Hindawi
Scientific Programming
Volume 2022, Article ID 9735392, 8 pages
https://doi.org/10.1155/2022/9735392

Research Article

Key Frame Extraction Method of Music and Dance Video Based on Multicore Learning Feature Fusion

Ping Yao

Zhengzhou University of Science and Technology, Zhengzhou, 451150, China

Correspondence should be addressed to Ping Yao; ieuniversity@163.com

Received 21 November 2021; Revised 16 December 2021; Accepted 17 December 2021; Published 17 January 2022

Academic Editor: Tongguang Ni

The purpose of video key frame extraction is to use as few video frames as possible to represent as much video content as possible, reduce redundant video frames, and reduce the amount of computation, so as to facilitate quick browsing, content summarization, indexing, and retrieval of videos. In this paper, a method of dance motion recognition and video key frame extraction based on multifeature fusion is designed to learn the complicated and changeable dancer motion recognition. Firstly, multiple features are fused, and then the similarity is measured. Then, the video sequences are clustered by the clustering algorithm according to the scene. Finally, the key frames are extracted according to the minimum amount of motion. Through the quantitative analysis and research of the simulation results of different models, it can be seen that the model proposed in this paper can show high performance and stability. The breakthrough of video clip retrieval technology is bound to effectively promote the inheritance and development of dance, which is of great theoretical significance and practical value.

1. Introduction

With the continuous progress of multimedia technology and computer network, video and images show more positive significance in daily life, and the amount of video image data is increasing geometrically [1]. Therefore, for video data, how to index it and finally retrieve it quickly and accurately has become an urgent demand [2]. At first, the way of human communication was sound and language, and then words and graphics appeared [3]. In modern civilized society, the emergence of digital products such as digital cameras and digital video cameras has further made images and videos a popular way of information exchange [4]. Video has developed into the main carrier of information dissemination, enriching people's lives and bringing opportunities for the vigorous development of artificial intelligence and big data industry [5]. Among them, computer vision plays a vital role. Motion recognition is a very challenging subject in the current research field of computer vision. Its purpose is to analyze video data using image processing [6–8] and classification recognition technology [9,10] to recognize human motion [11]. Effective fragment retrieval of dance video can help dance teachers arrange dance and assist Dance Teaching [12]. The breakthrough of dance video retrieval technology will effectively promote the inheritance and development of dance.

Every day, a large amount of video data is generated, and digital video is becoming more widely used in all aspects. With the large increase in video data, video database management systems have received a lot of attention and have a lot of potential [13]. Due to the large amount of video data, the current standard practice is to first detect and segment the video and then select several representative still image frames, referred to as key frames, from the lens to represent the visual content of the entire lens [14]. On the basis of video segmentation into shots, key frame extraction analyzes the color, texture, and other characteristics of image frames in the lens and finds the image frame that best represents the lens content based on the relationship between frames [15]. There are several common key frame extraction methods available today, but most of them are only effective for specific videos and cannot be applied to other videos, and the extracted key frames do not always represent the video's main content [16]. Using the key frame

as the index, extract the key frame set from the video sequence, summarize the original video content from high-level semantic information to low-level visual features, and retrieve the original video content. As a result, the number and quality of extracted key frames have a direct impact on the final search results' efficiency and accuracy [17]. Based on this, this paper proposes a multifeature fusion-based video key frame extraction method and applies it to dance action recognition.

Image retrieval is an early well-known video retrieval technology. It distinguishes the video by manually labeling some text descriptions or numbers. When retrieving the video, it uses the labeled label to search [18]. The content-based video retrieval method retrieves massive video data in the database according to the relationship between video content and context, in order to provide a visual feature algorithm that can automatically understand and recognize video in an unsupervised state [19]. Content-based video retrieval extracts the lowest level features to the high-level semantic features, analyzes and processes the video, automatically establishes an index of video data, and retrieves and browses according to the index [20]. The research results of motion recognition technology based on dance video not only are conducive to the analysis of dance video by dance professionals, but also can be used for teaching, protection, and excavation of artistic and cultural heritage. In addition, the research of motion recognition method based on dance video will also play a positive role in the research of human motion recognition in a large number of real and complex environments, enriching the application fields of motion recognition technology [21]. If the motion recognition technology is applied to the analysis of these music and dance videos, so as to obtain the organically related music and dance motion fragments, it can not only reduce the work intensity of dance professionals and facilitate the retrieval of music and dance video data, but also make the automatic dance arrangement system more efficient.

Firstly, this paper shows the feature extraction process and model in the key frame extraction method of music and dance video and then applies the feature fusion and recognition method to the key frame extraction of music and dance video. The simulation results show that the method proposed in this paper has high performance and accuracy.

