Publication Date: 30 June 2024
Archs Sci. (2024) Volume 74, Issue 3 Pages 190-197, Paper ID 2024330.
https://doi.org/10.62227/as/74330

Research on Modern Dance Performance Forms and Dance Movement Characteristics Based on Multi-scale Feature Fusion

Duoduo Wang[1,*], Yanting Liu[2] and Molin Li[2]
[1]School of Art and Design, Guilin University of Electronic Technology, Guilin, Guangxi, 541004, China.
[1]Guilin Street Dance Association, Guilin Dance Association, Guilin, Guangxi, 541004, China.
Corresponding authors: Duoduo Wang (e-mail: 18607736366@163.com).

Abstract Due to the complexity of dance performance forms, dance movement recognition is very difficult, and related research has received much attention. In this paper, a dance action feature recognition method based on multi-scale feature fusion is designed around modern dance performance forms and dance action features. The method carries out dance action feature extraction on the well-framed image of dance video by calculating the image entropy of optical flow map, combined with the audio feature sequence of audio feature extraction. Then the multi-core learning method is used to realize the multi-feature fusion of the dance action to complete the extraction and recognition of dance action features. Through the performance analysis of the model, the recognition accuracy of this paper's model on the NTU RGB-D 60 and NTU RGB-D 120 datasets stays above 90% and 85%, with an overall decrease of only 4.95% in the latter, which is the best performance among the compared algorithms. Experiments show that the proposed method in this paper has better dance movement recognition effect and generalization ability.

Index Terms multi-scale feature fusion, feature extraction, multi-core learning, dance movement feature recognition, performance analysis

I. Introduction

With the continuous improvement of social and economic level and the progress of science and technology, people constantly have a new pursuit of life. Relying on real life, modern dance can fully reflect people's thoughts and psychological state, and also reflect the penetration of different cultures [1]–[3]. Modern dance is different from folk dance, it can quickly integrate and break the barriers between different dance styles [4]. Even if the two are different, they can seek common ground while reserving differences, in order to promote the innovative development of different dance forms [5].

Modern dance as the 20th century originated in the western dance art, the development of its value is not only limited to a pure art form [6], [7], it accompanies the development of western modern culture growth, rebellion against tradition at the same time, it is a kind of artistic and cultural trend of innovation [8]. Modern dance, whether it is performing arts, spatial structure or the pursuit of artistic aesthetics, are in the seemingly reverse road to seek the return of life's most genuine, abstract, grotesque, individuality is synonymous with, it is precisely because modern dancers are always trying to break the traditional dance of the purity of the rules and regulations, and the courage to challenge the new things, the modern dance presents the difference between the traditional dance art style [9], [10].

With the continuous development of modern dance, it creates various movement techniques. Although dance is an art that mainly expresses thoughts and emotions with the help of rich body movements, the real emotions of modern people are still the final landing point of the ideological nature of movements. Therefore, the rich emotion gives the movement a strong impact and is very shocking [11]–[13]. As a newly emerging modern dance in the 20th century, compared with traditional dance, it is more unique, innovative and metaphorical in body language expression. As we all know, dance is an abstract body language, and its artistic performance has metaphorical characteristics [14].

Modern dance in today's cultural and artistic fields has been a significant position, as an art form, with a unique artistic posture for people to present "deviant" aesthetic enjoyment, as a cultural trend, it revolutionizes the development of art at the same time, but also triggered people to think more [15], [16].

The study, after the frame-splitting process of the dance video, carries out optical flow calculation on each image with good frame-splitting to obtain the motion characteristics of the dancer in the image, and then carries out image entropy calculation to obtain the entropy value sequence of the optical flow map. It is fused with the audio feature sequence obtained by audio feature extraction to realize the extraction and representation of dance movement features. On this basis, multi-feature fusion is used for dance action recognition, and considering the existence of heterogeneous feature fusion problems, multi-core learning is used to organically fuse multiple types of features for dance action recognition. NTU RGB-D 60 and NTU RGB-D 120 are selected as experimental datasets for training and testing to analyze the loss value curve and accuracy curve of the model. A series of comparison experiments are conducted on the two division criteria of NTU RGB-D 60 and NTU RGB-D 120 datasets, and solo and group dance situations are simulated to measure the accuracy of the model's dance action feature recognition in the two cases, and to test the performance of this paper's model in recognizing dance actions.

II. Dance movement feature recognition method based on multi-scale feature fusion

In this paper, an effective dance movement feature recognition method is proposed for dance video dataset, which firstly performs the extraction and processing of dance movement features, and then adopts the multi-core learning approach to organically fuse the three types of features for dance movement recognition research.

