Hindawi
Scientific Programming
Volume 2022, Article ID 6724892, 9 pages
https://doi.org/10.1155/2022/6724892

Research Article

High-Dynamic Dance Motion Recognition Method Based on Video Visual Analysis

Wanshu Luo and Bin Ning

School of Dancing, Shandong Youth University of Political Science, Jinan, Shandong 250103, China

Correspondence should be addressed to Wanshu Luo; 140123@sdyu.edu.cn

Received 9 November 2021; Revised 15 December 2021; Accepted 16 December 2021; Published 17 January 2022

Academic Editor: Tongguang Ni

In the field of computer vision, high-dynamic dance motion recognition is a difficult problem to solve. Its goal is to recognize human motion by analyzing video data using image processing and classification recognition technology. Video multifeature fusion has sparked a surge in research in a variety of fields. Several pixel points that can be distinguished and displayed in several adjacent images that can reflect their characteristics are referred to as multifeature fusion. It is responsible for a significant portion of the similarity results between the two video segments. Motion recognition relies heavily on video multifeature fusion, which has a direct impact on the robustness and accuracy of recognition results. The directional gradient histogram features, optical flow direction histogram features, and audio features extracted from dance video are used to characterize dance movements after all of the characteristics of dance movements have been considered. This paper focuses on the high-dynamic dance action recognition method based on video multifeature fusion, which aims to combine high-dynamic dance action recognition and video multifeature fusion.

1. Introduction

Video information has become widely used in many fields as a result of the development of computer vision and video image processing technology [1]. Computer vision is widely used in medical, transportation, and other fields as an auxiliary means of human vision and an important part of automation systems [2]. Users hope to retrieve and query specific action clips in high-dynamic dance video as easily and quickly as they retrieve and query text information and then obtain specific action clips of interest for playback and browsing [3]. Human posture estimation technology's application field of dance movement recognition is important. Dance movement recognition technology can assist dancers in identifying and correcting incorrect postures, as well as contributing to intelligent dance auxiliary training. Action recognition can help with game design in addition to dance learning and teaching [4]. Firstly, the actions in the dance image video are recognized, and then the action model is further established according to the recognition results to generate the character actions in the game, so as to greatly enhance the user experience effect. The research on human motion recognition in China started relatively late, but it developed rapidly. At present, many units are engaged in intelligent video analysis and research and have made some breakthroughs in related fields. Well known is the research on the key technology of intelligent surveillance video analysis carried out by the Institute of Automation, Chinese Academy of Sciences, which automatically analyzes the video captured from the camera, locates, tracks, and identifies the moving targets in it, realizes the management of daily surveillance, and makes timely response to abnormal situations. In the field of human motion recognition, some significant research achievements have been made after years of development. The research has also broadened from simple action analysis and recognition in a simple background to multiperson complex actions in a complex background. In the current research field of computer vision, motion recognition is a very difficult subject [5, 6]. Its goal is to recognize human motion [7] by analyzing video data [8] using image processing and classification recognition technology [9, 10].

In the research of motion recognition, video multifeature fusion is usually the first step [11]. After fully considering the characteristics of dance movements, the directional gradient histogram features, optical flow direction histogram features, and audio features extracted from dance video are used to characterize dance movements. Key frames are defined as some representative image frames with less redundancy in video. The directional gradient histogram feature is used to describe the local appearance and shape features of dance movements, and the optical flow direction histogram feature is used to describe the motion information of dance movements. In addition, the research on dance action recognition should also consider the impact of music on dance. Dance performers perform dance with the accompaniment of music, and the style of music is related to the type of dance. The key frame extraction has been made difficult by the variability of dance movements and the presence of too many redundant movements. This paper will calculate the optical flow of the image sequence of dance action video after framing in order to extract a set of key frames with less redundancy and can summarize the video content. This method can match large-distance actions and estimate optical flow for small objects. The accuracy of action recognition results and the robustness of action recognition methods are both influenced by the extracted features. Video multifeature fusion has sparked a surge in research in a variety of fields. Several pixel points that can be distinguished and displayed in several adjacent images that can reflect their characteristics are referred to as multifeature fusion [12, 13]. The directional gradient histogram features, optical flow direction histogram features, and audio features extracted from dance video are used to characterize dance movements after all of the characteristics of dance movements have been considered.

This fusion method produces new feature vectors with excessively high dimensions and redundant information. Although this method can improve classification effect to a degree, increasing the feature vector dimension increases computational cost, and the new features formed by fusion contain many features, resulting in the same weight of each feature [14, 15]. Such features with a poor classification effect are given the same weight as those with a good classification effect, affecting classification accuracy to some extent [16]. Video multifeature fusion plays a very important role in motion recognition, which directly affects the robustness and accuracy of recognition results [17, 18]. Video multifeature fusion is based on the trajectory formed by the connection of multiple feature points, and then extracting the trajectory based features is a good choice. It not only combines the global spatiotemporal volume features with local feature points, but also includes the temporal and partial spatial relationship between feature points, so that the extracted features can cover more abundant information of moving objects. The video multifeature fusion method can assign weights to each feature according to their contribution to classification through the decision-making mechanism, so it can let a variety of features give full play to their respective advantages, so as to improve the classification accuracy. Video multifeature fusion is a widely used fusion method in multifeature fusion.

