Enhancing Research on Dance Action Recognition by Integrating Collaborative CAD Through Multimodal Fusion

Xiaoqiang Yang[1*]

[1] School of Music and Dance, Hainan University, Haikou, 570228, China

Corresponding author: Xiaoqiang Yang, yangxiaoqiang1983@163.com

Abstract. Dance videos have gained significant popularity among scholars. The optimization of dance video segment retrieval practices could prove advantageous for dance instructors in organizing dance routines and facilitating dance instruction. This paper examines a dance action detection system that utilizes multi-feature fusion techniques, specifically emphasizing the directional gradient histogram feature, optical flow (OF) directional histogram feature, and audio signature feature. The regions of the human body encompassed by the frame are obtained by applying human body posture. The 3D-SIFT (Scale-Invariant Feature Transform) and OF features are obtained from their respective regions. SIFT is a technique that records fixed data regarding the human body. The inclusion of the temporal component enables the acquisition of continuous static information. This facilitates enhanced integration with optical fiber (OF) features, encompassing dynamic information about the human body. The results derived from the examination and evaluation performed on the dataset indicate that the algorithm exhibits the capacity to recognize the dance video database. In addition, it significantly improves the accuracy of detecting dance movements, allowing dancers to apply action correction features.

Keywords: Multi-feature fusion; Collaborative CAD; Dance motion recognition; Optical flow; Feature extraction
DOI: https://doi.org/10.14733/cadaps.2025.S5.229-243

1 INTRODUCTION

The field of computer vision presents notable difficulties in the realm of motion recognition research. The main goal is to utilize image processing and classification recognition methods to assess video footage and detect human movement. The field of motion recognition has gained prominence in recent years due to its substantial research value. This technology can be used in intelligent dance training to obtain the dancer posture skeleton map by extracting dancer picture features. The objective of this study is to determine the dancer's rhythmic movements and evaluate and correct their body alignment [1].

The field of human motion recognition holds significant potential for various applications, primarily evident in the following domains: Firstly, intelligent video surveillance is employed to enhance the effectiveness of monitoring. Additionally, the human-computer interaction system is employed to identify, analyze, and comprehend individuals' postures and behaviors, hence facilitating intelligent communication between humans and machines [2]. The field of motion recognition has evolved from basic hand or foot interactions to the analysis and acquisition of human body movements on a worldwide scale. In the realm of subtlety, there has been a notable increase in the identification rate of intricate activities, such as facial expression detection and gesture recognition. During the course of acquiring knowledge and comprehending content-based video retrieval, it has been observed that the retrieval of action videos has gained significant popularity. These videos typically encompass activities such as running, leaping, swimming, weightlifting, and similar activities [3-5]. Nevertheless, during the course of research, it has been observed that there exists a dearth of scholarly investigations pertaining to dance videos. This scarcity can be attributed to the inherent nature of dance as a dynamic and fluid form of bodily motion, wherein many dance genres exhibit distinct variations in limb movements and intensity fluctuations. Another factor to consider is the limited availability of dance data sets.

This paper mainly studies the dance action recognition method based on multi-feature fusion, the core of which is to extract various features that can describe moving objects from video sequences and analyze and recognize them. In this paper, the research of motion recognition mainly solves two major problems: motion representation and motion classification. Action representation is the extraction and description of action features, which means that the actions in the video are represented by features that are easy to distinguish. Action classification means that the features that are easy to distinguish and extracted from the video are classified by classification methods so as to achieve the purpose of classifying actions.

2 RELATED WORK

Initial attempts to estimate human posture primarily centered around the analysis of human contour features or component models. In reference [6], the Boosting classifier is utilized to extract the feature of the edge force field. Subsequently, a human posture estimation technique based on component detection is developed. In reference [7], a proposed model is presented for estimating human posture using a combination of histogram of oriented gradient (HOG) and color features. In reference [8], a global network is employed to identify basic key points, followed by the detection of intricate key points using RefineNet in order to estimate the attitude. Literature [9] suggests the use of motion contour to depict the overall attributes of a human action, as well as the utilization of motion energy diagram and motion history diagram to illustrate the process change associated with the precise execution of the action. The literature [10-11] expands the understanding of human physical attributes from a two-dimensional to a three-dimensional perspective and proposes a study approach that considers both spatial and temporal aspects of body features. The three-dimensional motion history map is represented in the literature [12] by utilizing a mixture of the original, forward depth, and backward depth.