A. Dance movement feature extraction and description

In action recognition research, the first thing to be done is usually feature extraction, which refers to the extraction of feature information from the action data set to describe the target action in the video, and is an essential step for action recognition research. This chapter describes the process of extracting key frames from a dance video, using the optical flow method to extract the motion features of the dance action in the dance video, and then using image entropy to calculate the entropy value of each optical flow image.

1) Feature extraction method based on motion features

Traditional optical flow algorithms are limited at some moments, and although they can track the target in the image, the optical flow computation is not very effective for discontinuities in the movement, occlusion phenomena, and large displacements.

In this paper, the LDOF method will be used to obtain the motion characteristics of the dance video. This is a good way for dance videos to accurately calculate every change of dance movements without missing the movement changes of those small body parts (hands, feet, etc.). The optical flow calculation method is shown in (1) to (6):

$$\begin{aligned} E\left(w\right) \quad = \quad & E_{cobbr}\left(w\right)+\gamma E_{grad}\left(w\right)+\alpha E_{smooblh}\left(w\right) \\ & +\beta E_{makk}\left(w,w_{l}\right)+E_{dect}\left(w_{l}\right), \end{aligned} \tag{1}$$

$$E_{codar}\left(w\right)=\int_{\Omega}\Psi\left(\left|I_{2}\left(x+w\left(x\right)\right)-I_{1}\left(x\right)\right|^{2}\right)dx, \tag{2}$$

$$E_{grad}\left(w\right)=\int_{\Omega}\Psi\left(\left|\nabla I_{2}\left(x+w\left(x\right)\right)-\nabla I_{1}\left(x\right)\right|^{2}\right)dx, \tag{3}$$

$$E_{snwowh}\left(w\right)=\int_{\Omega}\Psi\left(\left|\nabla\mu\left(x\right)\right|^{2}+\left|\nabla v\left(x\right)\right|^{2}\right)dx, \tag{4}$$

$$E_{maskh}\left(w\right)=\int\delta(x)\rho(x)\Psi\left(\left|w\left(x\right)-w_{1}\left(x\right)\right|^{2}\right)dx, \tag{5}$$

$$E_{desc}\left(w_{1}\right)=\int\delta(x)\left|f_{2}\left(x+w_{1}\left(x\right)\right)-f_{1}\left(x\right)\right|^{2}dx, \tag{6}$$

where, α, β and γ are adjustable weighting parameters and $E_{color}\left(w\right)$ is an assumption of luminance invariance for both color and grayscale images. The effect of illumination is unavoidable and therefore, to minimize the effect of illumination, a gradient constraint $E_{grad}\left(w\right)$ is added to this. which is then smoothed by $E_{smach}\left(w\right)$. The last two items are constructing descriptor matching and minimizing it by multivariate modeling and optimization.

2) Image entropy calculation based on optical flow map

Entropy calculation on image frames only takes into account the features of the image, but for dance videos, the change of movement is the key. Therefore, the method used in this paper is to calculate the entropy value of each optical flow image by taking the optical flow image derived from the above stage as the input of the stage. In this paper, the entropy value is calculated on grayscale images, although there is a loss of information compared to color images, the information required for entropy calculation has not changed, and there is an increase in the speed of operation, and it is also easy to store. The entropy value of the current optical flow map is calculated by following the chronological order. As shown in (7):

$$E\_img=-\sum_{k}^{m}p_{k}\log_{2}p_{k}, \tag{7}$$

where p_k represents the proportion of pixels in the image with a gray value of k, m is for the gray level (0-255), and E_img is the entropy value. The greater the amount of information contained in the image, the greater the entropy value, which means that there is a greater variation in the dance movements between frames T and $T+1$ of the current image.

3) Fusion of audio features with entropy sequences

The speed of dance movements has correlation with music. In this paper, the envelope feature curve of the music is fused with the entropy sequence to obtain an entropy sequence related to the music. The accompanying music is extracted from the video, the accompanying music in WAV format is read, and then the envelope features and energy features of the music are extracted to prepare for the subsequent feature fusion.

The extraction of the energy features of the audio starts with the frame splitting process, where the audio $x\left(j\right)$ is subjected to the windowing frame splitting process to get the knd frame of the audio. The audio signal is present in y, the length of y is N, the sampling rate is fs, the length of each take is $wlen$, the displacement of the two frames before and after is dis, and the overlap between the two frames is $olap = wlen - dis$. Therefore, for the audio signal of length N, the equation of the frame-splitting is shown in (8):

$$fs = \left(N - olap\right)/dis = \left(N - wlen\right)/dis + 1. \tag{8}$$

The average amplitude of the audio, i.e., the energy signature of the audio, is then calculated as shown in (9) and (10). In this way the audio feature sequence and the entropy value sequence can be aligned. Finally, feature fusion is performed by the product operation of the audio feature sequence and the entropy value sequence to obtain a music-related entropy value sequence:

$$y_k\left(j\right) = win\left(j\right) \cdot x\left(\left(k-1\right) \cdot dis + j\right), \tag{9}$$

where $1 \leq j \leq L, 1 \leq k \leq f$.

$$M\left(k\right) = \sum_{j=0}^{L-1} \left|y_k\left(j\right)\right|, 1 \leq k \leq f, \tag{10}$$

where $win\left(j\right)$ is the window function, $y_k\left(j\right)$ is the value of a frame, L is the frame length, dis is the frame shift length, and f is the total number of frames after splitting. In (10), $M\left(k\right)$ represents the characterization of the energy magnitude of a frame of audio.

