtained by the following equation:

$$P(X_t) = \sum_{i=1}^{k} \omega_{i,t} \bullet \eta\left(X_t, \mu_{i,t}, \sigma_{i,t}\right). \tag{8}$$

3.3.2. Model Update. Assume that the value of a pixel point in the new input frame image is X_t and determine whether the pixel matches the K Gaussian distributions established through the following equation:

$$\left|X_t - \mu_{i,t-1}\right| \le 2.5\sigma_{i,t-1}. \tag{9}$$

If one of K Gaussian distributions satisfies the condition of (9), the pixel point is judged to match it, and the mean, variance, and weight of the Gaussian distribution are updated, as shown in the following equation:

$$\begin{aligned} \omega_{i,t} &= (1 - \alpha)\omega_{i,t} + \alpha, \\ \mu_{i,t} &= (1 - \beta)\mu_{i,t-1} + \beta X_{i,t}, \\ \sigma_{i,t} &= (1 - \beta)\sigma_{i,t-1} + \beta\left(X_{i,t} - \mu_{i,t-1}\right)^T\left(X_{i,t} - \mu_{i,t}\right). \end{aligned} \tag{10}$$

Figure 3: Ballet to dance video database part of the picture display.

3.3.3. Foreground Detection. At the end of background model training, the K Gaussian distributions are arranged according to the size of the queue I and T, and the first B distributions with high priority are selected. Then the following equation is used to generate the background:

$$B = \arg\min\left(\sum_{k=1}^{b} \omega_k > T\right). \tag{11}$$

3.4. Building an Audio Dictionary. The bag-of-words model (BOW model) was first applied in the fields of text information processing and information retrieval and achieved good results. The main idea of this model is that given a text, the text is regarded as a random combination of words, independent of each other, without considering the order of words and sentences, grammar, and other factors in the text. This is done by counting the frequency of each word in the text and using the word frequency histogram to represent the text. In recent years, this method is widely used in motion recognition, image classification, and other research fields due to its simplicity and efficiency. Li et al., from Stanford University, applied this model in the field of image processing. In 2005, Dollar et al., for the first time, applied the word bag model in the study of action recognition.

In this paper, based on the idea of the word bag model, we use the extracted audio signature features to construct an audio dictionary. The process of establishing the voice frequency word bag model is as follows:

(1) Audio feature extraction. This process mainly extracts the required audio signature features from the extracted audio files.

(2) Build an audio dictionary. Firstly, the audio signature features extracted from the training set are clustered by the clustering algorithm, and the generated clustering center is the audio words. Then, the obtained audio words are combined as the audio dictionary.

(3) Quantification of features according to the audio dictionary. The Euclidean distance of each audio signature feature and each audio word in the audio dictionary is calculated, and the minimum distance is selected to complete the classification of audio signature features.

(4) Use of frequency histogram of audio words to represent each audio stream. The audio frequency range corresponds to the lower and top limits of human hearing, which is around 20 to 20,000 Hz.

4. Analysis of Experimental Results and Data

4.1. Data Set. The dance movement recognition algorithm based on multi-feature fusion will verify the validity of the algorithm on ballet to dance video database. Ballet to the