The utilization of local features in research on human motion recognition is advantageous due to their resistance to external environmental changes and enhanced anti-interference capabilities. The literature [13] presents a study methodology that utilizes spatio-temporal interest points, expanding the scope of interest points from a two-dimensional space to a three-dimensional space. The literature [14] expands upon the concept of gradient histogram by extending its application from two-dimensional space to three-dimensional space. It suggests that the three-dimensional gradient histogram can be employed to classify human actions in video. Furthermore, a more intricate technique has been established by reference [15], wherein the output of a linear dynamic system is utilized to represent low-level segments, also known as primitives. Previous studies have demonstrated that motion segmentation is inherent in the representation of state machines or motion graphs [16-17]. Nevertheless, the utilization of low-level segmentation yields segments that are partial, hence failing to accurately depict a comprehensive series of actions. According to the literature [18], three techniques are suggested for the automated segmentation of a lengthy sequence into action subsequences. The first two methods are applied in practice films, where the algorithm systematically examines all frames from start to finish and generates segmentation when breakpoints are detected. Literature [19] employs Canny edge detection to extract shape information from human body actions, which is then utilized to represent action edges. This approach aims to accomplish human body action recognition by matching comparable edges. The literature [20] expands upon the conventional Shape from Silhouette (SFS) method, which is only suitable for objects with rigid body motion. It also extends this method to hinged objects, allowing for the acquisition of shape and motion information for different parts of the human body. By solving simple motion constraint equations between hinged parts, the method estimates the position of human joints, thereby achieving the objective of motion recognition [2]. Literature [21] has modified the conventional approach of teaching attitude estimate and action recognition separately and merged them in a sequential manner. Instead, it has provided a framework that combines attitude estimation and action recognition. The precision of motion recognition has achieved a top-tier level, and the estimation of attitude has been enhanced. The literature literature [22] proposes a technique for action recognition using spatiotemporal tree sets. This method involves identifying hierarchical spatiotemporal trees from training data and subsequently constructing action models based on these trees to classify actions in movies. The utilization of hierarchical spatio-temporal trees for the representation of actions has the potential to enhance the robustness of middle-level features in representing activities. Nevertheless, the exponential search field poses challenges in terms of frequent identification and differentiation of tree structures.

3 RESEARCH METHOD

3.1 Common Features of Behavior Recognition

The dense trajectories (DT) algorithm and its improved dense trajectories (IDT) and space-time interest points (STIP) algorithms have been representative behavior recognition feature algorithms in recent years. The IDT algorithm is robust and performs well when moving targets. The recognition accuracy is widely used [23]. Since IDT performs feature extraction along the optical flow(OF) field trajectory, it contains more underlying feature information than STIP.

The global feature describes the whole frame information of the frame to be identified, which is represented by the OF feature and grayscale feature. The local feature describes the block information extracted from the image around the target to be identified, which mainly represents STIP and HOG features, etc. From the perspective of algorithm efficiency, most behavior recognition algorithms adopt the strategy of local feature extraction, sacrificing accuracy to obtain higher recognition speed.

3.1.1 HOG Feature

The main idea of the histogram of oriented gradient (HOG) feature is to get the edge feature of the target by counting the gradient direction histogram of the local area in the image to be recognized.

HOG feature maintains good invariance with the running shape and illumination change of the target, which makes it more suitable for human detection.

The HOG feature needs to calculate the pixel gradient, gradient amplitude, and gradient direction. The gradient calculation method of a pixel is shown in the following formula:

$$G_x(x,y) = H(x+1,y) - H(x-1,y)G_y(x,y) = H(x,y+1) - H(x,y-1) \tag{1}$$

The original image is connected to the $[-1,0,1]$ and $[1,0,-1]^T$ Gradient operators and the gradient values in horizontal and vertical directions are obtained. Formula (2) can calculate the gradient amplitude and direction at $(x \; and \; y)$.

$$G_x(x,y) = \sqrt{G_x(x,y)^2 + G_y(x,y)}\alpha(x,y) = tan^{-1}\left(\frac{G_y(x,y)}{G_x(x,y)}\right) \tag{2}$$

The image gradient is well-equipped to describe the contour and texture of the target, while the illumination change has little influence on the image gradient.

3.1.2 OF Characteristics

Optical flow(OF) is defined as the instantaneous velocity of the corresponding point movement of the space object on the imaging plane. The corner tracking algorithm extracts the changes of corner points to construct OF, and the improved dense point trajectory algorithm densely samples feature points at multiple scales and tracks them to construct OF.

In applications, it is very common to use histograms of oriented OF (HOF) to extract the features of the target OF. The OF histogram calculates the OF field corresponding to each image in the video and then calculates the angle histogram of the included angle of the OF to characterize the OF.

The extraction process of the OF histogram is as follows:

Calculate the velocity vector. Calculate the OF field corresponding to the pixel motion of each frame of the image, and then get the velocity vector corresponding to the pixel point:

$$v = [x,y]^T \tag{3}$$

Among them, v is the vector, composed of the magnitude information (intensity) and the direction of motion.

Statistical histogram. Calculate the angle between the horizontal axis and the vector OF, then get the projection size and count the histogram OF.

$$v = [x,y]^T, \theta = tan^{-1}\left(\frac{y}{x}\right) \tag{4}$$

When the angle falls on $-\frac{\pi}{2} + \pi\frac{b-1}{B} \leq \theta \leq -\frac{\pi}{2} + \pi\frac{b}{B}$, its amplitude $\sqrt{x^2+y^2}$ acts on the b bin $(1 \leq b \leq B)$ of the histogram, and finally, the histogram is normalized. Reducing the sensitivity of the HOF feature to the moving direction of the target is calculated based on the horizontal axis. The recognition effect is typically better when the histogram bin is more than 30.

3.2 Dance Motion Recognition Based on Multi-Feature Fusion

3.2.1 Feature Extraction

The Scale-invariant feature transform (SIFT) technique follows a sequential process of extracting key points, describing the information of these key points, and ultimately matching them. This paper presents an algorithm that integrates 3D-SIFT features with OF features. The primary purpose of utilizing 3D-SIFT features is to identify significant points in the image. Subsequently, the gradient direction and size of the entire neighborhood of these points are calculated. Finally, 3D-SIFT descriptors are generated through sub-histogram coding.