4) Keyframe Extraction for Dance Videos

The entropy value sequence after feature fusion is calculated, and the comparison operation is not based on the entropy value difference between two neighboring frames as the basis for finding the key frames, but the following formula is used to operate, and the obtained value is compared with the threshold value, and finally, the frame that is larger than the threshold value is selected as the key frame. For the ever-changing dance movements in a dance video, this can effectively extract the representative frame and make the keyframe set less redundant. As shown in (11):

$$V = \frac{\left|H_{cwrmt} - H_{key}\right|}{H_{key}}, \tag{11}$$

where H_{curnnt} denotes the entropy value of the current frame and H_{key} denotes the entropy value of the current keyframe.

B. Multi-feature Fusion and Recognition of Dance Movements

Based on the previously extracted features, a multi-feature fusion method for dance movement recognition is proposed, which applies multi-core learning to the study of multi-feature fusion methods.

1) Multi-core learning methods

Multi-core learning belongs to the category of kernel methods, and the main principle is to use the combination of multiple kernel functions instead of a single kernel function. In multicore learning, most of the multicore models built are linear combinations of multiple kernel functions to form a new kernel function.

The representation of linearly combined kernel functions is shown below:

$$k\left(x,z\right) = \sum_{j=1}^{M} \beta_j k_j\left(x,z\right), \beta_j \geq 0, \sum_{j=1}^{M} \beta_j = 1. \tag{12}$$

In (12) M denotes the number of kernel functions, $k_j\left(x,z\right)$ is the kernel function, and β_j is the corresponding weight of the kernel function. $k\left(x,z\right)$ is determined by the correlation of the features, and $\beta_j \geq 0$ is set in order to make $k\left(x,z\right)$ comply with Mercer's theorem. Therefore, the core problem of multikernel learning is to learn both the optimal kernel function parameters and the corresponding weights of the kernel functions during the optimization process. β Multikernel learning-based methods usually obtain better results than single-kernel learning-based methods, but the problem faced by multikernel learning is that its time and space complexity are too large. space complexity is too large. In recent years, researchers have proposed many solutions to the optimization problem of multikernel learning. The classical ones are based on semipositive programming, second-order conical programming and alternating optimization.

The method based on alternating optimization is divided into two main steps: the first step starts with a single kernel function learning, which first fixes the kernel function weights and solves the parameters of the support vector machine classifier. The second step is to fix the parameters of the support vector machine classifier and calculate the new kernel function weights. The above two steps are executed alternately until convergence. Based on the alternating optimization strategy, the multikernel learning problem is transformed into a semi-infinite linear programming problem (SILP):

$$\begin{cases} \text{max}_{d_a} \text{ min} \sum\limits_{m=1}^{q} d_m \left[\frac{1}{2}\sum\limits_i \alpha_i y_i \phi_m\left(x_i\right) - \sum\limits_i \alpha_i\right] \\ \text{s.t. } \sum\limits_i \alpha_i y_i = 0, C \geq \alpha_i \geq 0 \quad \forall\, i \\ \sum\limits_m d_m = 1, d_m \geq 0 \,\forall\, m \end{cases} \tag{13}$$

(13) can be solved using alternating optimization, where the authors update the kernel function weights using a tangent plane approach, where the lower boundary of the tangent plane for each kernel function weight is computed with the parameters of the support vector machine classifier derived from the previous iteration. The SILP-based algorithm optimizes the kernel function linear combination weights and the single kernel SVM parameters by using a two-layer loop, respectively, before updating the kernel function weights. Although the efficiency of solving the SILP problem is significantly improved, the algorithm is not stable enough and therefore also difficult to be applied to large-scale solving.