2. Related Work

Literature [22] extracts the features of edge force field by Boosting classifier and designs a human posture estimation algorithm based on component detection. Literature [23] proposed an appearance model combining histogram and color features to estimate the pose of dance movements. However, due to the complexity of human posture changes, it is difficult for traditional methods to achieve effective posture estimation. Therefore, the method based on deep learning [9,24,25] is gradually used for human posture estimation. Literature [26] designed an hourglass-shaped neural network structure to extract multiscale features and identify dance movements. Literature [27] proposed a method to obtain human skeleton map by partial affinity domain. In addition, many dance movement recognition algorithms based on deep learning have been proposed one after another. Literature [28] takes out a fixed number of image frames at the first frame, the first frame, the second frame, or the equally spaced positions as key frames. Literature [29] selects multiple key frames according to the significant changes between frames. Firstly, the first frame of the shot is taken as the key frame, and then the difference between the previous key frame and the remaining frames is calculated. If the difference is greater than a certain threshold, another key frame is selected. Literature [30] proposed a method of extracting key frames based on shot activity. Firstly, the histograms of internal frames and reference frames were calculated, and then the activity marks were calculated. According to the curve of activity, the frame with local minimum is regarded as the key frame.

These methods often do not consider the change and complexity of the visual content in the lens. Most of the more complicated methods measure the similarity between any two frames in the shot by means of some underlying features such as color, texture, and motion information and divide all frames in the shot into different classes by combining threshold or clustering and then select representative frames from each class as key frames. Therefore, this paper proposes a method of dance motion recognition and video key frame extraction based on multifeature fusion, which is used to learn complex and changeable dance motion recognition. Through the steps of preprocessing, classifying, and indexing video data, a practical, convenient, and economical video retrieval system is developed, and the mechanism of video information retrieval and browsing scheme is improved.

3. Key Frame Extraction Method of Music and Dance Video

The relationship between video data units in terms of operation is unclear. The relationship between video segments is complex and difficult to define precisely, which introduces a slew of new issues into the setup and operation of a video database. It is difficult to process unstructured video data directly because it is difficult to measure the similarity between two unstructured data [31]. The successful application of motion recognition technology in other fields provides us with a sufficient theoretical foundation to apply it to dance video motion recognition. Currently, there are a large number of music and dance video materials, and professionals must spend a significant amount of time listening to and looking at these dance video materials, which is clearly inefficient. A specific action category is thought to have generated the image sequence in the video. As a result, the single-layer motion recognition method is primarily concerned with how to represent and match videos [32]. One or more frames of images that reflect the main information content in a group of shots and can express the shot content succinctly are known as key frames. Because each shot is taken in the same scene, each frame of images in the same shot contains a lot of the same information. Feature extraction is usually the first step in motion recognition research.

The feature extraction process mainly consists of three parts: the first part is to extract directional gradient histogram features by using the method of accumulating edge features; the second part mainly extracts the directional histogram features of optical flow from the dance data set; the third part extracts the corresponding audio stream files from the dance action videos and then extracts the audio signature features from the audio stream files. The specific feature extraction process is shown in Figure 1.

To achieve the data compression effect, only the key frames of the shot can be stored due to storage capacity. Second, key frames are used to represent shots, similar to keywords in text retrieval, so video shots can be processed by image retrieval technology. 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 the dance action video after framing in order to extract a set of key frames with less redundancy and can summarize the video content. For smaller objects, this method can match movements with large displacement and estimate optical flow. At the moment, there is not much of a difference in the visual characteristics and content of the image frame.

When a video stream is segmented into a series of semantically independent shots, although the amount of data that needs to be analyzed and processed is segmented, the amount of image data in the shots is still huge. To reduce the amount of data in video index, it is more important to facilitate users to retrieve video information and improve retrieval efficiency. It is necessary to extract one or more key frames from a shot according to the complexity of the shot content [33]. Since the shot is composed of frame images that are continuous in time and highly relevant in content, the most irrelevant frames can be selected as the key frames of the shot to contain the most information. The specific algorithm is to let $f_i (i = 1, 2, \ldots, N)$ be the feature vector of the i-th frame of a shot with N frames of images and define the correlation coefficient between feature vectors f_i and f_j as

$$\rho_{ij} = \frac{C_{ij}}{\sigma_i \sigma_j}. \tag{1}$$

Here, $C_{ij} = (f_i - m)(f_j - m)$, $\sigma_i^2 = C_{ii}$, and m is the mean vector. Select the k frame $r_1, r_2, \ldots, r_k \in \{1, 2, \ldots, N\}$ with the smallest correlation as the key frame ($k << N$):

$$R\left(f_{r_1}, f_{r_2}, \Lambda, f_{r_k}\right) = \left\{ \sum_{i=1}^{r_{k-1}} \sum_{j=1}^{r_k} \left(\rho_{r_i, r_k}\right)^2 \right\}^{\frac{1}{2}}. \tag{2}$$

The main problem of the above method is that the amount of calculation is too large, because it is necessary to calculate the correlation for any two frames. The method is simplified, and 1 to 3 frames of images are automatically extracted as key frames according to the different characteristics of the lens.