2. Related Work

According to [19], some authoritative journals and important academic conferences in relevant fields around the world, such as IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) and International Journal of Computer Vision (IJCV), include motion analysis and recognition in video as one of the main research contents. A major visual monitoring project was established by the US Defense Advanced Research Projects Agency, which included many well-known universities. Its main research focus is on the intelligent analysis and comprehension of surveillance videos in a variety of settings, including battlefields and everyday settings. This study can achieve human motion region segmentation, multiperson tracking, and basic human motion analysis and recognition. Literature [20] shows that, by tracking the trajectory formed by several points on the human body, the recognition of simple actions such as walking and running is realized, which opens the door of action recognition research. Literature [21] points out that, in the process of learning and understanding content-based video retrieval, action video retrieval has become the mainstream and action video generally involves running, jumping, swimming, weightlifting, etc. Literature [22] shows that there is too little research on dance video because dance is a continuous and changeable limb movement, and different types of dance will have limb movements with different intensity changes. Literature [23] mentions that, through the big data analysis method, the main idea of content-based video retrieval technology is, given an image, similarity matching carried out according to the content in the image to find videos with the same or similar content. Literature [24] pointed out that image retrieval is a well-known video retrieval technology earlier. It distinguishes videos by manually marking some text descriptions or numbers, and when searching videos, it uses marked labels to search. Literature [25] reveals that, according to research, a large number of domestic and international scientific research institutions and related scholars are dedicated to the study of motion recognition based on video and have made significant contributions to the field's development. Domestic motion recognition research began relatively late, according to [26], using the big data analysis method. Many domestic universities and scientific research institutions have conducted motion recognition research as the application of motion recognition becomes more widespread. According to [27], features with a poor classification effect have the same weight as those with a good classification effect, which will affect classification accuracy to some extent.

3. Theory and Technology Related to High-Dynamic Dance Movement Recognition

3.1. Dance Movement Recognition Theory. Dance motion recognition has advanced to the level of motion recognition, as opposed to low-level motion recognition such as gesture recognition and simple body motion recognition. The term “flexible force sensor” refers to the material in which it is made. The force between irregular contact surfaces can be measured with a flexible force sensor. The flexible force sensor is an array of pressure sensitive points that can be seen from the outside. The first step in high-dynamic dance movement recognition was to use computers to perceive people’s movements, and then it progressed to recognizing simple movements like walking, running, and jumping. High-dynamic dance movement recognition has evolved into video movement recognition, which recognizes human movements in videos as a series of simple movements. There will be multiple types of actions in each video frame sequence, and each type of action will contain multiple simple actions. Every simple action is made up of a variety of human body positions. The connection of transition actions between actions allows all types of actions to be coherent and smooth. The single step gait cycle’s total pressure curve looks like a saddle curve during normal walking. The curve has three distinct points: xmax1 represents total pressure when the heel touches the ground, xmin represents minimum total plantar pressure when the entire foot touches the ground, and xmax2 represents total plantar pressure when the front foot steps off the ground. Figure 1 depicts this.

Therefore, it is difficult to obtain high recognition accuracy when a simple limb localization algorithm is used in dance movement recognition. It not only combines the global spatiotemporal volume features with local feature points, but also includes the temporal and partial spatial relationship between feature points, so that the extracted features can cover more abundant information of moving objects. In the stage of choreography, different forms of choreography should be carried out according to the characteristics of music. Similarly, in the soundtrack, it is necessary to match appropriate accompaniment music according to the rhythm and style of dance movements. The speed of dance movements is related to music. This paper will extract the accompaniment music from the video, read the accompaniment music in wav format, and then extract the envelope feature and energy feature of the music to prepare for the subsequent feature fusion, which is very helpful to the event detection of music. Dance movements usually have complex and large-scale changes, and recognizing dance movements requires a deep learning model [28, 29] to grasp the needs of each scale information when extracting features. Figure 2 is the result of envelope feature extraction for the accompaniment music of the dance video Theodora _ Africa _ wide. It shows the whole appearance characteristics of the audio signal.

The plantar pressure data acquisition platform of the flexible force sensor designed in this paper dynamically acquires the plantar pressure information of the tester by using the characteristics of the flexible force sensor in real time. However, the high-dynamic dance moves are continuous and smooth, and it is difficult to accurately find the segmentation points, which is the main reason for the unsatisfactory segmentation of video motion sequences. Furthermore, the shortest data cannot contain the most human motion information simply by extracting key frames for motion sequence segmentation. A complete dance performance is made up of several groups of movements, each of which is made up of several simple movements such as flowers and pressing steps, and each simple movement is made up of several basic posture sequences such as raising hands and kicking legs.

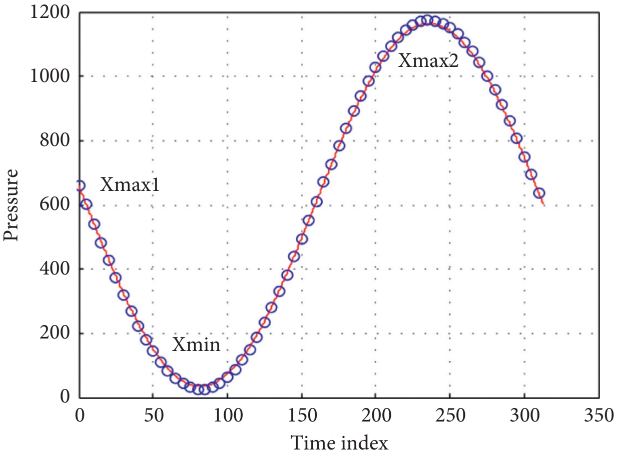

Figure 1: Characteristic points of total plantar pressure curve.