Based on the idea of alternating optimization strategy, Simple MKL algorithm uses gradient descent to update the kernel weights. At the same time, the objective function of multikernel learning is rewritten:

$$\begin{cases} \min\limits_{f_n,b,\xi,d} \frac{1}{2}\sum\limits_{m}\frac{1}{d_m}\left\|f_m\right\|_{H_a}^2 + C\sum\limits_{i}\xi_i \\ \text{s.t}y_i\sum\limits_{m} f_m\left(x_i\right) + y_i b \geq 1-\xi_i\ \forall \\ i\xi_i \geq 0\ \forall\ i, \sum\limits_{m} d_m = 1, d_m \geq 0\ \forall\ m \end{cases} \tag{14}$$

The f_m in (14) denotes the mapping function which will map the data to the corresponding regenerative kernel Hilbert space. Meanwhile d_m constrains the paradigm of f_m. The smaller the value of d_m, the smoother f_m. The dyadic form of this objective function is as follows:

$$\begin{cases} \min_{d_a}\max_{\alpha} -\frac{1}{2}\sum\limits_{i,j}\alpha_i\alpha_j y_i y_j\sum\limits_{m} d_m K_m\left(x_i,x_j\right) + \sum\limits_{i}\alpha_i \\ \text{s.t}\ \sum\limits_{i}\alpha_i y_i = 0\ \forall\ i \\ C \geq \alpha_i \geq 0\ \forall\ i \\ \sum\limits_{m} d_m = 1, d_m \geq 0\ \forall\ m \end{cases} \tag{15}$$

(15) differs from the semi-infinite linear programming method in that it utilizes the smoothing property of the objective function and uses a gradient descent method to update the kernel function weights. The kernel function weights calculation is obtained by the following equation:

$$D_m = \begin{cases} 0 & \text{if } d_m = 0 \text{ and } \frac{\partial J}{\partial d_m} - \frac{\partial J}{\partial d_p} > 0 \\ -\frac{\partial J}{\partial d_m} + \frac{\partial J}{\partial d_\mu} & \text{if } d_m > 0 \text{ and } m \neq \mu \\ \sum\limits_{g\neq\mu,d_i>0}\left(\frac{\partial J}{\partial d_v} - \frac{\partial J}{\partial d_\mu}\right) & \text{if } m = \mu \end{cases} \tag{16}$$

In (16), D_m denotes the direction of gradient descent. When updating the kernel function weights each time, the use of linear search can quickly obtain the optimal kernel function weights through a suitable search step size.

2) Dance movement recognition based on feature fusion

Reference to the parallel branch structure of the Shufflenet network, using the different size of the convolution core, capture the fine-grained characteristics of the image and the context information of the department, and the multi-scale characteristics of the different scale characteristics of the different scale of the different scales are improved by the feature splicing and the channel shuffle, and the classification performance of the different scales is improved.

Considering the limited ability of each class of features to distinguish dance movements individually, this paper adopts a linear weighted combination of directional gradient histogram features, optical flow directional histogram features, and audio features in a multicore learning approach to fuse them in order to realize the mutual complementation of multiple classes of features to improve the recognition ability of the classifier.

Assume that there are p dance movement $x_1, x_2, \ldots, x_p$ and category $y_1, y_2, \ldots, y_p$ in the dance dataset. meanwhile, the G kernel functions corresponding to the HOG features are defined as $k_g\left(x_i,x_j\right)$, $g = 1,2,\ldots,G$, the F kernel functions corresponding to the HOF features are defined as $k_f\left(x_i,x_j\right), f = 1,2,\ldots,F$, and the M kernel functions corresponding to the audio signature features are defined as $k_m\left(x_i,x_j\right)$, $m = 1,2,\ldots,M$. In this paper, linear combinations of the kernel functions incorporating the three features mentioned above can be expressed by (17):

$$k\left(x_i,x_j\right) = \sum_{g=1}^{G}\beta_g k_g\left(x_i,x_j\right) + \sum_{f=1}^{F}\beta_f k_f\left(x_i,x_j\right) + \sum_{m=1}^{M}\beta_m k_m\left(x_i,x_j\right). \tag{17}$$

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

In the support vector machine based on multikernel learning, the task of the training phase of the multikernel learning model is to learn to solve the weights of each kernel function β and the parameters of the support vector machine classifier itself α and b. Based on the idea of Simple MKL algorithm in the previous section, the objective function of the algorithm in this paper is defined as shown in (18):

$$\begin{cases} \min\limits_{\beta_g,\beta_f,\beta_M,\alpha,b} J = \frac{1}{2}\sum\limits_{g=1}^{G}\beta_g\alpha^T K_g\alpha + \frac{1}{2}\sum\limits_{f=1}^{F}\beta_f\alpha^T K_f\alpha \\ \quad +\frac{1}{2}\sum\limits_{m=1}^{M}\beta_m\alpha^T K_m\alpha + C\sum\limits_{i}\xi_i \\ \text{s.t}\ y_i\left[\sum\limits_{g=1}^{G}\beta_g K_g\left(x_i\right) + \sum\limits_{f=1}^{F}\beta_f K_f\left(x_i\right) + \sum\limits_{m=1}^{M}\beta_m K_m\left(x_i\right)\right]\alpha \\ \quad +y_i b \geq 1-\xi_i\ \forall\ i, \\ \xi_i \geq 0\ \forall\ i, \sum\limits_{g=1}^{G}\beta_g + \sum\limits_{f=1}^{F}\beta_f + \sum\limits_{m=1}^{M}\beta_m = 1 \end{cases} \tag{18}$$