Let f denote a frame of image and $S = \{f_m, m = 1, 2, \ldots, N\}$ denote a shot with N frames. Take image frames f_1, $f_{N/2}$, and f_N as candidate key frames. Define the distance between two images f_i and f_j as

$$D\left(f_i, f_j\right) = \sum_{x,y} \left| f_i(x, y) - f_j(x, y) \right|. \tag{3}$$

When extracting key frames, first calculate the distance between two candidate key frames, namely, $D(f_1, f_{N/2}), D(f_1, f_N), D(f_{N/2}, f_N)$. Compare them with a predetermined threshold T. If they are both smaller than T, it means that they are relatively close. At this time, take $f_{N/2}$ as the key frame. If they are all larger than t, it means that there is a big gap between them. At this time, all three frames are regarded as key frames. In other cases, take the two images with the largest distance as key frames.

Key frames are digital images that contain the most intuitive information summary for users of video retrieval systems. The summary should as much as possible express the main content of the shot, so that the user understands the content to be expressed in the video from the start. The envelope and music energy features of music will be extracted in this paper, and the music feature and entropy sequence will be fused to produce a music-related entropy sequence. Figure 2 depicts the main flow of video key frame extraction.

Use the optical flow calculation method to obtain the movement characteristics of dance videos:

$$\begin{aligned}
E(w) &= E_{\text{color}}(w) + \gamma E_{\text{grad}}(w) + \alpha E_{\text{smooth}}(w) \\
&\quad + \beta E_{\text{match}}(w, w_l) + E_{\text{desc}}(w_1), \\
E_{\text{color}}(w) &= \int_{\Omega} \psi\left(\left|\nabla I_2(x + w(x)) - \nabla I_1(x)\right|^2\right) \mathrm{d}x, \\
E_{\text{grad}}(w) &= \int_{\Omega} \psi\left(\left|\nabla I_2(x + w(x)) - \nabla I_1(x)\right|^2\right) \mathrm{d}x, \\
E_{\text{smooth}}(w) &= \int_{\Omega} \psi\left(|\nabla \mu(x)|^2 + \left|\nabla v(x)^2\right|\right) \mathrm{d}x, \\
E_{\text{match}}(w) &= \int \delta(x) \rho(x) \psi\left(\left|w(x) - w_1(x)\right|^2\right) \mathrm{d}x, \\
E_{\text{desc}}(w_1) &= \int \delta(x) \left| f_2(x + w_1(x)) - f_1(x) \right|^2 \mathrm{d}x.
\end{aligned} \tag{4}$$

Here, α, $ß$, and γ are adjustable weight parameters, and $E_{\text{color}}(w)$ is the assumption of brightness invariance, which is applicable to both color images and grayscale images. The influence of light is inevitable. Therefore, in order to reduce the influence of light, it is necessary to add gradient constraint $E_{\text{grad}}(w)$ on this basis and then smooth it through $E_{\text{smooth}}(w)$. The last two items are to construct descriptor matching and find its minimum value through variable models and optimizations.

Calculate the entropy value of the current optical flow diagram in chronological order:

$$S = -\sum_{k}^{m} p_k \log_2 p_k. \tag{5}$$

Here, p_k represents the proportion of pixels with a gray value of k in the image, m represents the gray level, and S is

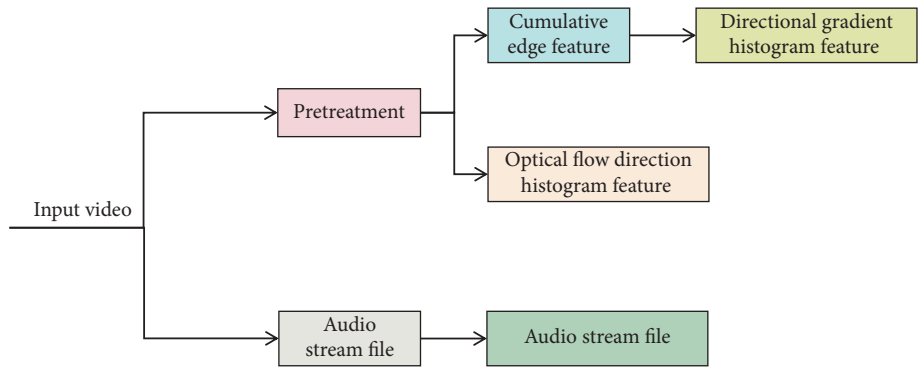

Figure 1: Feature extraction process.

