J. COMBIN. MATH. COMBIN. COMPUT. 124 (2025) 699–718

Journal of Combinatorial Mathematics
and Combinatorial Computing
www.combinatorialpress.com/jcmcc

Combinatorial Press

Research on the mechanism of emotion expression in dance based on machine learning models

Guojing Tan[1,✉], Jianan Wang[1]

[1] *School of Performing Arts, Sichuan University of Media and Communications, Chengdu, Sichuan, 610000, China*

ABSTRACT

The body language of dancers is vital for conveying emotion. In this study, Kinect is used to detect and track dancers' movements, and we develop two models: a dance action recognition model based on skeleton data and a dance emotion recognition model using an Attention-ConvLSTM. The action recognition model achieves 88.34% accuracy—reaching its best performance after just 40 iterations—while the emotion recognition model reaches an accuracy of 98.95%. Our analysis shows that features such as eigenvalue speed, skeleton pair distance, and inclination effectively differentiate emotions, although certain emotions (e.g., Excited vs. Pleased and Relaxed vs. Sad) can be confused. Notably, the leg's skeletal points significantly influence emotion expression. Ultimately, the study establishes a dance emotion expression mechanism through coordinated movement changes of the head, hands, legs, waist, and torso.

Keywords: kinect, action recognition, attention-ConvLSTM, emotion recognition, dance action

1. Introduction

Dance, as an ancient and vibrant art form, has been an important medium for people's emotional expression and communication since ancient times [8, 2]. Dance is mainly composed of the performer's skillful movements, dance vocabulary, performance emotions and other elements, which requires the dance performer to show the beauty of dance and express the feelings of the dance performer in the wonderful physical form [12, 14, 15]. Good dance emotion expression often requires dance performers to use accurate dance vocabulary, facial expression, etc., the character image of the dance work, the theme of thought, the spirit of the content, etc., image presented in front of the audience, and then successfully caused the emotional resonance of the dance viewers. In dance performance,

✉ Corresponding author.
E-mail address: 13982280543@163.com (G. Tan).

Received 11 September 2024; Accepted 09 February 2025; Published Online 19 March 2025.

DOI: 10.61091/jcmcc124-46

good emotional expression can strengthen the emotional thrust of the dance performance, highlight the character image of the dance performance, and enhance the artistic infectivity of the dance performance [4, 16, 25, 18]. Through the integration of exquisite skills and emotions, dancers show their inner emotional world through the form of dance, so that the audience can deeply feel the emotional information conveyed by the dance. Emotional communication is the purpose of the art of dance, and reasonable expression of emotion can substantially help the actor in the process of dance performance to enhance the dance expression and deepen the understanding of the dance performance process, so that the actor in the dance performance successfully shaped a complete and full character image [1, 22, 19, 7].

Dance, as a kind of silent art, uses the human body as a medium to convey feelings and meanings with the body. Chen et al. [6] starts from the innovation and expression of dance works, the role in cross-cultural communication, and the fusion of technology. Toppen [24] found that students' social and emotional learning skills can be enhanced through dance, and that students' focus on dance helps students forget any differences between them, and dance helps them get along better in the form of positive learning. Li [13] describes the current situation, definition and importance of dance education and dance art performance, as well as examining the impact of dance art performance on students, pointing out that a dance work is a vehicle for artists to express their emotions and thoughts through the language of dance, and the results of the study have a certain significance for the cultivation of students' art appreciation and aesthetic interests. Stutesman and Goldstein [20] used a mixed-method approach to qualitatively analyze the components of dance that convey emotion and found that there are four main components: narrative content, social interaction, emotional portrayal intention and the texture of the movement, and quantitative tests showed that the emotional portrayal intention is an important factor in the audience to accurately perceive the emotion.

In addition, Bernardi et al. [3] pointed out the importance of the movement expression of emotional experience for dance, finding through dance motion capture experiments that inter-individual differences in experienced emotions were significantly correlated with whole-body acceleration curves, and that the combination of dance movement and music enhanced dancers' sense of pleasure. Borowski [5] reviewed hundreds of papers on social and emotional competence and dance with the aim of exploring the mechanisms by which the dance experience promotes the occurrence of social and emotional competence development, and came up with four key influencing elements: self-suggestion, nonverbal expression and communication, embodied cognition and learning, and synchronicity and supportive learning environments. Zhao [26] specially designed a CBRSF model composed of machine learning technology, for emotion analysis and expression algorithms in the context of dance movements, and experimentally verified that the model can dynamically adjust the dance movements through real-time emotional cues, and thus realize nuanced emotional expression. Sun and Wu [21] synthesized the Kinect 3D sensor, feature extraction, gesture estimation and theoretical knowledge, and constructed an "arousal-emotion" emotion model based on the fusion neural network model, aiming at solving the problem of focusing on skills but not emotion in sports dance training, and integrating movement and emotion, thus improving the efficiency of dance training.