The key points refer to local extremum sites that possess directional information across spatial representations of varying scales. The purpose of scale space is to apply specific rules to the original image, resulting in the creation of a pyramid-shaped sequence that represents spatial information at several scales. The schematic representation of the scale space is depicted in Figure 1. The left image depicts a two-dimensional scale space, whereas the right image illustrates a three-dimensional scale space with the inclusion of a time dimension.

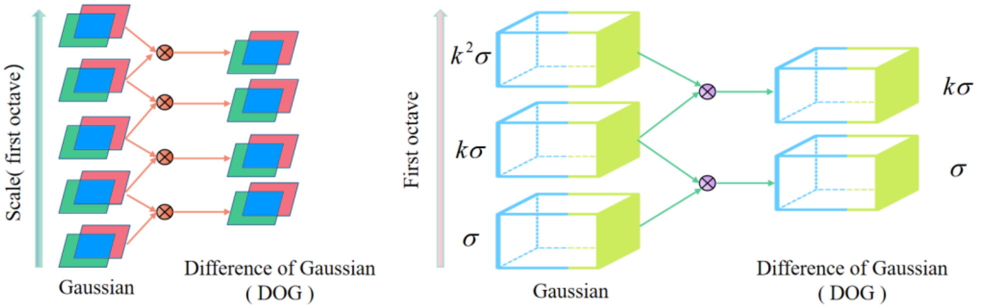

Figure 1: Scale-space schematic diagram.

Once the scale space has been constructed, crucial points are chosen by selecting local extremum points. The majority of the crucial elements are situated in the periphery and apex of objects in the image, specifically, the points exhibiting significant alterations. The concept of locality pertains to the spatial arrangement of pixels, encompassing the nin region that is centered on the pixel inside the same frame image, as well as the nin region that is located in the same place within its neighboring images. Upon conducting a comparison, the key point in the current scale is determined by selecting either the largest or smallest value inside the neighborhood.

This paper employs a combination of pixel OF value extraction from the previous section and a threshold OF value of 0.3 to filter out pixels with asymmetric local curvature of DOG. The initial value is set at 0.3, but the final threshold value is determined through experimental analysis. If the selected key point of value is found to be lower than the predetermined threshold value, it will be eliminated.

The assignment of orientation defines the direction and magnitude of the gradient for each key point. The dimensions and orientation of each pixel's 2D gradient are precisely specified as:

$$m_2D(x,y) = \sqrt{L_x^2 + L_y^2}, \theta(x,y)\, tan^{-1}\left(\frac{L_y}{L_x}\right) \tag{5}$$

x, y is the coordinate of the pixel in the image, and L_x, L_y is obtained by finite difference approximate calculation:

$$L_x \approx L(x+1,y,t) - L(x-1,y,t) L_y \approx L(x,y+1,t) - L(x,y-1,t) \tag{6}$$

The 3D-SIFT is similar to the 2D-SIFT calculation method, but the 3D-SFT needs to calculate the spatiotemporal gradient $\left(L_x, L_y, L_t\right)$. Besides $L_x,$ ⬚⬚ Ly, and L_t also use finite difference calculation, which is approximate as follows:

$$L_t \approx L(x,y,t+1) - L(x,y,t-1) \tag{7}$$

Then, use L_x, L_y and to calculate the gradient size and direction of 3D:

$$m_3D(x,y,t+1) = \sqrt{L_x^2 + L_y^2 + L_t^2} \tag{8}$$

$$\theta(x,y,t) = tan^{-1}\left(\frac{L_y}{L_x}\right) \tag{9}$$

$$\varphi(x,y,t) = tan^{-1}\left(\frac{L_t}{\sqrt{L_x^2+L_y^2}}\right) \tag{10}$$

Because the $\sqrt{L_x^2 + L_y^2}$ is positive, $\varphi \in \left(-\frac{\pi}{2}, \frac{\pi}{2}\right)$ always exists, and a unique θ,φpair represent(⬚)s each corner, so two values represent the gradient direction of each pixel in three dimensions.

When using the meridian parallel method, it should be noted that bins need to be normalized by solid angle ω. The normalized value of bin area is added to each bin, also called solid angle. If the stereo angle is not normalized, the direction histogram will get the wrong weight. The solid angle is calculated as follows:

$$\omega = \int_{\varphi}^{\varphi+\Delta\varphi} \int_{\theta}^{\theta+\Delta\theta} sin\,\theta\, d_\theta\, d_\varphi = \Delta\varphi \int_{\theta}^{\theta+\Delta\theta} sin\,\theta\, d_\theta = \Delta\varphi[-\cos\theta]_{\theta}^{\theta+\Delta\theta} = \Delta\varphi\big(\cos\theta - \cos\theta\,(\theta+\Delta\theta)\big) \tag{11}$$

The normalized values are added to the histogram as follows:

$$hist\left(i_\theta, i_\varphi\right) + \frac{1}{\omega} m_{3D}(x', y', t') e^{\frac{-\left((x-x')^2-(y-y')^2-(t-t')^2\right)}{2\sigma^2}} \tag{12}$$

In which $(x, y, and\ t)$ represent the coordinates of the point of interest, open paren x prime,y prime,t prime, and close paren represent the coordinates of the pixel added to the direction histogram, and the peak value of the histogram is the main direction. The main peak is stored because it can be used to rotate the neighborhood of critical points and create rotation-invariant features.