Eq. (18) in $K_g\left(x_i\right) = \left[k_g\left(x_i,x_1\right),\ldots,k_g\left(x_i,x_p\right)\right]$, $K_f\left(x_i\right) = \left[k_f\left(x_i,x_1\right),\ldots,k_f\left(x_i,x_p\right)\right]$ and $K_m\left(x_i\right) = \left[k_m\left(x_i,x_1\right),\ldots,k_m\left(x_i,x_p\right)\right]$. According to the idea of Simple MKL algorithm, a gradient descent algorithm is used to minimize the objective function and thus learn to solve the optimal parameters, the specific process is to compute the classifier parameters α and b in each iteration, given the kernel function weights β, and then compute the new kernel function weights β, given α and b. Therefore, the classification function of a multikernel-based learning support Vector Machine based classification function is shown below:

$$y = F\left(x\right) = \left[\sum_{g=1}^{G}\beta_g K_g\left(x\right) + \sum_{f=1}^{F}\beta_f K_f\left(x\right) + \sum_{m=1}^{M}\beta_m K_m\left(x\right)\right]\alpha + b. \tag{19}$$

In addition, (19) is a two-class classification function and the recognition problem to be solved in this paper is a multi-class classification problem. So it is necessary to convert the two-class classification problem into a multi-class classification problem. The multi-class problem in this paper is converted into a joint binary classification problem, i.e., for each category in the dataset all the dance moves belonging to this category are labeled as positive, and other dance moves are labeled as negative. According to the Simple MKL algorithm, assuming that there is p class of dance moves, p classifiers of the two-class SVM need to be trained. Therefore, the objective function for multi-class classification also becomes as shown in (20) below:

$$J = \sum_{p=1}^{P} J_p \left(\beta_g, \beta_f, \beta_m, \alpha_p, b_p\right). \tag{20}$$

Eq. (20) where J_p is the pnd support vector machine two-class classifier. The output is the dance action of p and the negative sample is the dance action whose category is not p. Finally the algorithm in this paper obtains the movement categories when performing multi-class classification according to the following (21) formula:

$$y = \arg\max_{y_p} F_p\left(x\right). \tag{21}$$

Dance works in the form of dance performance can be divided into solo, double, triple, group dance. The double dance mostly adopts the coherent technique of lifting compound, and the triple dance is richer than the double dance in terms of level, which contains the performance form of the solo dance and joins the performance form of the double dance at the same time, and the performance form of the group dance is more colorful and diversified. The use of dance action feature recognition model to recognize different dance performance forms may have different effects, which will be studied in the following performance analysis.

III. Performance Analysis of Dance Movement Feature Recognition

In order to evaluate the dance movement feature recognition model proposed in this paper, experiments are conducted on two skeleton datasets NTU RGB+D 60 and NTU RGB+D 120, which are widely used for a wide range of applications in dance movement recognition work.

A. Experimental data set

The NTU RGB+D 60 dataset is a large dataset containing 60 different action types widely used in the field of skeleton based action recognition. The dataset contains a total of 56,880 samples, of which 40 categories are daily behaviors, 9 categories are health-related behaviors, and 11 categories are two-person interactive actions. The dataset provides two segmentation criteria: one is Cross-Subject, which divides the training and test sets according to the person ID, and contains a training set of 40,320 segments from 20 subjects, and a validation set of 16,560 segments from 20 other subjects. The second is Cross-View, which divides the training and test sets by camera, and the dataset contains a training set of 37920 clips sampled from two cameras, and a validation set of 18960 clips from another camera.

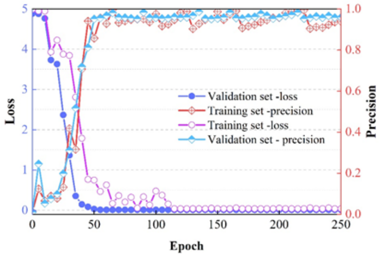

Figure 1: Model’s loss value curve and accuracy curve

The NTU RGB+D 120 dataset extends the NTU RGB+D 60 dataset by providing 114480 video clips executed by 106 subjects from 155 viewpoints. Similar to NTU RGB+D, two division criteria are proposed for the dataset: one is Cross-Subject, in which the dataset contains a training set of 63,026 clips and a validation set of 50,922 clips, respectively. The second is Cross-setup, the dataset in this division criterion contains a training set of 54,471 fragments and a validation set of 59,477 fragments, respectively.