The article uses Kinect to detect and track the dancer's limbs, by constructing a dance action recognition model based on skeleton information, which uses the DWT algorithm to match the extracted dance action features with the standard action for recognition. The ConvLSTM dance action emotion recognition model is then optimized by adding the attention mechanism to construct the Attention-ConvLSTM based dance action emotion recognition model to recognize the emotion expressed by the dance action. Subsequently, the action recognition performance of the dance action recognition model based on skeleton information and the emotion recognition performance of the dance action emotion recognition model based on Attention-ConvLSTM are tested respectively. Finally, the emotional similarity is measured by the distance of the emotional state of the dance action, and the effect of different feature parameters and skeletal points on the emotional similarity is investigated. And based on the analysis results, the emotion expression mechanism of dance language is constructed.

2. Dance movement recognition model based on skeleton information

2.1. Kinect-based human detection and tracking technology

2.1.1. Human detection and tracking technologies.

(a) Human detection technology

In general, the classification of foreground targets mainly includes classification methods based on appearance, shape, features and other information. Specifically the commonly used classification methods are as follows:

(i) Adjacent frame difference method

$$D_x(x,y) = \left|\; f_{k-1}(x,y) - f_k(x,y) \;\right|. \tag{1}$$

Eq. (1) is a differential operation on the image. Ideally this operation can effectively extract moving objects, but in practice it is difficult to ensure the quality of the captured video, and the surrounding environment and the camera itself will bring a lot of noise to the capture results. Therefore, it is necessary to process the results of the differential operation to eliminate the effect of noise. Noise usually obeys a Gaussian distribution, so we can overcome the effect of noise on the differential operation by setting the threshold value, see Eq. (2):

$$\begin{cases} D_x(x,y) = 0, D_x(x,y) < \tau, \\ D_x(x,y) = 1, D_x(x,y) \geq \tau. \end{cases} \tag{2}$$

(ii) Background difference method

The background difference method is to do the difference operation between the current frame and the background to distinguish the foreground from the background. In the process of modeling the actual background, in order to estimate and recover the background and also want to increase the accuracy and reliability of the background model to a certain extent. Usually, the single Gaussian/multi-Gaussian background modeling method and the mean and median background modeling method are used.

Consider the video image $I\,(x,y,t)$ as consisting of a moving target $m\,(x,y,t)$ and a background $b\,(x,y,t)$, then:

$$I(x,y,t) = m(x,y,t) + b(x,y,t). \tag{3}$$

From Eq. (3):

$$m(x,y,t) = I(x,y,t) - b(x,y,t). \tag{4}$$

However, because of the presence of noise, the result calculated by Eq. (4) contains the differential image $d\,(x,y,t)$ in addition to the noise information $n(x,y,t)$:

$$d(x,y,t) = I(x,y,t) - b(x,y,t) - n(x,y,t). \tag{5}$$

(iii) Optical flow method

The optical flow method begins by projecting a 3D model into a 2D plane, combining the motion target of the 3D model with the motion target of the 2D image. The motion of each pixel point is determined according to the change in intensity of the pixel point in the time domain.

(b) Human body tracking technology

A tracking method that monitors detected targets and analyzes and predicts their trajectories before tracking the target of interest is called target tracking. People usually use feature-based, contour-based and other methods for human body tracking.

(i) Feature-based tracking

This refers to tracking the human body by combining features such as texture, shape, edges and color of the region. Point features and corner features of moving contours are commonly used features.

(ii) Active contour-based tracking

Given a confined curved contour of the human body first, the motion of this curved contour is tracked next. The focus of this method is the need to accurately give the human body contour.

(iii) Region-based tracking

This method divides the human body into several regions, including the head, arms, torso, etc., and then tracks each region separately according to the ratio, and then combines the tracking ratio for overall tracking.

(c) Model-based human body tracking

This refers to the tracking based on the geometric model of the human body, and the commonly used methods are 2D contour method, three-dimensional model method and line drawing method.

2.1.2. Kinect-based human detection and tracking technology.

1) Kinect depth map principle and measurement. The imaging principle of Kinect is based on optical encoding technology, and its principle is shown in Figure 1.

Therefore, we project the scattered structured light in the space to be detected, and then complete the marking of the entire space. For any object appearing in this space, we only need to obtain the scattering information on the surface of the object, and then we can obtain the coordinate information of the object.