3.2.2 Multi-Feature Fusion Module

The neural network architecture employed in this study bears a resemblance to ResNet [24], with both models incorporating a five-stage feature extraction process. The representation of convolution characteristics with varying scales consists of five stages, namely C1, C2, C3, C4, and C5. The composition of each stage consists of many residual blocks (RBs), which ultimately extract characteristics to produce thermal maps of significant locations in human posture. Figure 2 displays a backbone network that is constructed using Feature Pyramid Networks (FPN).

The backbone network demonstrates proficiency in accurately identifying basic visible key points. However, it encounters challenges when attempting to identify key points related to human posture inside intricate environments, including those that are concealed. In this paper, a multi-feature fusion module is devised to address the demand for more comprehensive feature information in determining the location of complicated key points.

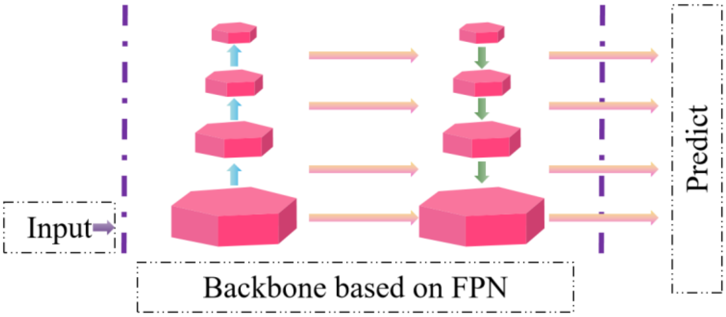

Figure 2: Schematic diagram of backbone network based on FPN.

The backbone network based on FPN is employed to identify basic key point estimation, while the multi-feature fusion module is utilized to handle more intricate key point estimation. The structural arrangement is depicted in Figure 3.

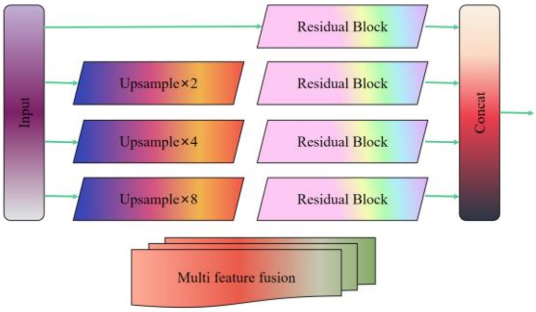

Figure 3: Multi-feature fusion structure diagram.

This research employs RB to augment the feature extraction of each layer, hence acquiring more comprehensive semantic characteristics. Simultaneously, the enhancement of local features is achieved through the use of up-sampling operations, which aim to improve the feature resolution at each stage. Ultimately, every characteristic from FPN is combined using the P operation.

FPN extracts distinctive characteristics throughout the training process and finds crucial locations inside the human skeleton. The FPN stage largely emphasizes the completion of essential and straightforward tasks. The multi-feature fusion module aims to improve the comprehension of complex and significant elements, such as occlusion and concealment. The process involves gathering knowledge of the features present in each layer of the FPN and subsequently merging them, leading to the retrieval of the heat map of the human skeleton.

3.2.3 Fusion of Audio Features and Entropy Sequences

Dance is an art with human body movements as its primary expression. Among the dance works with different themes, the relationship between dance and music is the closest, and the quality of a dance work and the accompaniment of music in dance play a key role. Without good music, it isn't easy to produce beautiful dance works [25]. In this paper, the authors propose a method for extracting the accompaniment music from a video and utilizing the extracted music in WAV format for further analysis. The purpose of this extraction is to capture the characteristics of the accompaniment music, which can be beneficial for event detection related to the music.

Firstly, the energy feature of audio is extracted by framing, and then the audio $x(j)$ is windowed and framed to get the kth audio.

The audio signal is stored in y, tits length is cap N, the sampling rate is f s, the length is w l e n every time, the displacement of the front and back frames is d i. s, and the overlapping part between the two frames is o l a. o equals w l e n minus d i. s. Therefore, for an audio signal with a length of cap N, the formula of framing is shown in formula (13).

$$fs = (N - olap)/dis = (N - wlen)/dis + 1 \tag{13}$$

Then, calculate the average amplitude of audio, that is, the energy characteristics of audio, as shown in the following formula.

$$y_k(j) = win(j) \cdot x\big((k-1) \cdot dis + j\big), 1 \leq j \leq L, 1 \leq k \leq f \tag{14}$$

$$M(k) = \sum_{j=0}^{L-1} |y_k(j)|, 1 \leq k \leq f \tag{15}$$

$win(j)$ is the window function, $y_k(j)$ is the value of one frame, $L$is is the frame length, dis is the frameshift length and f is is the total number of frames after framing. In formula (15), $M(k)$ represents the characterization of the energy of an audio frame.

In this way, the audio feature sequence can be aligned with the entropy sequence. Finally, through the product operation of audio feature sequence and entropy sequence, feature fusion is carried out to obtain a music-related entropy sequence.

The aforementioned procedures are applied to obtain the envelope feature sequence and energy feature sequence of voice recordings. In order to analyze the correlation between music and dance movements, it is crucial to possess knowledge regarding the duration of the dance video, the number of image frames, and the frame rate. Furthermore, the utilization of the standard deviation enables the execution of interval operations, hence facilitating the calculation of the associated audio value per second. In order to construct a feature, the audio value is finally mixed with an entropy sequence.