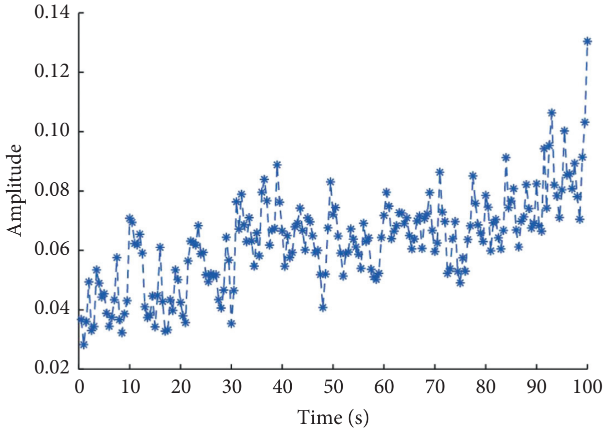

Figure 2: Dance video Theodora_ Afraid_ Envelope extracted from accompaniment music wide.

3.2. Dance Movement Recognition Technology. The initial stage of high-dynamic dance motion recognition is to extract some features of images or video sequences and realize motion recognition through feature matching. Later, researchers found that the combination of low-level features and high-dynamic dance pose features can improve the accuracy of motion recognition. There are roughly two ways to obtain high-dynamic dance posture information. One is to accurately obtain the information of each joint coordinate, human skeleton, motion trajectory, and so on through motion capture equipment. Another method is to obtain the approximate joint positions and skeletons of the human body in images or videos by the method of human posture estimation. Dance movements usually have complex and large-scale changes, and recognizing dance movements requires a deep learning model to grasp the needs of each scale information when extracting features. According to the actual situation, the change trend of human posture joint position is basically the same, and there are obvious changes at the beginning and end of the action. At frame 600, i.e., after that, the curve changes disorderly, and the change cannot be accurately judged through the curve in the figure. It is necessary to obtain the concise and clear change trend of the whole pose sequence through further calculation. The position change trend of 30 joint points in human posture sequence is shown in Figure 3.

Although the change trend of human posture can be roughly seen in the figure, it is impossible to calculate and obtain the segmentation position, so it is necessary to obtain a curve that is visible and can calculate the segmentation position through further calculation. In this paper, the posture trend is expressed by calculating the average value of 30 joint positions.

$$\text{mean} = \frac{\left(\sum_{j=1}^{19} f_i\right)}{19}. \quad (1)$$

It has been discovered that using the mean value does not result in a significant change in the positions of the 30 joints in the video sequence frames but does result in a stable change trend for the regions with more chaotic changes. Figure 4 depicts the changing trend of the average position of the actor's 30 joint points over time.

Dancers often have obvious preparation time and stop time when performing dance movements. During this period, the dancers' posture basically does not change, and the distance between consecutive frames changes gently. However, most of the adjacent minima to maxima can only represent a simple movement, but not necessarily a complete dance movement. Because a complex dance movement is mostly composed of several simple movements, there are usually several extrema in a dance movement sequence.

Therefore, after finding each extreme value in the curve of continuous frame change trend, this paper uses cubic spline interpolation function to fit the curve. Firstly, the maximum value of continuous frame change data in video is obtained, and the upper envelope of data sequence is fitted by cubic spline function, and then the lower envelope of data sequence is fitted by cubic spline function. Finally, the mean value between the upper envelope and the lower envelope is calculated as the final fitting result of the current data series. The upper envelope, lower envelope, and final results of curve fitting are shown in Figures 5 and 6, respectively.

We can see the obvious segmentation points between various dance movements in a dance video after we get the fitting results. A simple action sequence can be determined by two adjacent minima, with the minima position indicating the segmentation position of the action sequence frame. Because the fitted curve may still have inaccurate segmentation positions, an action sequence must be determined using the dual control method of the difference between adjacent maxima and minima and the number of frames in the action sequence. According to statistics from actual recorded dance situations, the dance action is a rhythmic action with a time limit of 10 seconds in the data collection process and a video frame rate of two eight beats per second, making it impossible to complete a dance action in one second at normal speed.

It is difficult to quickly and accurately retrieve specific action segments in high-dynamic dance video, and only the inherent video information contained in high-dynamic dance video data can be used to retrieve specific action segments temporarily. In order to realize the analysis and management of high-dynamic dance videos, it is necessary to study the design method of retrieval system for specific action segments in videos. People's orientation, the arrangement of limbs, and the relationship between adjacent joints all need to be inferred and identified from global context information, and local information can be accurately located. Motion recognition based on high-dynamic dance is also divided into two categories. One is to directly use the main joint information in human motion posture for similarity calculation and determine the action category by

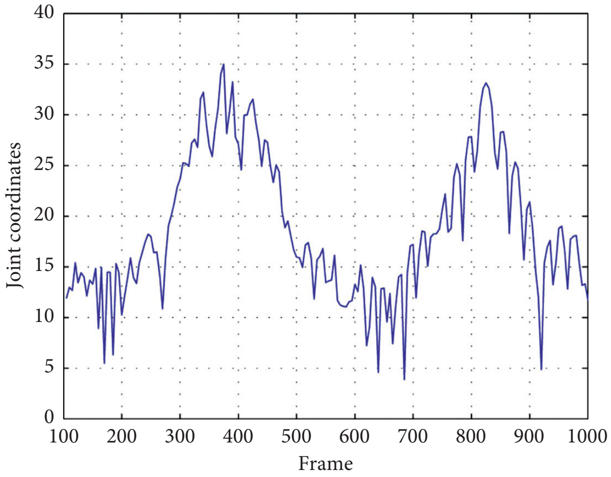

Figure 3: Trend chart of joint changes in continuous pose of actors.