B. Loss and Accuracy Analysis

The performance of the proposed dance movement feature recognition model is verified by training and testing on the dataset. The loss value curve and accuracy curve of the model are shown in Fig. 1. As the number of training times of the model increases, the loss value in the training set fluctuates more in the first 100 cycles, which indicates that the learning is continuously adjusted during the training process with the aim of making the loss of the dance movement feature recognition model converge, and corresponding to the training set, the accuracy rate in the validation set also has the same change. The convergence process of the loss value in the validation set is smoother, and the accuracy in the validation set is more stable, and the accuracy of the dance action feature recognition in the training set and validation set remains above 90% and 95%, respectively.

C. Comparative experiments

Using NTU RGB-D 60 dataset and NTU RGB-D 120 dataset, the results, i.e., recognition accuracy, of the dance movement feature recognition model proposed in this dissertation are demonstrated and compared and analyzed with other movement recognition algorithms proposed so far based on human skeleton data.

1) Experimental analysis of NTU RGB-D 60

The accuracies of different dance movement recognition methods on NTU RGB-D 60 dataset are shown in Table 1. The dance movement feature recognition model proposed in this paper obtains 90.16% and 95.73% recognition accuracies on Cross-subject and Cross-view segmentation criteria of NTU RGB-D 60 dataset. Compared with the Ind-RNN network model based on recurrent neural network, it is 7.36% and 8.15% higher on the two metrics, respectively. Compared with the 3SCNN network model based on convolutional neural network, it is 1.49% and 4.15% higher in two metrics, respectively. The results compared with other action recognition algorithms based on human skeleton data illustrate the superiority of the dance action feature recognition model in this paper.

Comparing the recognition accuracy results of the Cross-subject and Cross-view partitioning standards based on the NTU RGB-D 60 dataset, it can be found that the results of all recognition accuracy on the Cross-view evaluation criteria are higher than those of the Cross-subject evaluation criteria, with the difference between the two reaching 12.02% and 4.33% at the minimum. The reason lies in the NTU RGB-D 60 dataset, which is participated by 40 people of different ages and postures, through 3 Microsoft KinectV2 cameras, set at 3 levels of the same height but with pitch angles of 0°, 45°, -45°There were only 3 changes in the viewing angle, and among the 40 subjects, the tall, short, fat and thin were different, although the same movements were made, there were still many differences in details of the same movements of different subjects, resulting in lower recognition results than those of the Cross-view evaluation criteria.

Then the model's performance is analyzed in depth for the specific classification of the 60 categories of movements under the Cross-subject and Cross-view division criteria, and Figure 2 shows the recognition accuracy of each dance movement category under the NTU RGB-D 60 dataset, with the vertical coordinates of the vertical coordinate being the specific movement numbers and the horizontal coordinates of the horizontal coordinates being the recognition accuracies of the individual movements. The model can accurately recognize most of the movement categories in Cross-subject and Cross-view division criteria, and the recognition accuracy of most movements is above 80%. Focusing on specific action categories, the accuracy rates of categories 11 (70.71%), 12 (65.87%), 29 (69.85%), 30 (75.43%), 41 (67.61%), and 44 (68.89%) are significantly lower than the other categories in the Cross-subject segmentation criterion.The accuracy rates of the Cross-view segmentation criterion for categories 11 (66.62%), 12 (74.53%), 29 (71.04%), 30 (73.12%), 34 (68.89%), and 41 (72.03%) were significantly less accurate than the other categories.

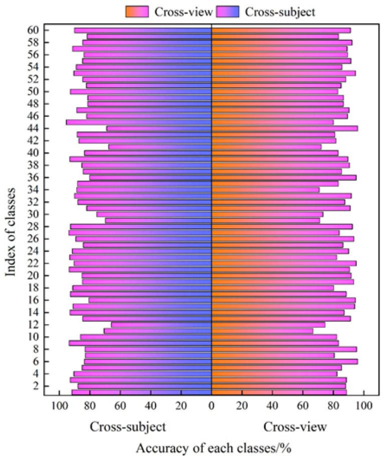

Figure 2: Identification accuracy of various dance movements in the NTU RGB-D 60

2) Experimental analysis of NTU RGB-D 120

In addition to the NTU RGB-D 60 dataset, this paper also experimented and analyzed on its extended version NTU RGB-D 120 dataset, and the accuracy rates of different dance movement recognition methods on the NTU RGB-D 120 dataset are shown in Table 2. Compared with the NTU RGB-D 60 dataset, this dataset has a larger number of movement categories and more complex characters and scenes, so basically, the recognition accuracy of all the mainstream methods on this dataset has decreased significantly. In the NTU RGB-D 120 dataset, the recognition result of the dance action feature recognition model proposed in this paper is 85.21% on the Cross-subject criterion and 86.38% on the Cross-setup criterion, which are both maintained at a better level and show more obvious advantages than other methods. This proves the superiority of the dance movement feature recognition model proposed in this paper in terms of performance.