4 DISCUSSION AND ANALYSIS OF RESULTS

4.1 Dance Data Set

This research utilizes two dance data sets, specifically the DanceDB data set and the FolkDance dataset developed by the author's laboratory. The DanceDB dataset incorporates emotional signals throughout each dance category. The data set of FolkDance dance is partitioned into four distinct groups. Each group has many subdivided dance motions that exhibit a wealth of action categories. Furthermore, each group of dance movements is characterized by its inherent complexity and level of difficulty.

The DanceDB dance data collection consists of 48 dance videos with fixed backgrounds and camera angles for each film. The image has a frame rate of 20 frames per second (fps), and each frame has dimensions of 490*330 pixels. While the current data set has a limited number of types, it presents certain issues, such as the potential for the moving target and background to get readily intertwined within the movie. The published dance action data set in the scientific field of dance action analysis is of exceptional quality, making it suitable for evaluating the efficacy of the method suggested in this study.

The FolkDance dance data set is a dance dataset that has been personally generated by the laboratory. Vicon employs motion capture equipment to gather professional dance motion films. Throughout the entire process of producing the data set, as outlined in the data set production plan, the ultimate objective is to create four distinct groups of folk dance motions. A comprehensive collection of 84 dancing videos was documented, with the background and camera angle in each movie being predetermined and unalterable. The video images have a frame rate of 20 frames per second (fps), and each frame has dimensions of 480*350 pixels.

4.2 Experimental Setup

The experimental environment is as follows:

CPU : Intel(R) Core(TM) i5-3230M @ 2.6GHZ.

Memory: 4G.

Operating system: Windows, 32-bit.

Development tools: Visual Studio 2012, OpenCV 2.4.8. OpenCV is a computer vision library written in C or C++ language and has realized many common graphics and computer vision algorithms.

4.3 Experimental Result Analysis

4.3.1 Recognition Effect of Different Features Extracted from Different Human Body Regions

This paper extracts RGB features, SIFT features, Flow features, and recognition rates after SIFT and Flow features from different parts of the upper body, lower body, and whole body regions. Different human body regions extract different feature recognition effects, as shown in Figure 4.

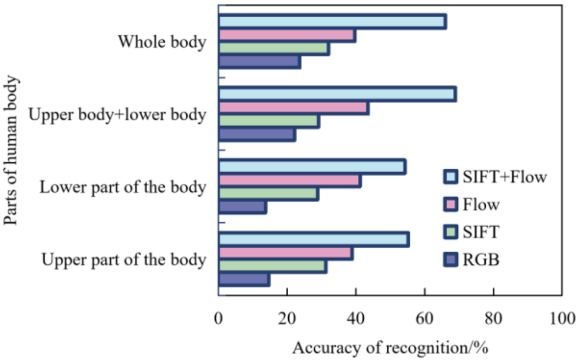

Figure 4: Recognition effect of features extracted from different human body regions.

An examination of recognition rates obtained through several ways indicates that the rate of recognizing actions is pretty consistent when using either the upper body or lower body for presenting features. Nevertheless, the lower body has a slightly elevated rate of recognition. The utilization of a mixture of human body regions yields the highest percentage of recognition. In contrast, the rate at which full body regions are directly recognized is lower when compared to the mixture of human body regions.

The rationale behind this finding is that the frequency of dances incorporating both upper-body and lower-body motions is rather low, but dances containing lower-body movements are more commonly observed. The reciprocal contact between the upper body area and the lower body region is apparent within the context of whole-body movement. Consequently, it is typical to employ both lower-body and upper-body movements when assigning dance-related movements. Hence, the utilization of several training sessions holds the capacity to enhance the accuracy of recognition.

4.3.2 Analysis of Key Frame Extraction Method

According to the keyframe extraction method, the key frame of the dance video is extracted. Compared with the method in literature [22], in the experiment, because the dance action of the dance video and its accompaniment music set off against each other, the corresponding entropy value will change when the accompaniment music changes with time so that it can effectively distinguish the heavy beat from the beat with a large amount of exercise.

In the experiment, the threshold selection is to select a keyframe set with a high evaluation coefficient through continuous iteration according to the evaluation criteria. Figures 5 and 6 show the comparison between the key frame extraction results of reference [22] and the key frame extraction results of this method.

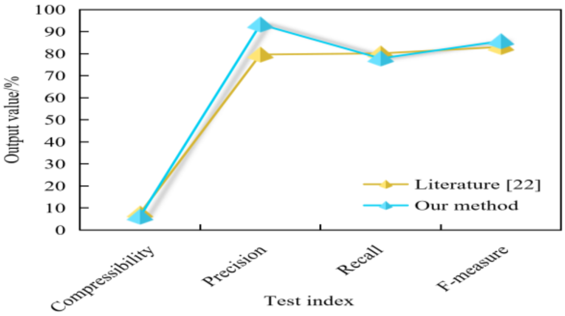

Figure 5: Comparison of essential frame extraction methods (danceDB data set).