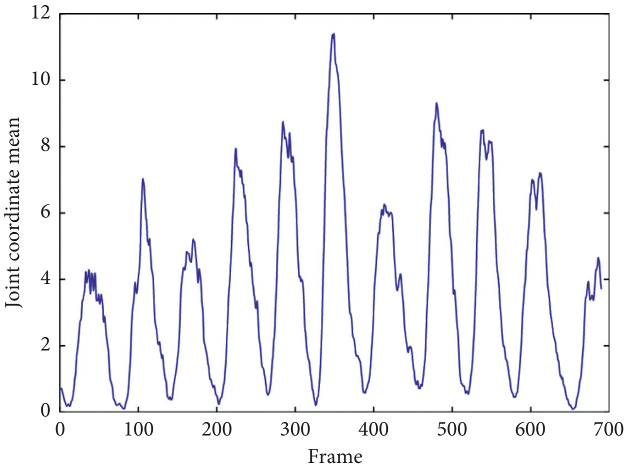

Figure 4: Average change trend of actors' continuous posture joints.

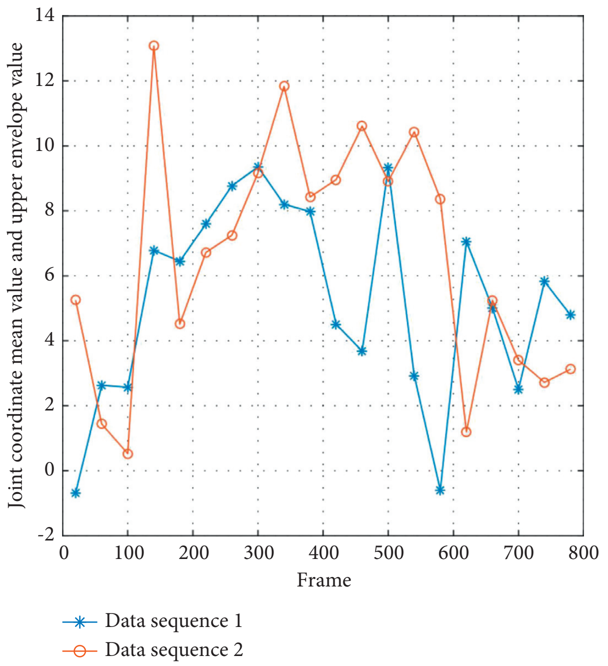

Figure 5: Curve fitting result diagram of envelope.

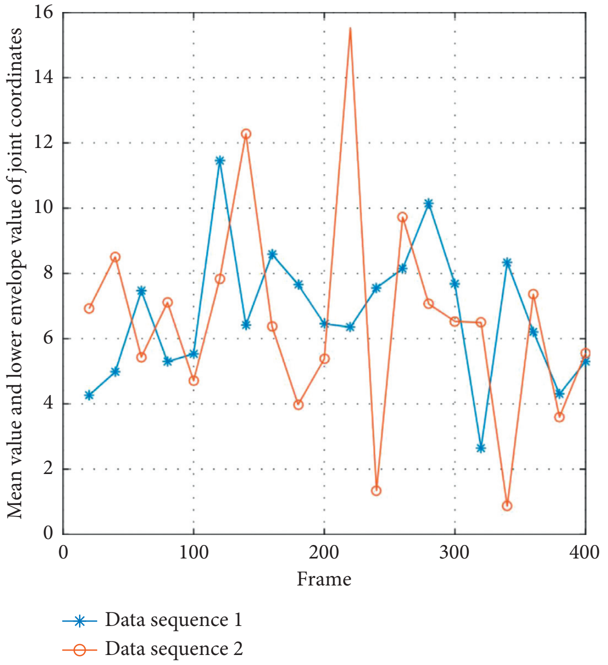

Figure 6: Lower envelope fitting results.

matching with the similarity between known actions, which is mostly used for human motion recognition in images. The other is to segment the human body in the image by using the obtained human posture, obtain the local image with each main joint position as the key, and then extract some features of the local image for recognition.

3.3. Action Recognition Method. At present, motion recognition methods are mainly divided into two categories: single-layer method and hierarchical method. Single-layer based methods usually regard actions as the characteristic category of video and use classifiers to identify actions in video. Image sequences in video are considered to be generated by specific action categories. The hierarchical method mainly identifies high-level actions by identifying simple actions or low-level atomic actions in the video. High-level complex actions can be decomposed into a sequence of subactions, which can continue to be decomposed as high-level actions until they are decomposed into atomic actions. The classification of action recognition methods is shown in Figure 7.

3.3.1. Single-Layer Method. There are two kinds of motion recognition methods based on single layer: spatiotemporal method and sequence model method. The main difference between spatiotemporal method and sequential model method is how to treat the time dimension. Different actions need to fully detect all kinds of actions, and the least crossing is better. When there are multiple types of human actions in the video, manual video editing shall be carried out and then processed according to the video containing single type of actions. Before collecting video data, most researchers choose to plan the type of action to be collected and the timing of various action videos. The sequence model method is usually better than the time and space method because it considers the sequence relationship of actions. In reality, the video cannot match the dataset's style during the research process. In most cases, the video contains a wide range of actions, and the human posture is consistent across all of them, with no discernible posture change interval. The contour feature based on the action energy map is used by the global feature, while the target cell is used by the local feature. Finally, the feature points are classified using the support vector machine's multiclass classification method.