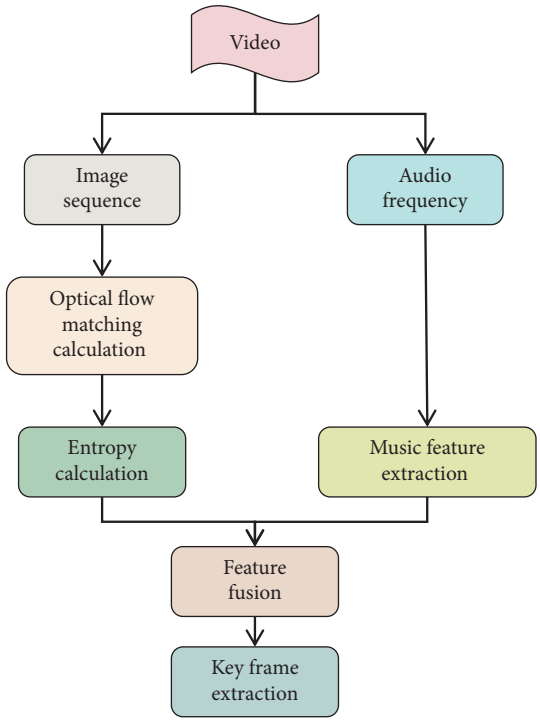

Figure 2: Dance key frame extraction process.

the entropy value. The greater the amount of information contained in the image, the greater the entropy value.

In the process of correspondence between audio and dance movements, the length of dance video, the frame number of images, and the frame rate of video are known. Then, the standard deviation is used to carry out interval operation to obtain the corresponding audio value per second, and the audio value and entropy value sequence are subjected to feature fusion.

4. Feature Fusion and Recognition

Although multiple key frames can more effectively describe the information expressed by shots than a single one, the likelihood of repeated or redundant video frames increases dramatically as the number of key frames increases. As a result, the focus and difficulty of key frame extraction technology is how to select the appropriate key frames that can not only represent the shot information, but also improve retrieval efficiency and reduce the amount of video index data. Optical flow directional histogram features are used to describe the motion information of dance movements, while directional histogram features are used to describe the local appearance and shape features of dance movements. Furthermore, the influence of music on dance should be considered when studying dance action recognition. All dancers perform with music playing in the background, and the type of music is related to the type of dance. Audio features, on the other hand, contain a lot of information, making them an important auxiliary feature that can help reduce the impact of self-occlusion on dance movements. Figure 3 depicts the multicore learning feature fusion process.

Suppose there are p dance moves $x_1, x_2, \ldots, x_p$ and category $y_1, y_2, \ldots, y_p$ in the dance data set. At the same time, the G kernel functions corresponding to the Histogram of Oriented Gradient (HOG) feature are defined as $k_g(x_i, x_j)$, the F kernel functions corresponding to the Histograms of Oriented Optical Flow (HOF) feature are defined as $k_f(x_i, x_j)$, $f = 1, 2, \ldots, F$, and the M kernel functions corresponding to the audio signature feature are defined as $k_m(x_i, x_j)$, $m = 1, 2, \ldots, M$. The linear combination of the kernel function combining the above three characteristics can be expressed by the following formula:

$$S = -\sum_{k}^{m} p_k \log_2 p_k. \tag{6}$$

Formula (6) satisfies $\beta_g \geq 0 \forall g, \beta_f \geq 0 \forall f, \beta_m \geq 0 \forall m, \sum_{g=1}^{G} \beta_g + \sum_{f=1}^{F} \beta_f + \sum_{m=1}^{M} \beta_m = 1$. β_g, β_f, and β_m are the weights of the corresponding kernel functions.

In order to express the content of the shot as completely as possible, conservative principles will be adopted when extracting key frames. When analyzing a video, if all the image frames at every moment are used, too many redundant image frames will be used. Therefore, people think of extracting key frames from thousands of image frames.

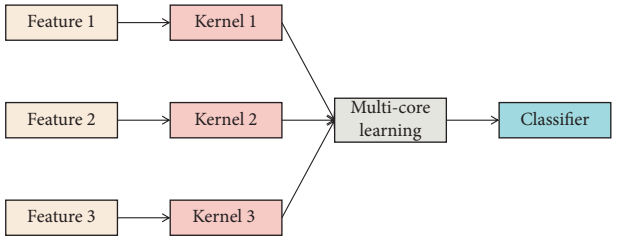

Figure 3: Multicore learning feature fusion process.

The use of key frames greatly reduces the amount of data in video index and also provides an organizational framework for searching and browsing videos.

5. Result Analysis and Discussion

In the past, motion recognition methods that relied on a single feature could only describe one aspect of human motion in video, but they could not effectively describe human motion. As a result, motion recognition research has turned to the multifeature fusion method. Combining different features can more comprehensively describe human motion in video, resulting in a better recognition effect. The dance video retrieval database is made up of key frames and video clips that have been summarized. Kernel functions play different roles in classification depending on the problem. The goal of multicore learning is to improve the classification effect by giving different kernel functions reasonable weights and combining multiple kernel functions to describe features more thoroughly. The data set is used to extract directional gradient histogram features, optical flow directional histogram features, and audio features for the dance motion recognition method described in this paper. Figure 4 shows the entire appearance feature of the audio signal as a result of envelope feature extraction from dance video accompaniment music.