It can be noticed that some action recognition algorithms provide recognition accuracy on both datasets, which are compared with the method in this paper, the comparison of different action recognition algorithms on NTU RGB-D dataset is shown in Figure 3. The horizontal coordinates in the figure represent the action recognition algorithms and the method used in this paper that provide accuracy on both NTU RGB-D 60 and NTU RGB-D 120 datasets, the vertical coordinate is the action recognition accuracy, and the dashed line represents the difference between the action recognition accuracies of the same action recognition method on the NTU RGB-D 60 dataset and the NTU RGB-D 120 dataset. From the figure, it can be found that the dance action feature recognition model proposed in this paper drops only 4.95% accuracy on the NTU RGB-D 120 dataset, and it also has the relatively highest accuracy on the two datasets, which also reflects to some extent that the dance action feature recognition model in this paper has strong generalization performance.

Figure 4 shows the model's recognition accuracies for each action category on the Cross-subject and Cross-setup delineation criteria for the NTU RGB-D 120 dataset. Firstly, in action numbers 1-60, low accuracies are observed in positions corresponding to the results of NTU RGB-D 60 dataset, which should confirm each other's results. In action numbers 61-120, both Cross-subject and Cross-setup division criteria showed low accuracies in actions 71 (70.26%, 64.38%), 72 (68.5%, 61.96%), 73 (64.51%, 58.34), and 74 (65.15%, 57.69).

D. Recognition analysis under solo and group dance

In order to verify the accuracy of this paper's method in extracting the movement features of modern dance performances, 10 single video samples and 10 multi-person video samples are selected in the dataset to simulate the dance

Method	Cross-subject (%)	Cross-view (%)
Part-aware LSTM	63.21	72.87
TCN	75.68	83.62
Ind-RNN	82.80	87.58
ST-GCN	81.04	88.49
HCN	86.16	90.51
PB-GCN	88.04	92.37
TSRJI	74.59	83.63
AS-GCN	81.75	93.77
3SCNN	88.67	91.58
3s RA-GCN	87.66	94.01
PGCN-TCA	87.48	92.08
Sem-GCN	86.79	93.97
SGN	86.35	95.39
PR-GCN	85.03	90.44
PeGCN	84.56	92.87
Our	**90.16**	**95.73**

Table 1: The accuracy of different dance action recognition methods in the NTU RGB-D 60

Method	Cross-subject (%)	Cross- Setup (%)
Part-aware LSTM	28.65	29.47
TSRJI	69.14	64.53
3SCNN	82.04	82.11
3s RA-GCN	81.93	83.04
Gimme Signals	71.35	72.62
SGN	78.65	80.58
Mix-Dimension	81.32	82.67
ST-TR-agcn	83.06	83.91
Our	**85.21**	**86.38**

Table 2: The accuracy of different dance action recognition methods in the NTU RGB-D 120

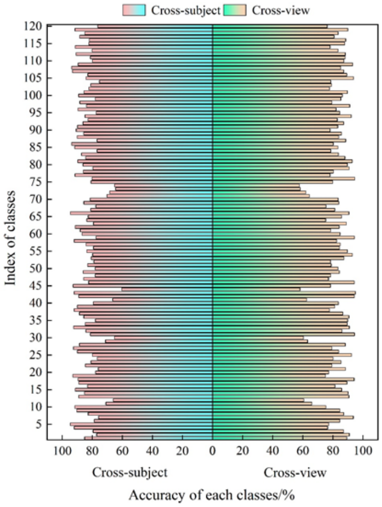

Figure 3: Identification accuracy of various dance movements in the NTU RGB-D 120

movements during solo dance and group dance, respectively. The accuracy of action feature recognition in solo dance and group dance in different datasets is tested, in which the test results of solo dance are shown in Fig. 5 and the accuracy test results of group dance are shown in Fig. 6. In solo dance, the accuracy of dance action feature recognition of the design method in this paper is high, and the overall accuracy is higher than 90%. In group dance, the accuracy of dance movement feature recognition of the method designed in this paper is still high, and the overall accuracy is higher than 85%, but it is slightly lower than the test data in solo dance.