3.3.2. Hierarchical Approach. Hierarchy-based methods usually use single-level or low-level subactions to identify high-level complex actions. A high-level complex action can be decomposed into several subaction sequences, and subactions can be decomposed as high-level actions until they are decomposed into atomic actions. At the same time, if manual video editing is used, it will take a lot of time to face professional dance movements, and nonprofessional researchers cannot guarantee the accuracy of video editing. The hierarchical method is closely related to the single-layer method to a certain extent. For example, the single-layer method can not only be used for low-level or atomic action recognition, but also be extended to the action recognition method of hierarchical model. Hierarchical methods are generally divided into three types: statistics based methods, grammar based methods, and description based methods. In addition, this paper also tested the recognition time within the range of 30 ~ 200 template images and found that the recognition time is directly proportional to the number of template images, ranging from 0.25 s to 1.75 s, as shown in Figure 8. Because the more the template images, the larger

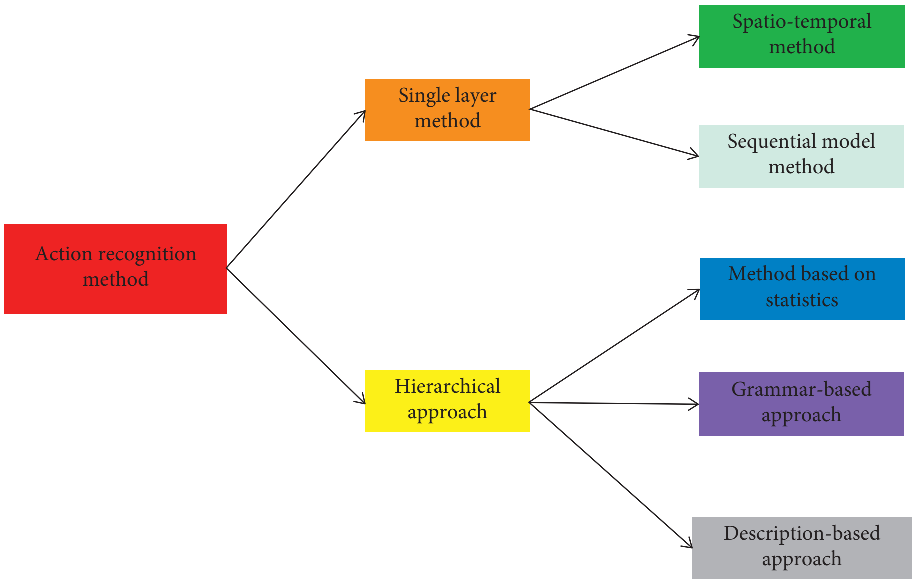

Figure 7: Classification of action recognition methods.

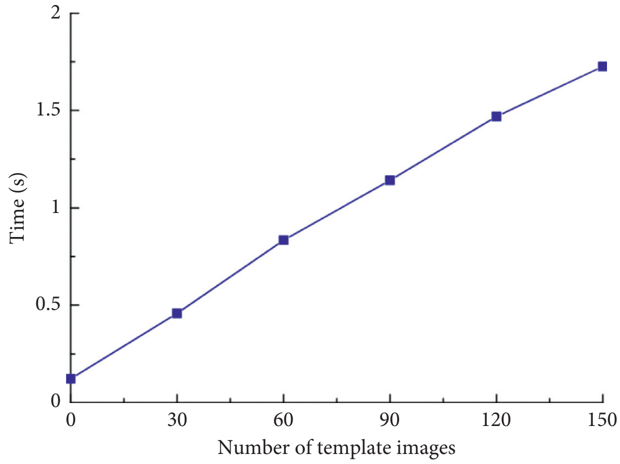

Figure 8: Identification time consumption diagram.

the scale of the template matrix, and the greater the amount of calculation during recognition operation, and the time required for calculating the distance vector between each row of the matrix and the test image is basically the same. Therefore, the recognition time increases linearly with the number of template images.

Build the model. Here, assuming that the value of a pixel at time t is XT, the probability of occurrence of XT can be obtained from

$$P\left(X_t\right) = \sum_{t=1}^{K} \omega_{i,t} \cdot \eta\left(X_t, \mu_{i,t}, \sigma_{i,t}\right), \quad (2)$$

where $\omega_{i,t}$ is the weight of the ith Gaussian distribution at time t, $\eta\left(X_t, \mu_{i,t}, \sigma_{i,t}\right)$ is the corresponding probability density function, $\mu_{i,t}$ is the corresponding mean, and similarly $\sigma_{i,t}$ is the variance. Meanwhile, the specific expression of $\eta\left(X_t, \mu_{i,t}, \sigma_{i,t}\right)$ is shown in

$$\begin{aligned} &\eta\left(X_t, \mu_{i,t}, \sigma_{i,t}\right) \\ &\quad = \frac{1}{\sqrt{2\pi\left|\sigma_{i,t}\right|}} e^{-(1/2)\left(X_t - \mu_{i,t}\right)^T \sigma_{i,t}^{-1} \left(X_t - \mu_{i,t}\right)}. \end{aligned} \quad (3)$$

Firstly, the pixel values of the first frame of the video are assigned to the mean value of K Gaussian distributions; secondly, a larger value is assigned to their variance, and their weights are assigned the same value.

4. Video Multifeature Fusion and Recognition

4.1. Analysis of Related Problems. In the past methods of motion recognition, using a single feature can only describe one aspect of human motion in video but cannot describe human motion effectively. With the continuous evolution of dance types and forms and the increasing number of dance videos, how to browse dance videos quickly and effectively is the main problem. The movement of music and dance video is complex and changeable, and there are many repetitive movements, which brings trouble to the analysis and recognition of dance movements. The main difference between video multifeature fusion methods is the selected features and fusion strategies. Therefore, video multifeature fusion method has become a research direction of motion recognition. Fusion of different video features can describe human movements in the video more comprehensively, thus achieving better recognition effect. At present, video multifeature fusion can be divided into feature level fusion and decision level fusion.