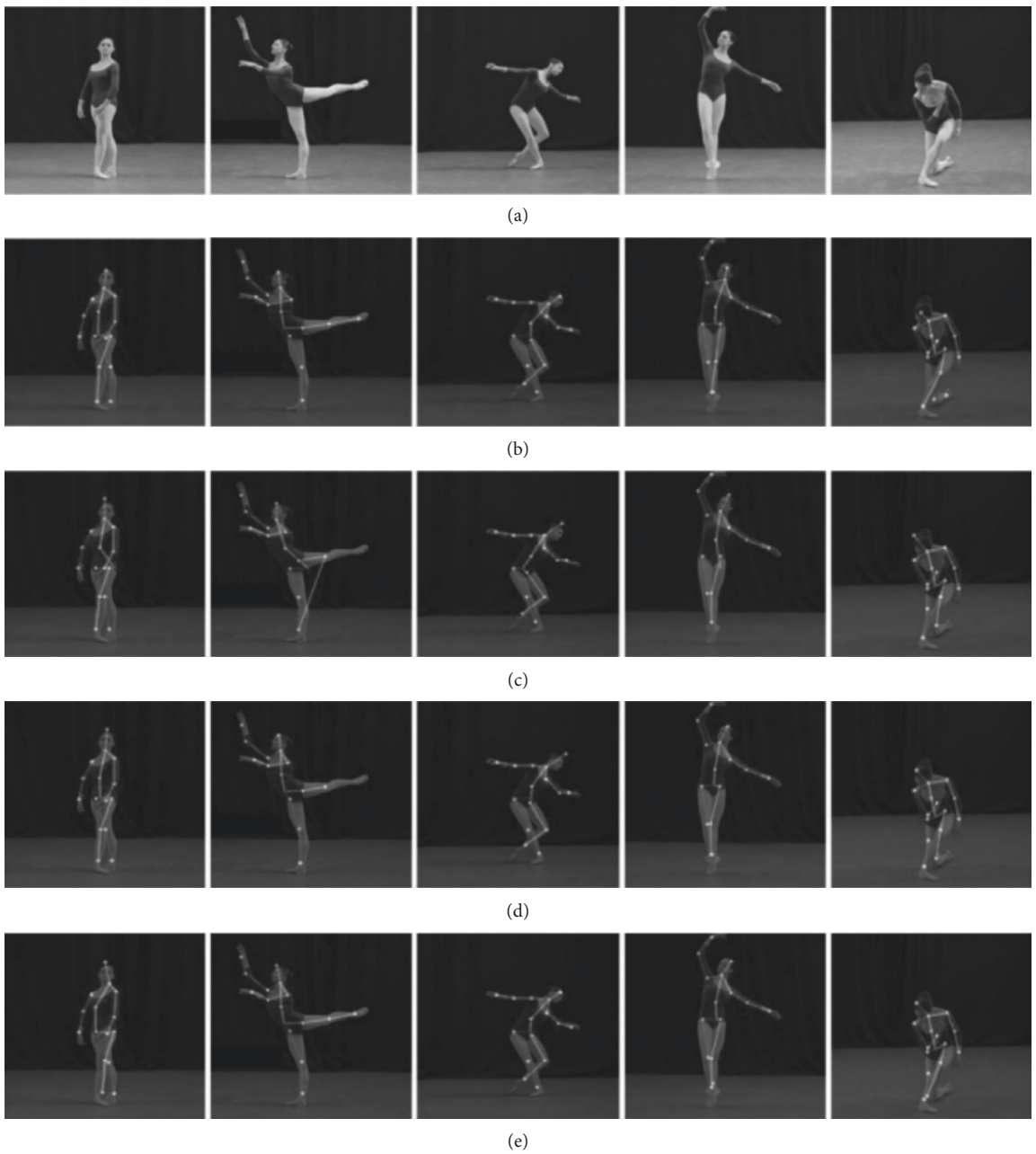

Figure 4: The effect of the model.

database is decomposed from 250 dance movement videos to obtain some key-frame pictures of ballet dance movements. Figure 3 select some pictures as displayed.

4.2. Evaluation Indexes of the Model. This paper uses Percentage of Correct Key-points (PCK) to evaluate each key opinion. The spirit of PCK is that by setting the threshold value, the standardized distance between the projected value and the real value of human key points is less than the proportion of the threshold value, and the lower the PCK value, the higher the accuracy of the algorithm. In this paper, AP, FLOP, and Param Size are used to evaluate the overall performance of the new algorithm. The fraction of OKS scores of all images larger than the set threshold is calculated using AP to evaluate the algorithm's correctness.

4.3. Evaluation of Dance Database. The dance movement recognition algorithm based on multi-feature fusion will be tested on the ballet to dance video database, and the test results are shown in Figure 4. However, for the dance movements in column 2 in Figure 4, because the skeletal scale of dancers' postures varies greatly, the algorithm based on hourglass considers the left foot as hidden by the right foot and directly predicts the left and right ankles at key points of the right ankle. The FPN-based approach, on

Table 1: Accuracy analysis of each key point of the model.

Methods	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Average
Hourglass	95.1	88.3	82.7	79.8	84.1	87.6	77.8	85.6
FPN	95.8	89	82.5	80.2	83.5	86.8	77.4	85.9
Our method	96.2	89.7	83.2	81.6	84.7	88	78.5	86.3

the other hand, is even less optimal, as it explicitly ignores the critical point of the left ankle. The algorithm in this paper has richer semantic features after multi-feature fusion and successfully predicts the left ankle. The dance action in column 5 of Figure 4 is more complex, with dancers' limbs contracting, huge variations in skeleton scale, and varying degrees of occlusion at important places. Hourglass- and FPN-based algorithms misidentify the left leg as the right and the right leg as the left because the dancer's left and right legs are crossed. Although the algorithm in this paper failed to predict the left ankle, it did predict the right leg, and there was no misidentification of important regions of the knee due to the crossing of the left and right legs.

This paper uses Practice Conversion Kit (PCK) index to evaluate each key point of the new algorithm, and the evaluation results are shown in Table 1. As can be seen from Table 1, in the head prediction, the PCK values of the three algorithms are all above 95%. The lowest PCK value of the hourglass algorithm is 95.1%, and the highest PCK value of this algorithm is 96.2%. This shows that the method also improves the regression of simple key points. The prediction in shoulder key points also shows that the algorithm improves the simple key points, and the PCK value of shoulder key points in this method increases by 1.4% compared with the algorithm based on hourglass. However, for wrist and ankle key points, the PCK value of the hourglass-based algorithm is 79.8% and 77.8%, respectively. The PCK value of the FPN-based algorithm is slightly improved for the wrist (80.2%) but slightly decreased for the ankle (77.4%). In this algorithm, a multi-feature fusion module is added, which not only deepens feature extraction but also integrates multi-channel features, which enriches feature extraction semantics.

Compared with the hourglass method, the algorithm improves 1.8% and 0.7%, proving that the new algorithm can effectively improve the prediction of complex key points. But, on the other hand, the ankle key points only increased by 0.7%, indicating that the detection of hidden key points obscured by the human body in 2D images is extremely difficult.