The dance video is composed of a series of dance movements, and the coherent dance movements reflect more or less amount of movement. The motion information in the dance video is expressed by optical flow, and then the information in each optical flowchart is counted by entropy. Entropy sequence and music features are fused to obtain a music-related entropy sequence. Then, the key frames are selected by the threshold, and when the threshold is set, it will be compared with the key frame set selected by several users to select the best threshold suitable for the video. The matching of feature points starts from the first frame of the query segment, and the frames in the query segment are sequentially compared with the frames in the key frame set. Then select a key frame that is most similar to the frame of the query fragment. Video clips can be described by one or several key frames. No matter what level of similarity matching, if there are some dissimilar parts between the query segment and some subsegments, shots, or frames in the video segment, such segments are discontinuous. When this situation produces more dissimilarities, it indicates that the similarity between the two fragments is lower. Adaptive key frame extraction algorithm based on unsupervised

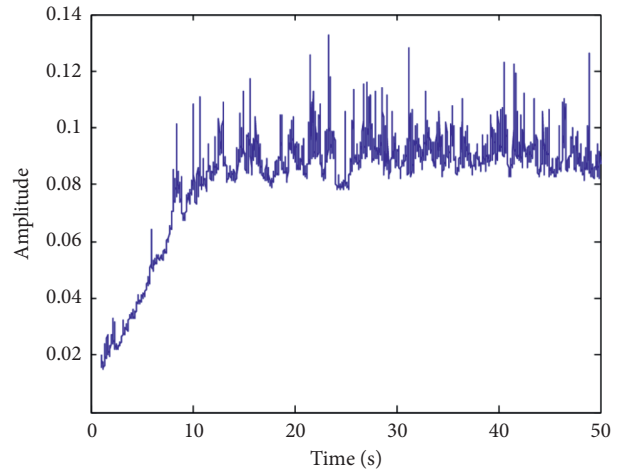

Figure 4: Envelope extracted from accompaniment music of dance video.

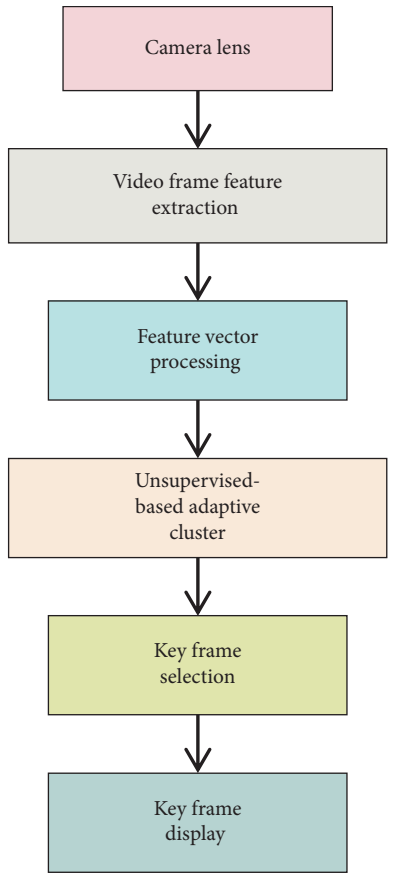

Figure 5: Algorithm implementation process.

clustering is used to extract key frames for different types of shots, as shown in Figure 5.

Different combinations of repetitive dance movements are frequently used to create dance videos. Similar dance movements will appear in various types of dance videos during this process, and eventually, everyone will follow this choreography pattern. We can discover that dance

movements in dance videos are closely related to music through segment retrieval of dance videos. All of the video frames in the shot are treated as independent subclusters at the start of merging, and pairwise similarity is calculated. The two most similar subcategories, that is, the least similar, are chosen and merged into a new subcategory. Merge according to this cycle, then wait until an automatic merging stop rule is satisfied before getting the final clustering result. We should extract the features of the image frames in the video first, then process the features of the image frames with the algorithm, and finally extract the key frames, regardless of which algorithm we use to extract the key frames. Figure 6 shows a simulation comparison of image key frame extraction reliability optimization.

The motion features of video describing human motion information are often essential features in the research based on motion recognition, and the optical flow method is usually used to extract the features in some related motion recognition research at present. Texture is a value calculated from the texture image, which quantifies the features of gray scale changes inside the texture. Generally, texture features are related to the position, direction, size, and shape of the texture but have nothing to do with the average gray level. The purpose of feature extraction is to transform the spatial structure difference of random texture or geometric texture into the difference of feature gray value and use some mathematical models to describe the texture information of the image, including the smoothness, sparseness, and regularity of the image area. The linear regression curve is calculated according to the stepwise multiple linear regression equation, as shown in Figure 7.