4.1.1. Feature Layer Fusion. Feature layer fusion refers to the combination of various features to form new features; for example, there are two features $F1$ and $F2$, and a new feature vector $F3 = (f1, f2)$ is formed after feature layer fusion.

4.1.2. Integration of Decision-Making Level. The main idea of decision level fusion is to create corresponding classifiers for each feature, respectively, and then fuse the results of each feature classifier according to the selected decision mechanism to get the final classification result. The above two fusion strategies are facing certain problems. In recent years, scholars need to apply multicore learning to the research of multifeature fusion methods.

4.2. Dance Movement Recognition Method Based on Feature Fusion. The directional gradient histogram features, optical flow directional histogram features, and audio features extracted in this paper describe the characteristics of dance movements from the aspects of the appearance and shape of human dance movements in dance videos, the movement of human dance movements, and the assistance of audio features. Video clip retrieval entails locating a video clip that is similar to the query clip in the video and then determining the video clip's location in various ways. The Balletto dance video database will be used to test the effectiveness of the dance motion recognition algorithm based on video multifeature fusion. The feature fusion method has shown to be effective in the field of image classification, and it has since been applied to motion recognition research. Learning by using multiple kernel functions in training is a simple understanding. Multicore learning can effectively fuse various heterogeneous features by linear combination of kernel functions in the process of learning classifiers, so that different features can complement each other to improve recognition accuracy. Because the features used in our research contain heterogeneous features, multicore learning can effectively fuse various heterogeneous features by linear combination of kernel functions in the process of learning classifiers. The first step in high-dynamic dance motion recognition is to extract some features from images or video sequences and then match them to achieve motion recognition. Figure 9 depicts the multicore learning feature fusion process.

In practical application, according to the content and structure characteristics of the video, comprehensively measure the above points and allocate the weight. Based on this, we can find the video clip with high similarity to the query clip, as shown in Figure 10.

Therefore, it is assumed here that there are p dance movements $x1$, $X2$,..., XP and categories $Y1$, $Y2$,..., YP in the dance dataset. At the same time, the G kernel functions corresponding to the hog feature are defined as kg (Xi, XJ), $g = 1, 2, \ldots, G$, the f kernel functions corresponding to the HOF feature are defined as KF (Xi, XJ), $f = 1, 2, ..., F$, and the M kernel functions corresponding to the audio signature feature are km (Xi, XJ), $M = 1, 2, ..., M$. In this paper, the linear combination of kernel functions integrating the above three features can be expressed by the following formula:

$$\begin{gathered} k\left(x_i, x_j\right) = \sum_{g=1}^{G} \beta_g k_g\left(x_i, x_j\right) + \sum_{f=1}^{F} \beta_f k_f\left(x_i, x_j\right) + \sum_{m=1}^{M} \beta_m k_m\left(x_i, x_j\right), \\ \beta_g \geq 0 \forall g,\ \beta_f \geq 0 \forall f,\ \beta_m \geq \forall m, \quad \sum_{g=1}^{G} \beta_g + \sum_{f=1}^{F} \beta_f + \sum_{m=1}^{M} \beta_m = 1. \end{gathered} \tag{4}$$

β_g, β_f, and β_m are the weights of the corresponding kernel functions, respectively.

Arbitrarily select the two action sequences X and y after segmentation, and the lengths are L1 and L2, respectively. X and y are, respectively, expressed as

$$\begin{aligned} X &= \left\{x_1, x_2, \ldots, x_m, \ldots, x_{l1}\right\}, \\ Y &= \left\{y_1, y_2, \ldots, y_n, \ldots, y_{l2}\right\}. \end{aligned} \tag{5}$$

The mth frame of x is represented by xm, and nth frame of y is represented by yn. If $l1 = l2$, the interval between two action sequences will be calculated directly. If the two action sequences are not equal in length, the distance will be calculated after they are aligned by dynamic programming method.

In support vector machine based on multicore learning, the task of multicore learning model training stage is to learn to solve the weight β of each kernel function and the parameters α and B of support vector machine classifier. Based on the idea of simple MKL algorithm proposed by Palaiahnakote et al. introduced in the previous section, the objective function of the algorithm in this paper is defined as follows:

$$\begin{gathered} \begin{cases} \dfrac{\min}{\beta_g, \beta_f, \beta_m, a, b} & J = \dfrac{1}{2} \displaystyle\sum_{g=1}^{G} \beta_g \alpha^T K_g \alpha + \dfrac{1}{2} \sum_{f=1}^{F} \beta_f \alpha^T K_f \alpha + \dfrac{1}{2} \sum_{m=1}^{M} \beta_m \alpha^T K_m \alpha + C \sum_i \xi_i \\ \text{s.t.} & y_i \left[\displaystyle\sum_{G=1}^{G} \beta_g k_g\left(x_i\right) + \sum_{f=1}^{F} \beta_f K_f\left(x_i\right) + \sum_{m=1}^{M} \beta_m K_m\left(x_i\right) \right] \alpha + y_i b \geq 1 - \xi_i \forall i, \end{cases} \\ \xi_i \geq 0 \forall i, \quad \sum_{g=1}^{G} \beta_g + \sum_{f=1}^{F} \beta_f + \sum_{m=1}^{M} \beta_m = 1, \end{gathered} \tag{6}$$

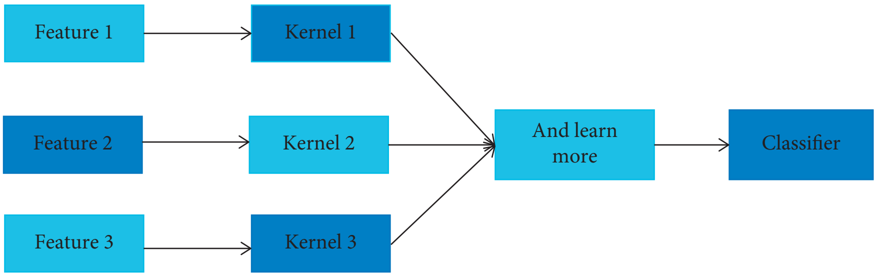

Figure 9: Schematic diagram of multicore learning feature fusion process.