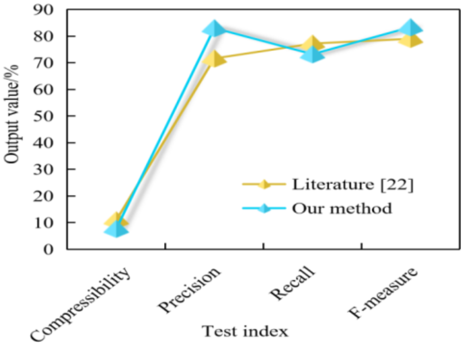

Figure 6: Comparison of essential frame extraction methods (folk dance data set).

Figures 5 and 6 clearly demonstrate that this method extracts fewer important frames compared to reference [22], while achieving a higher F-measure. The analysis of the DanceDB data set reveals a compression ratio of 6.1% and an F-measure score of 85.71%. The results obtained from conducting experiments on a different dataset indicate that the compression ratio achieves a value of 7.7%, while the average F-measure value reaches 83.37%. These findings suggest that the suggested strategy exhibits more robustness.

In terms of video clip retrieval, this study primarily focuses on searching video clips based on keyframes. In terms of key frame extraction, the equidistant shot segmentation method and histogram equalization are employed to select keyframes. Additionally, it searches video clips using color correlation graphs, in contrast to the approaches described in literature [23], literature [24], and literature [25]. This approach has limited efficacy in key frame extraction for dance videos, hence posing challenges in precisely segmenting frames within such recordings. A minimal disparity was observed in the color correlation diagram. Figures 7 and 8 display the comparative findings between reference [37] and this technique.

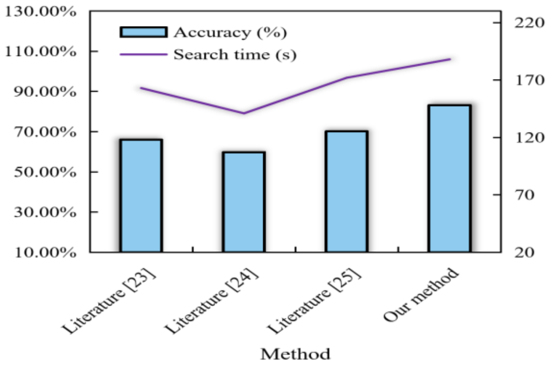

Figure 7: Comparison results of video segment retrieval (danceDB data set).

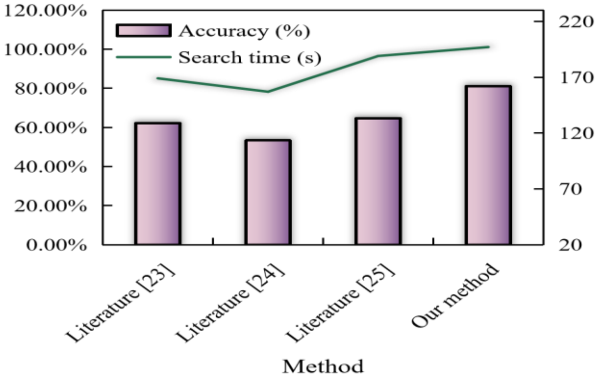

Figure 8: Comparison results of video segment retrieval (folk dance data set).

Figures 7 and 8 demonstrate that the video clip retrieval method described in this research outperforms existing publications in terms of accuracy. However, there is a noticeable disparity in retrieval time. This demonstrates the precision of key frame extraction, hence enhancing the accuracy of later video retrieval. Moreover, there are more or fewer video clips in different categories of dance videos, which are similar to certain video content in different categories but express different themes or emotions. That is to say, the same dance movements and different combinations will produce a new dance video.

4.3.3 Performance Comparison of Different Features

Firstly, feature vectors of TDH, HOG, HOF, and MBH of all training videos are extracted to obtain feature vector sets of all training videos about the four descriptors. Then, the K-means clustering method is used to cluster on each feature vector set to receive multiple clustering centers, constituting a visual dictionary of each descriptor. Like most literatures, we set the number of clustering centers of each feature type as 4000; that is, the dictionary size of each descriptor is 4000.

Fig. 9 is the recognition results of various single feature descriptors and combined feature descriptors on each data set, reflecting the contribution of different features to the recognition results.

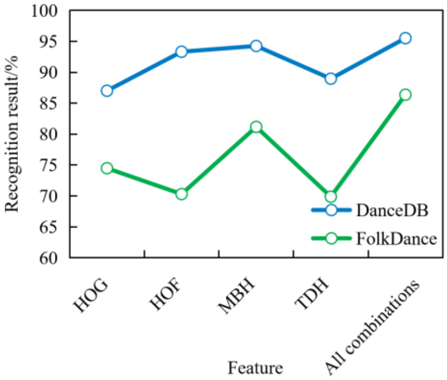

Figure 9: Performance comparison of different features.

In Fig. 9, the static appearance features described by HOG perform better on the DanceDB data set with a simple background than on a data set with a complex background, such as the FolkDance data set. HOF describes speed information, and it can be expected that when it is used in motion recognition, it can get more distinguishable motion information than HOG.

MBH, as an improved feature of HOF, obviously performs best on each data set, especially in the case of camera motion, such as the FolkDance data set, which is about 10% higher than HOF, while for fixed camera and less camera motion, such as DanceDB data set, the discrimination between HOF and HoF is not significantly improved.

TDH describes the characteristics of the track itself well, and it is easy to extract the correct movement track on the data set with a simple background and simple action, such as KTH. Only the distribution characteristics of the track itself can provide even more distinguishable information than HOG, and the extraction method is relatively simple.