FS denotes a method of searching that uses F7 layer features, HS denotes a method of searching that uses coarse searching, and HFS denotes a method of searching that goes from coarse to fine. Figure 8 depicts the effect of various algorithmic features on the retrieved video key frames. Clearly, HFS is the first to demonstrate good retrieval accuracy, implying that the video GIS data retrieval algorithm can obtain richer detail features of video GIS key frames. When retrieving more than 13 images, however, the retrieval effect of HS outperforms that of FS. This demonstrates that the binary code generated by this algorithm has a high level of discrimination and contains a lot of semantic information.

For a video containing multiple shots, the key frame fidelity is the average of the fidelity of each shot and the group of key frames. The purpose of extracting key frames is to use as few video frames as possible to represent as much video content as possible, so the higher the compression rate, the more effective this key frame extraction method. Make the estimate continuous at the threshold. The shrinking trend of the adaptive nonlinear curve is shown in Figure 9.

The purpose of key frame extraction is to replace the whole video with few image frames, so as to facilitate the viewer to quickly browse the content of the whole video and reduce the amount of video data, thus making the video processing more convenient and faster. Therefore, the effectiveness of the key frame extraction method should be considered first, and then its computational complexity and

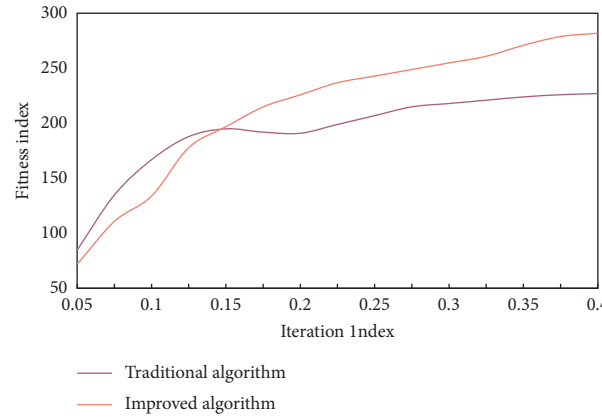

Figure 6: Image key frame extraction optimization simulation comparison.

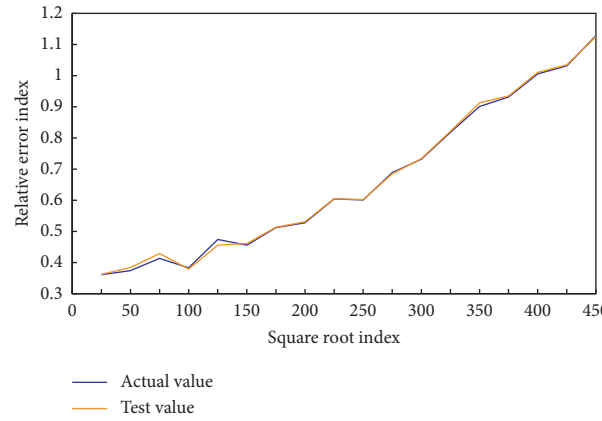

Figure 7: The relationship between the actual value and the calculated value of stepwise linear regression.

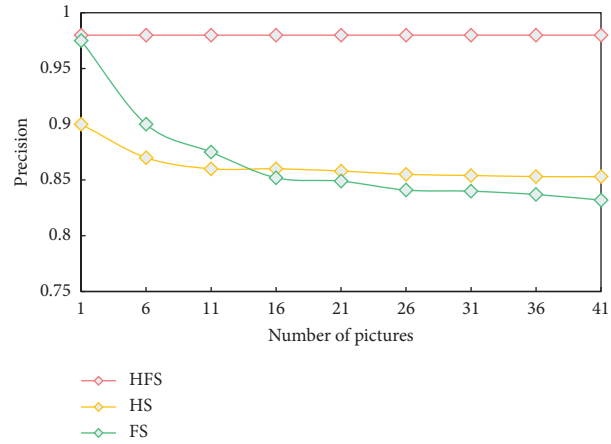

Figure 8: Retrieval effect of different network characteristics.

efficiency should be considered on the premise of effectiveness. The experimental results of the comparison between this method and the benchmark method in four dance combinations are shown in Figure 10. In the recognition of four dance combinations, the recognition rate of this method is higher than that of the benchmark method. The

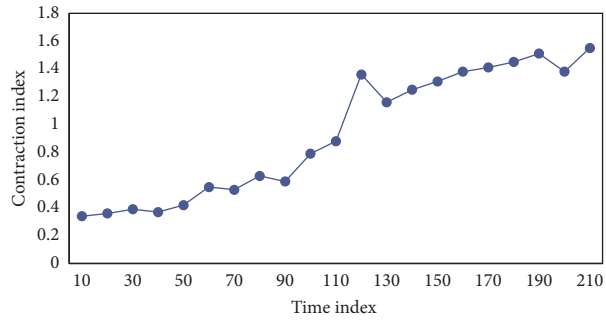

Figure 9: Adaptive nonlinear curve shrinking trend.