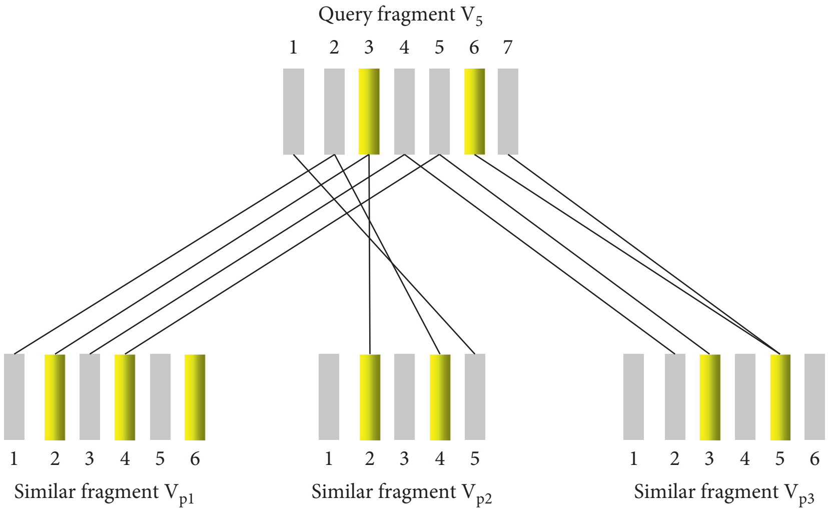
Query fragment V_5
1 2 3 4 5 6 7
1 2 3 4 5 6
Similar fragment V_{p1}
1 2 3 4 5
Similar fragment V_{p2}
1 2 3 4 5 6
Similar fragment V_{p3}

Figure 10: An example of video segment similarity matching.

where kg (*Xi*) = [kg (*Xi*, *X*1), ..., kg (*Xi*, XP)], KF (*Xi*) = [KF (*Xi*, X1), ..., KF (*Xi*, XP)], and km (*Xi*) = [km (*Xi*, X1), ... According to the idea of simple MKL algorithm, the gradient descent algorithm is used to minimize the objective function and learn to solve the optimal parameters. The specific process is that, in each iteration, the classifier parameters α and B are calculated by giving the kernel function weight β; then, given α and B, a new kernel function weight β is calculated. Therefore, the classification function based on multicore learning support vector machine is as follows:

$$\begin{aligned} y &= F(x) \\ &= \left[\sum_{g=1}^{G} \beta_g K_g(x) + \sum_{f=1}^{F} \beta_f K_f(x) + \sum_{m=1}^{M} \beta_m K_m(x)\right]\alpha + b. \end{aligned} \tag{7}$$

In practical application, according to the content and structure characteristics of the video, comprehensively measure the above points and allocate the weight. Based on this, we can find the video clip with high similarity to the query clip.

5. Conclusions

The large amount of video data on the network is due to the wide range and fast speed of video transmission. The retrieval system improves precision and recall and has good practical performance, thanks to the density distribution, compactness, dispersion, and similarity between specific action segments in high-dynamic dance video. The diversity and repeatability of video become a difficult problem in the field of video retrieval when using content-based video retrieval. The related video retrieval research in this paper is focused on changeable dance videos. The single-layer method and hierarchical method are introduced first in terms of action recognition methods. Different actions must be fully detected in order to fully detect all types of actions, and it is preferable to cross at least once. When a video contains a variety of human actions, manually edit the video and then process it as if it contained only one type of action. Subaction sequences can be decomposed as high-level actions, and subactions can be decomposed as high-level actions until they are decomposed into atomic actions. As a result, motion recognition research has become increasingly popular in recent years, and it is now widely used in a variety of fields such as intelligent monitoring, human-computer interaction, and virtual reality. In a nutshell, human motion recognition research, particularly complex movement research, is still in its infancy, and many aspects such as methods and real time need to be improved. Despite the fact that an increasing number of researchers are working on this project, ongoing efforts are required to achieve the goal of advanced intelligent recognition and detection.

Data Availability

The data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] X. Wang, G. Yan, H. Wang et al., “Semantic annotation for complex video street views based on 2d–3d multi-feature fusion and aggregated boosting decision forests,” *Pattern Recognition*, vol. 62, pp. 189–201, 2017.

[2] W. Yang, F. Tong, X. Gao, C. Zhang, G. Chen, and Z. Xiao, “Remote sensing image compression evaluation method based on neural network prediction and fusion quality fidelity,” *Mobile Information Systems*, vol. 2021, no. 4, 9 pages, Article ID 9948811, 2021.

[3] H. Song, W. Xu, D. Liu, L. Bo, L. Qingshan, and N. M. Dimitris, “Multi-stage feature fusion network for video super-resolution,” *IEEE Transactions on Image Processing*, vol. 30, pp. 2923–2934, 2021.

[4] M. Güder and N. K. Iekli, “Multi-modal video event recognition based on association rules and decision fusion,” *Multimedia Systems*, vol. 24, no. 1, pp. 55–72, 2018.

[5] Y. Cai, J. Liu, Y. Guo, S. Hu, and S. Lang, “Video anomaly detection with multi-scale feature and temporal information fusion,” *Neurocomputing*, vol. 423, no. 5, pp. 264–273, 2021.