5. Conclusion

Motion recognition is a very challenging subject in the field of computer vision, and it is also one of the research hotspots in the field of computer vision. The purpose of the research based on motion recognition is to use image processing and classification recognition technology to analyze video data to recognize human movements, which has high research value. Therefore, the research on motion recognition is a very popular research direction in recent years, which has been widely applied in many fields such as intelligent monitoring, human-computer interaction, virtual reality, and so on. This paper mainly studies the selection and representation of features and multi-feature fusion methods in dance movement recognition. The main work of this paper is summarized as follows: (1) firstly, the classical literature on the representation and recognition methods of human motion features in videos is studied, in which the direction gradient histogram features, optical flow direction histogram features, word bag model, and audio features involved in dance videos are studied in detail. Secondly, the latest literature in the field of motion recognition is studied in depth to understand the latest motion recognition research trends, and the motion recognition methods are classified and sorted out. (2) Aiming at the research and analysis of the characteristics of dance movements at the same time, this paper proposes an effective feature extraction method, which divides the video of dance movements into equal sections. Then, the edge features of all video images in each segment were added to one image, and the directional gradient histogram features were extracted. Finally, a set of directional gradient histogram feature vectors were used to characterize the appearance and shape features of dance movements in the video. (3) This paper does research on dance movement recognition. This paper extracts the feature of direction gradient histogram, optical flow histogram, and audio frequency and uses the method of multi-feature fusion for dance movement recognition. At the same time, considering the problem of heterogeneous feature fusion, this paper uses multi-kernel learning method to organically integrate three kinds of features for dance movement recognition.

Data Availability

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

Conflicts of Interest

The author declares that there are no conflicts of interest.

References

[1] J. Myers, S. Lephart, Y.-S. Tsai, T. Sell, J. Smoliga, and J. Jolly, "The role of upper torso and pelvis rotation in driving performance during the golf swing," *Journal of Sports Sciences*, vol. 26, no. 2, pp. 181–188, 2008.

[2] A. Covaci, C.-C. Postelnicu, A. N. Panfir, and D. Talaba, "A virtual reality simulator for basketball free-Throw Skills development," in *Proceedings of the Technological Innovation for Value Creation*, pp. 105–112, 2012.

[3] X. Wei, R. Liu, and Q. Zhang, "Review of the techniques for motion capture data processing[J]," *Computer Aided Drafting, Design and Manufacturing*, vol. 22, no. 1, pp. 1–11, 2012.
[4] J. Jun Xin, C. Chia-Wen Lin, and M. Ming-Ting Sun, "Digital video Transcoding," *Proceedings of the IEEE*, vol. 93, no. 1, pp. 84–97, 2005.
[5] S. Marius-Calin, P. N. Ralf, B. Ronan, and F. Pascal, "Local and global skeleton fitting techniques for optical motion capture [J]," *LNCS*, vol. 26, no. 11, pp. 26–40, 1998.
[6] Q. Dang, J. Yin, B. Wang, and W. Zheng, "Deep learning based 2D human pose estimation: a survey," *Tsinghua Science and Technology*, vol. 24, no. 6, pp. 663–676, 2019.
[7] T. Pfister, J. Charles, and A. Zisserman, "Flowing ConvNets for human pose estimation in videos[J]," in *in Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile*, pp. 1913–1921, December 2015.
[8] S. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines[C]," in *Proceedings of the computer Vision and Pattern Recognition*, pp. 4724–4732, Chicago, 2016.
[9] A. Newell, K. Yang, and J. Deng, "Stacked Hourglass networks for human pose estimation," in *Proceedings of the Amsterdam: European Conference on Computer Vision*, pp. 483–499, Paris, 2016.
[10] Z. Cao, T. Simon, S. Wei, and Y. Sheikh, "Realtime Multi-Person 2D pose estimation using part affinity fields[C]," in *Proceedings of the Computer Vision and Pattern Recognition*, pp. 3210–3222, Beijing, 2016.
[11] Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu, and J. Sun, "Cascaded pyramid network for Multi-Person pose estimation [C]," in *Proceedings of the Computer Vision and Pattern Recognition*, pp. 356–363, Shanghai, 2017.
[12] W. Yang, S. Li, W. Ouyang, H. Li, and X. Wang, "Learning feature pyramids for human pose estimation [C]," in *Proceedings of the International Conference on Computer Vision*, pp. 1290–1299, Guangzhou, 2017.
[13] A. F. Bobick and J. W. Davis, "The recognition of human movement using temporal templates," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 3, pp. 257–267, 2001.
[14] Y. Hu, L. Cao, F. Lv, S. Yan, Y. Gong, and T. S. Huang, "Action detection in complex scenes with spatial and temporal ambiguities[C]," in *Proceedings of the International Conference on Computer Vision*, pp. 128–135, IEEE, 2009.
[15] H. Qian, Y. Mao, W. Xiang, and Z. Wang, "Recognition of human activities using SVM multi-class classifier," *Pattern Recognition Letters*, vol. 31, no. 2, pp. 100–111, 2010.