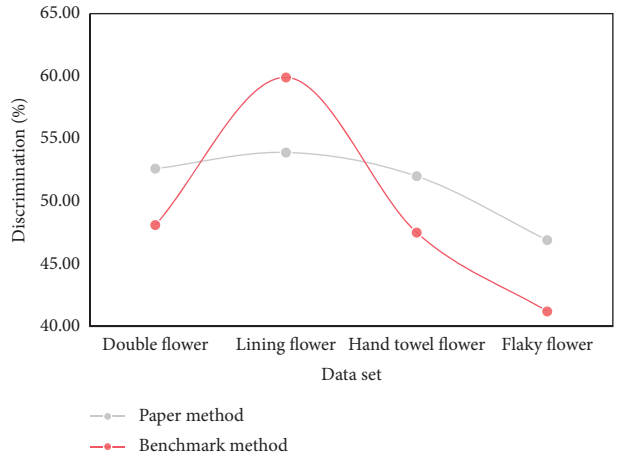

Figure 10: Comparison of the recognition rate.

recognition rate of this method is 53.9%, which is lower than 59.9% of the benchmark method. In other combinations, the performance of this method is better than that of the benchmark method, especially when the similarity of dance movements in the combination of towel and flower is too high.

When the dance movements are too complicated and there are similar movements and self-occlusion, the benchmark method based on trajectory feature fusion can not accurately represent the dance movements. The fusion algorithm in this paper can avoid the above influence to a certain extent, thus improving the recognition rate of dance, and it also verifies the effectiveness of the algorithm. The experimental results show that not only is the algorithm in this paper relatively simple in calculation, but also the extracted key frames can effectively summarize the main content of the video, realize video compression and storage, and lay a good foundation for video retrieval and video summarization. From the perspective of the future development trend of video data mining, the requirement for video data processing technology is getting higher and higher because of the contradiction between the large amount of calculation of video processing and the short retrieval time expected by users. Moreover, with the continuous development of information technology, it is more and more difficult to process video data with various features.

6. Conclusions

The dance contains far too many repetitive dance moves, which will slow down retrieval speed when searching. As a result, this paper proposes a method for extracting key frames from music and dance videos. First, the framed video's optical flow is calculated, and the video's motion features are extracted. Music and dance are inextricably linked. The corresponding audio in the video is then extracted, along with its features. This paper presents an unsupervised automatic video key frame extraction method. This method uses simple cyclic merging to cluster video frames and creates an automatic merging stop rule to stop merging when the clustering results are optimized. There is no need to set parameters or prior knowledge in advance for this video frame clustering process. The results of the experiments on multiple test videos show that the extracted key frames can effectively represent the video's main visual content.

Although some progress has been made in dance motion recognition research in this paper, the recognition rate of dance video motion recognition research is currently low, owing to the complexity of the dance motion and the inadequacy of existing methods for dance motion recognition. Because of the complexity of dance movements, the dance data set we created at this point only considers solo dance situations, ignoring changing stage scenes and other factors. More research will be done in the future on how to apply music theory-related content to create a more accurate mapping between music and dance movements, in order to improve dance movement recognition accuracy. [33].

Data Availability

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

Conflicts of Interest

The author declares no conflicts of interest.

References

[1] N. Ma, X. Shi, D. Qin, and C. Liu, "A key frame extraction method for music and dance videos," *Journal of System Simulation*, vol. 30, no. 7, pp. 2801–2807, 2018.

[2] K. Yang, H. Song, K. Zhang, and J. Fan, "Deeper Siamese network with multi-level feature fusion for real-time visual tracking," *Electronics Letters*, vol. 55, no. 13, pp. 742–745, 2019.

[3] R. Chi, Z. M. Lu, and Q. G. Ji, "Real-time multi-feature based fire flame detection in video," *IET Image Processing*, vol. 11, no. 1, pp. 31–37, 2017.

[4] A. N. Gong, M. Ding, and F. Dou, "Multi-feature fusion music emotion classification method based on DBN," *Computer Systems Applications*, vol. 26, no. 9, 164 pages, 2017.

[5] T. Liu, "Electronic music classification model based on multi-feature fusion and neural network," *Modern Electronic Technology*, vol. 41, no. 19, p. 5, 2018.

[6] 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.

[7] J. Kong, H. Wang, X. Wang, X. Jin, X. Fang, and S. Lin, "Multi-stream hybrid architecture based on cross-level fusion strategy for fine-grained crop species recognition in precision agriculture," *Computers and Electronics in Agriculture*, vol. 185, Article ID 106134, 2021.

[8] W. Cai, B. Zhai, Y. Liu, R. Liu, and X. Ning, "Quadratic Polynomial Guided Fuzzy C-Means and Dual Attention Mechanism for Medical Image Segmentation," *Displays*, vol. 70, Article ID 102106, 2021.