[6] B. Yin, M. Lv, and Y. Wei, “Multi-feature fusion for thermal face recognition,” *Infrared Physics & Technology*, vol. 77, pp. 366–374, 2016.

[7] X. Ning, F. Nan, S. Xu, L. Yu, and L. Zhang, “Multi-view frontal face image generation: a survey,” *Concurrency and Computation: Practice and Experience*, Article ID e6147, 2020.

[8] P. S. Lamba, D. Virmani, and O. Castillo, “Multimodal human eye blink recognition method using feature level fusion for exigency detection,” *Soft Computing*, vol. 24, no. 5, Article ID 16829, 2020.

[9] C. Wang, J. Zhou, and B. Xiao, “Uncertainty estimation for stereo matching based on evidential deep learning,” *Pattern Recognition*, vol. 124, Article ID 108498, 2021.

[10] Z. Huang, Y. Liu, and C. Zhan, “A novel group recommendation model with two-stage deep learning,” *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2021.

[11] M. Zhao, A. Jha, Q. Liu et al., “Faster mean-shift: GPU-accelerated clustering for cosine embedding-based cell segmentation and tracking,” *Medical Image Analysis*, vol. 71, Article ID 102048, 2021.

[12] H. Tao and L. Xiaobo, “Automatic smoky vehicle detection from traffic surveillance video based on vehicle rear detection and multi-feature fusion,” *IET Intelligent Transport Systems*, vol. 32, no. 2, 2019.

[13] M. Gao, W. Cai, and R. Liu, “AGTH-Net: attention-based graph convolution-guided third-order hourglass network for sports video classification,” *Journal of Healthcare Engineering*, vol. 2021, Article ID 8517161, 10 pages, 2021.

[14] R. Liu, W. Cai, G. Li, X. Ning, and Y. Jiang, “Hybrid dilated convolution guided feature filtering and enhancement strategy for hyperspectral image classification,” *IEEE Geoscience and Remote Sensing Letters*, vol. 5, 2021.

[15] L. Hua, J. Xue, and L. Zhou, “An automatic MR brain image segmentation method using a multitask quadratic regularized clustering algorithm,” *International Journal of Health Systems and Translational Medicine*, vol. 1, no. 2, pp. 44–58, 2021.

[16] M. Luo, X. Chang, L. Nie, Y. Yang, A. G. Hauptmann, and Q. Zheng, “An adaptive semisupervised feature analysis for video semantic recognition,” *IEEE Transactions on Cybernetics*, vol. 48, pp. 648–660, 2017.

[17] S. Ying, Y. Weng, B. Luo et al., “Gesture recognition algorithm based on multi-scale feature fusion in RGB-D images,” *IET Image Processing*, vol. 14, pp. 3662–3668, 2020.

[18] X. Jiamin, S. Palaiahnakote, L. Tong, L T. Chew, and U. Seiichi, “A new method for multi-oriented graphics-scene-3D text classification in video,” *Pattern Recognition*, vol. 49, pp. 19–42, 2016.

[19] W. Liyuan, Z. Jing, Y. Jiacheng, and Z. Li, “Porn streamer recognition in live video based on multimodal knowledge distillation,” *Chinese Journal of Electronics*, vol. 30, no. 6, pp. 1096–1102, 2021.

[20] P. Tang, H. Wang, and S. Kwong, “G.-Ms2F: GoogLeNet based multi-stage feature fusion of deep CNN for scene recognition,” *Neurocomputing*, vol. 225, pp. 188–197, 2016.

[21] K. V. V. Kumar, P. V. V. Kishore, and D. Anil Kumar, “Indian classical dance classification with adaboost multiclass classifier on multifeature fusion,” *Mathematical Problems in Engineering*, vol. 2017, Article ID 6204742, 18 pages, 2017.

[22] M. Li, Z. Miao, and W. Xu, “A CRNN-based attention-seq2seq model with fusion feature for automatic labanotation generation,” *Neurocomputing*, vol. 454, no. 23, 2021.

[23] L. Hai, W. Xiang, and Z. B. Wei, “Infrared head pose estimation with multi-scales feature fusion on the IRHP database for human attention recognition - sciencedirect,” *Neurocomputing*, vol. 411, pp. 510–520, 2020.

[24] B. Xiao, J. Zhao, and C. Zhao, “Video text detection based on multi-feature fusion,” *Journal of Intelligent and Fuzzy Systems*, vol. 37, no. 2, pp. 1–12, 2019.

[25] H. Wang, S. K. Nguang, and J. Wen, “Robust video tracking algorithm: a multi-feature fusion approach,” *IET Computer Vision*, vol. 12, no. 5, pp. 640–650, 2018.

[26] X. Zhai, “Dance movement recognition based on feature expression and attribute mining,” *Complexity*, vol. 2021, no. 21, 12 pages, Article ID 9935900, 2021.

[27] Q. Guan, S. Ren, and L. Chen, “A spatial-compositional feature fusion convolutional autoencoder for multivariate geochemical anomaly recognition,” *Computers & Geosciences*, vol. 156, Article ID 104890, 2021.

[28] H. Tang, H. Liu, W. Xiao, and N. Sebe, “Fast and robust dynamic hand gesture recognition via key frames extraction and feature fusion,” *Neurocomputing*, vol. 331, pp. 424–433, 2019.

[29] Y. Fang, X. Zhang, and D. Zhou, “Improve inter-day hand gesture recognition via convolutional neural network-based feature fusion,” *International Journal of Humanoid Robotics*, vol. 18, 2021.