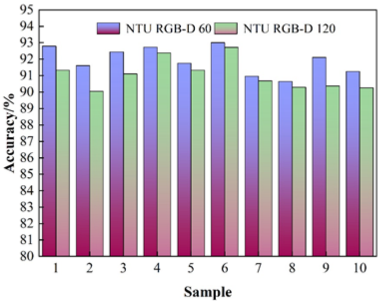

Figure 4: The action characteristics of the solo dance are accurate

IV. Conclusion

In this paper, oriented to the modern dance performance form, we propose a dance action feature recognition method based on multi-scale feature fusion, and conduct a performance study on NTU RGB-D 60 and NTU RGB-D 120 datasets. The research results are as follows:

1) The dance action feature recognition accuracy of this paper's model is above 90% and 95% in training and validation, respectively, and the loss value and accuracy in the training set fluctuate greatly, while the validation set is smoother and more stable. The recognition accuracy of this paper's model stays above 90% and 85% in

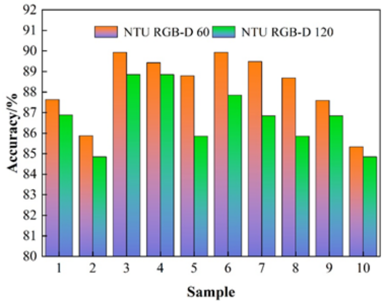

Figure 5: The accuracy of the action characteristics of the group dance

the tests of solo dance and group dance.

2) The action recognition accuracies of this paper's model on the NTU RGB-D 60 dataset are above 90%, and on the NTU RGB-D 120 dataset are above 85%, which are better than other comparative algorithms, and the latter's accuracies have only decreased by 4.95% as a whole, reflecting the better performance of this paper's model in action recognition and stronger generalization ability.
3) Although the research on dance movement recognition in this paper has achieved some results, the recognition rate of the current multi-person dance video movement recognition research is not very high, mainly because the multi-person dance movements are too complex, and the existing methods are still not well suited for dance movement recognition. Therefore, further research and improvement are needed for future dance movement recognition research.

Due to some technical limitations, the following points will be improved:

1) To strengthen the study of feature fusion mechanism and explore more characteristics and integration strategies that improve the accuracy of dance action characteristics.
2) The optimization algorithm is analyzed and improved, especially in the processing of large-scale data sets.
3) The extension model is studied in different styles of dance, different performance forms (such as solo dance, double dance and dance), especially for the identification of dance movements and more detailed experimental verification and technical optimization.

References

[1] Mantellato, M. (2020). (re)playing shakespeare through modern dance: youri vámos's romeo and juliet. *Cahiers Elisabethains, 102*(1), 018476782091379.

[2] Casini, S. (2019, July). Phantasmata of Dance: Time and Memory within Choreographic Constraints. In *Forum for Modern Language Studies* (Vol. 55, No. 3, pp. 325-338). Oxford University Press.

[3] James, C. (2017). Modern Dance Certificate. *The Yale Review, 105*(2), 10-11.

[4] Lucía C. Acevedo. (2021). Patrizia veroli and gianfranco vinay (eds), music-dance: sound and motion in contemporary discourse. *Dance Research, 39*(2), 274-277.

[5] Mollenhauer, J. (2021). 'what's in a name?' taxonomic choices in the field of dance studies. *Dance Research, 39*(1), 89-105.

[6] Porter, A. (2020). Anthony Crickmay and the Art of Dance Photography. *Dance Research, 38*(1), 1-6.

[7] Ritchie, I. (2023). Dance of light and the when an archttect turns to art. *Architectural Design* (5), 93.

[8] Laemmli, W. E. (2017). Paper dancers: art as information in twentieth-century america. *Information & Culture, 52*(1), 1-30.

[9] Galli Stampino, M. (2017). A Theatre of Diplomacy. International Relations and the Performing Arts in Early Modern France.

[10] Cisneros, R. , Crawley, M. L. , & Whatley, S. (2020). Towards hybridity: dance, tourism and cultural heritage. *Performance Research, 25*(4), 125-132.

[11] Van Zile, J. (2021). Approaches to Dance (3): Naïveté and Curiosity. *Dance Research, 39*(2), 264-273.

[12] Hall, J. (2020). Judith butler and a pedagogy of dancing resilience. *Journal of Aesthetic Education, 54*(3), 1-16.

[13] Zhang, D. (2021). Intelligent recognition of dance training movements based on machine learning and embedded system. *Journal of Intelligent and Fuzzy Systems*(1), 1-13.

[14] Shi, Y. (2022). Stage performance characteristics of minority dance based on human motion recognition. *Mobile Information Systems, 2022*(1), 1940218.

[15] Siegmund, G. (2017). Ramsey burt, ungoverning dance. contemporary european theatre dance and the commons. *Dance Research, 35*(2), 277-278.

[16] Paramana, K. (2017). The contemporary dance economy: Problems and potentials in the contemporary neoliberal moment. *Dance Research, 35*(1), 75-95.