[9] M. Gao, R. Liu, and J. Mao, "Noise robustness low-rank learning algorithm for EEG signal classification," *Frontiers in Neuroscience*, vol. 15, p. 1618, 2021.

[10] 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*, pp. 1–5, In press.

[11] P. G. Bhat, B. N. Subudhi, T. Veerakumar, V. Laxmi, and M. S. Gaur, "Multi-feature fusion in particle filter framework for visual tracking," *IEEE Sensors Journal*, vol. 20, no. 5, pp. 2405–2415, 2020.

[12] 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.

[13] 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.

[14] M. Jian, S. Zhang, L. Wu, S. Zhang, X. Wang, and Y. He, "Deep key frame extraction for sport training," *Neurocomputing*, vol. 328, no. 7, pp. 147–156, 2019.

[15] R. Wang, J. Hu, J. Yang et al., "Video key frame extraction based on mapping and clustering," *Journal of Image and Graphics*, vol. 21, no. 12, p. 10, 2016.

[16] Z. Lan, S. Dan, and Y. Li, "Road surveillance video key frame extraction algorithm based on correlation coefficient," *Journal of Chongqing Jianzhu University: Natural Science Edition*, vol. 35, no. 1, p. 6, 2016.

[17] Y. U. Wang, R. Wang, and J. Yang, "A new adaptive video key frame extraction method," *Journal of Hefei University of Technology: Natural Science Edition*, vol. 39, no. 11, p. 6, 2016.

[18] J. Wang and X. Lu, "Video key frame extraction algorithm based on semantic correlation," *Computer Engineering and Applications*, vol. 57, no. 4, p. 7, 2021.

[19] M. Zhong and Y. Zhang, "Key frame extraction method of vehicle surveillance video based on visual saliency," *Computer Technology and Development*, vol. 29, no. 6, p. 6, 2019.

[20] J. Liang and H. Wen, "Video key frame extraction and video retrieval based on deep learning," *Control Engineering*, vol. 26, no. 5, p. 6, 2019.

[21] P. T. Sheeba and S. Murugan, "Hybrid features-enabled dragon deep belief neural network for activity recognition," *The Imaging Science Journal*, vol. 66, no. 5-6, pp. 355–371, 2018.

[22] H. Yasin, M. Hussain, and A. Weber, "Keys for action: an efficient keyframe-based approach for 3D action recognition using a deep neural network," *Sensors*, vol. 20, no. 8, p. 2226, 2020.

[23] I. Mademlis, A. Tefas, and I. Pitas, "A salient dictionary learning framework for activity video summarization via key-frame extraction," *Information Sciences*, vol. 432, pp. 319–331, 2018.

[24] Y. Huang, L. Cheng, L. Xue et al., "Deep adversarial imitation reinforcement learning for QoS-aware cloud job scheduling," *IEEE Systems Journal*, pp. 1–11, 2021, In press.

[25] Q. Liu, T. Xia, L. Cheng, M. Van Eijk, T. Ozcelebi, and Y. Mao, "Deep reinforcement learning for load-balancing aware network control in IoT edge systems," *IEEE Transactions on Parallel and Distributed Systems*, vol. 33, no. 6, pp. 1491–1502, 2021.

[26] H. Xiaoli and Y. Gao, "Multi-level extraction algorithm for mutual information entropy of video key frames under CUDA framework," *Journal of University of Electronic Science and Technology of China*, vol. 047, no. 5, pp. 726–732, 2018.

[27] R. Liang, Q. Zhu, and J. Hu, "Video key frame extraction based on the idea of multi-layer core set cohesion," *Computer Engineering and Design*, vol. 37, no. 6, pp. 1567–1572, 2016.

[28] M.-A. Li, J.-F. Han, and J.-F. Yang, "Automatic feature extraction and fusion recognition of motor imagery EEG using multilevel multiscale CNN," *Medical, & Biological Engineering & Computing*, vol. 59, no. 10, pp. 2037–2050, 2021.

[29] S. Kannappan, Y. Liu, and B. Tiddeman, "DFP-ALC: automatic video summarization using distinct frame patch index and appearance based linear clustering," *Pattern Recognition Letters*, vol. 120, no. 4, pp. 8–16, 2019.

[30] J. Sun and Y. Li, "Multi-feature fusion network for road scene semantic segmentation," *Computers & Electrical Engineering*, vol. 92, no. 12, Article ID 107155, 2021.

[31] J. Zhong, Y. Zhang, Y. Pang, and X. Li, "Hypergraph dominant set based multi-video summarization," *Signal Processing*, vol. 148, no. 7, pp. 114–123, 2018.

[32] X.-B. Shi, X. Shi, and N. Ma, "A key frame extraction method for music and dance videos," *Journal of System Simulation*, vol. 30, no. 7, p. 7, 2017.

[33] M. Li, "On the integration of dance in music teaching in secondary vocational schools," *Education Modernization*, vol. 5, no. 18, pp. 292-293, 2018.