The features of the combination can improve the recognition accuracy relatively, especially in complex backgrounds and actions, such as the FolkDance data set, by about 5%. In contrast, the accuracy of recognition of simple backgrounds and actions is relatively less improved. This proves the effectiveness of the feature extraction method based on multi-feature fusion.

Then, this paper compares the influence of the improved word bag model and the general word bag model on recognition accuracy, as shown in Table 10.

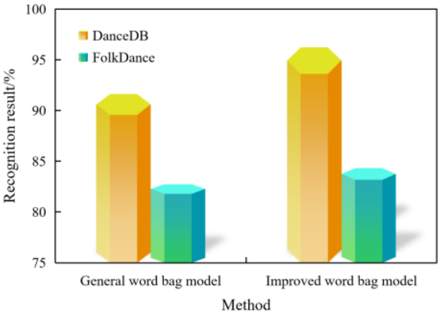

Figure 10: Influence of improved word bag model on recognition.

Figure 10 shows that on each data set, using the improved word bag model to increase the spatiotemporal distribution among features has improved the recognition accuracy compared with that without spatiotemporal distribution. This proves that using the multi-channel word bag model in this paper can improve the recognition accuracy and partially compensate for the word bag model's deficiency.

4.3.4 Accuracy Analysis of Key Points in the Model

In this paper, the new algorithm's key points are evaluated using the index of percentage of correct key points (PCK), and the evaluation results are shown in Figure 11.

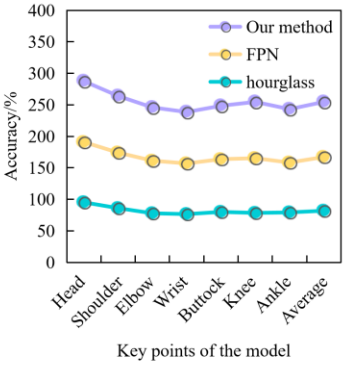

Figure 11: Evaluation results.

Figure 11 clearly demonstrates that the PCK values of all three methods exceed 95% in accurately predicting the head. The hourglass-based algorithm exhibits a minimum PCK value of 95.3% and a maximum PCK value of 96.7%. This finding demonstrates that the utilization of this approach can enhance the regression of basic important variables to a certain degree.

The prediction of the shoulder key points also demonstrates the algorithm's enhancement of the basic key points. In contrast to the hourglass-based approach, the shoulder key points of this method exhibit a 1.6% rise in the PCK value. The hourglass-based algorithm achieves PCK values of 76.6% and 79.3% for the crucial sites of the wrist and ankle, respectively. The PCK values of the FPN-based method exhibit a minor increase of 80.4% for the wrist while experiencing a slight drop of 79.1% for the ankle.

This technique incorporates a multi-feature fusion module that enhances feature extraction and combines several features, resulting in a more comprehensive extraction of feature semantics. Conversely, the ankle key points experienced a mere 0.8% increase, suggesting that identifying the concealed key points in 2D images is exceedingly challenging.

5 CONCLUSIONS

The study presents a dance movement identification algorithm that utilizes multi-feature fusion to enhance the system's ability to identify intricate dancers' movements accurately. In the context of motion representation, the concept of multi-feature fusion is employed to develop a comprehensive feature extraction method that enhances the description of moving objects in video. To mitigate the impact of inaccurate posture estimation on action detection, human posture is employed to identify image regions and extract information from each region. This work employs a multi-feature fusion approach to extract directional gradient histogram features, histogram features, and audio features for the purpose of dance action recognition research. The analysis of the dance video library reveals that the algorithm significantly enhances the model's precision in forecasting the crucial elements of the obstructed and dynamic dance action skeleton. The design of a neural network architecture capable of handling multimodal inputs is critical to this research. Experimenting with different fusion strategies, such as early fusion, late fusion, or attention mechanisms, can provide insights into the most effective ways to combine Collaborative CAD.

Xiaoqiang Yang, https://orcid.org/0009-0005-6662-9092

REFERENCES

[1] Chen, G.; Ge, K.: A Fusion Recognition Method Based on Multi-feature Hidden Markov Model for Dynamic Hand Gesture, Computational Intelligence and Neuroscience, 2020(12), 2020, 8871605. https://doi.org/10.1155/2020/8871605

[2] Gong, F.; Li, C.; Gong, W. et al. A Real-Time Fire Detection Method from Video with Multi-feature Fusion, Computational Intelligence and Neuroscience, 2019(1), 2019, 1-17. https://doi.org/10.1155/2019/1939171

[3] Gu, Y.; Liu, M.; Sheng, W.: et al. Sensor Fusion Based Manipulative Action Recognition, Autonomous Robots, 45(3), 2021, 1-13. https://doi.org/10.1007/s10514-020-09943-8

[4] Ho, N. H.; Yang, H. J.; Kim, S. H.: et al. Multimodal Approach of Speech Emotion Recognition Using Multi-Level Multi-Head Fusion Attention-based Recurrent Neural Network, IEEE Access, PP(99), 2020, 1-1. https://doi.org/10.1109/ACCESS.2020.2984368

[5] Jia, H.; Qu, M.; Wang, G.: et al. Dough-Stage Maize (Zea mays L.) Ear Recognition Based on Multiscale Hierarchical Features and Multi-feature Fusion, Mathematical Problems in Engineering, 2020(2), 2020, 1-14. https://doi.org/10.1155/2020/9825472

[6] Jing, T.; Xiaoqiang, Y.: Dance Movement Recognition Based on Modified GMM-Based Motion Target Detection Algorithm, Security and Communication Networks, 2022(6023784), 2022, 12. https://doi.org/10.1155/2022/6023784

[7] Liang, C.; Liu, D.; Qi, L.: et al. Multimodal Human Action Recognition with Sub-Action Exploiting and Class-Privacy Preserved Collaborative Representation Learning, IEEE Access, 8, 2020, 39920-39933. https://doi.org/10.1109/ACCESS.2020.2976496

[8] Liu, W.; Li, S.: Human Motion Target Recognition Using Convolutional Neural Network and Global Constraint Block Matching, IEEE Access, PP(99), 2020, 1-1. https://doi.org/10.1109/ACCESS.2020.2986305

[9] Luo, D.: Modeling and Simulation of Athlete's Error Motion Recognition Based on Computer Vision, Complexity, 2021(1), 2021, 1-10. https://doi.org/10.1155/2021/5513957

[10] Luo, Jin-m.; Luo, J.; LI, Yan-m.: et al. Face Recognition Algorithm Based on Multi-feature Fusion Convolution Neural Network, aeronautical computing technique, 049(003), 2019, 40-45.

[11] Mao, L.; Wang, N.; Wang, L.: et al. Classroom Micro-Expression Recognition Algorithms Based on Multi-Feature Fusion, IEEE Access, PP(99), 2019, 1-1. https://doi.org/10.1109/ACCESS.2019.2917230

[12] Ni, X.; Wang, H.; Meng, F.: et al. LPI Radar Waveform Recognition Based on Multi-resolution Deep Feature Fusion, IEEE Access, PP(99), 2021, 1-1. https://doi.org/10.1109/ACCESS.2021.3058305

[13] Rui, S.; Qiheng, H.; Weiming, L.: et al. Video-Based Person Re-Identification via Combined Multi-Level Deep Feature Representation and Ordered Weighted Distance Fusion, Acta Optica Sinica, 39(9), 2019, 0915006. https://doi.org/10.3788/AOS201939.0915006

[14] Sharif, M.; Khan, M. A.; Tahir, M. Z.: et al. A Machine Learning Method with Threshold Based Parallel Feature Fusion and Feature Selection for Automated Gait Recognition, Journal of Organizational and End User Computing, 32(2), 2020, 67-92. https://doi.org/10.4018/JOEUC.2020040104

[15] Tang, F.; Lu, X.; Zhang, X. et al. Deep Feature Tracking Based on Interactive Multiple Model, Neuro Computing, 333(MAR.14), 2019, 29-40. https://doi.org/10.1016/j.neucom.2018.12.035

[16] Tong, M.; Zhao, M.; Chen, Y.: et al. D~3-LND: A Two-Stream Framework with Discriminant Deep Descriptor, Linear CMDT and Nonlinear KCMDT Descriptors for Action Recognition, Neurocomputing, 325(JAN.24), 2019, 90-100. https://doi.org/10.1016/j.neucom.2018.09.086

[17] Tu, M.: Gesture Detection and Recognition Based on Pyramid Frequency Feature Fusion Module and Multiscale Attention in Human-Computer Interaction, Mathematical Problems in Engineering, 2021(7), 2021, 1-10. https://doi.org/10.1155/2021/6043152

[18] Wang, H.; Li, J.: Human Action Recognition Algorithm Based on Multi-feature Map Fusion, IEEE Access, PP(99), 2020, 1-1. https://doi.org/10.1109/ACCESS.2020.3017076

[19] Wang, P.: Research on Sports Training Action Recognition Based on Deep Learning, Scientific Programming, 2021(7), 2021, 1-8. https://doi.org/10.1155/2021/3396878

[20] Wen, M.; Wang, Y.: Multimodal Sensor Motion Intention Recognition Based on Three-Dimensional Convolutional Neural Network Algorithm, Computational Intelligence and Neuroscience, 2021(1), 2021, 1-11. https://doi.org/10.1155/2021/5690868

[21] Weon, I. S.; Lee, S. G.; Ryu, J. K.: Object Recognition Based Interpolation with 3D LIDAR and Vision for Autonomous Driving of an Intelligent Vehicle, IEEE Access, PP(99), 2020, 1-1. https://doi.org/10.1109/ACCESS.2020.2982681

[22] Yu, J.; Gao, H.; Yang, W.: et al. A Discriminative Deep Model with Feature Fusion and Temporal Attention for Human Action Recognition, IEEE Access, PP(99), 2020, 1-1.

[23] Zhai, X.: Dance Movement Recognition Based on Feature Expression and Attribute Mining, Complexity, 2021(21), 2021, 1-12. https://doi.org/10.1155/2021/9935900

[24] Zhang, D.; He, L.; Tu, Z.; et al.: Learning Motion Representation for Real-Time Spatio-Temporal Action Localization, Pattern Recognition, 103(1), 2020, 107312. https://doi.org/10.1016/j.patcog.2020.107312

[25] Zhu, X.; Yan, W.; Chen, D.: et al. Research on Gesture Recognition Based on Improved GBMR Segmentation and Multiple Feature Fusion, Journal of Computer and Communications, 07(7), 2019, 95-104. https://doi.org/10.4236/jcc.2019.77010