Kinect needs to be calibrated separately for RGB and IR cameras. The spatial point p raw value d_r can be derived from the correspondence between the color image and the depth information:

$$d_r = K \tan\left(H \cdot d + L\right) - O. \tag{6}$$

The depth map is labeled with different colors depending on the distance between Kinect and the target.

After obtaining the depth of the image, the world coordinates of the point can be obtained, and

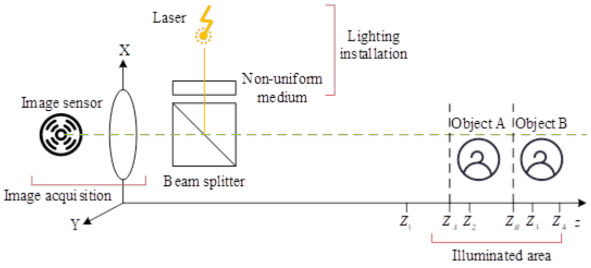


Fig. 1. Depth imaging system

the depth coordinate (x_d, y_d, z_d) corresponds to the world coordinate (x_w, y_w, z_w) as:

$$\begin{cases} x_w = (x_d - \frac{w}{2}) \cdot (z_w + D') \cdot F \cdot (\frac{w}{h}), \\ y_w = (y_d - \frac{h}{2}) \cdot (z_w + D') \cdot F, \\ z_w = d, \end{cases} \tag{7}$$

where $D' = -10, F = 0.002$, the resolution $w \times h$ of Kinect is 1920×1080.

2) Kinect-based human joint point recognition. Kinect is capable of recognizing and tracking the human skeleton [23]. Kinect first recognizes the 25 joint point coordinates of the human body, establishes the human skeletal structure, and combines the depth information to realize the representation of the human skeleton structure in three-dimensional space, and the cross type of the skeleton joints is shown in Figure 2.

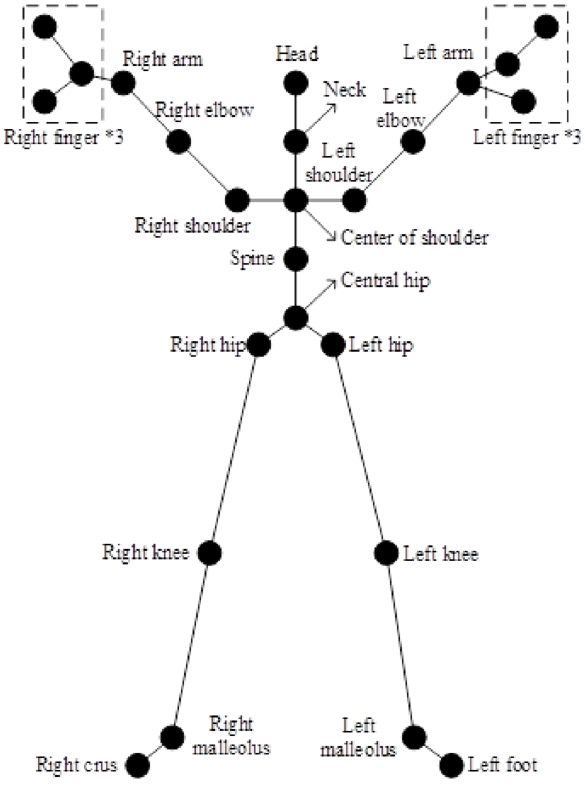


Fig. 2. Human skeleton node model identified by Kinect

Human body joint point recognition consists of the following three parts:

a. Remove the background, first find the possible areas of the human body. The outline of the human body is extracted using edge detection. The specific method is to use the distance from the Kinect sensor to analyze.

b. Recognition of important parts of the human body, including the head, arms, legs, and body torso.

c. Human body joints recognition, in Kinect the human body will be connected by joints, so Kinect analyzes the front and side joints to determine the coordinate position situation of the human body, the human body node recognition steps are shown in Figure 3.

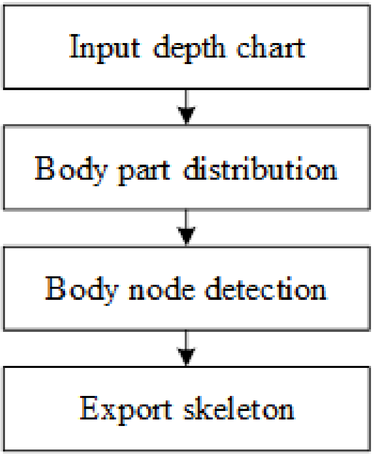


Fig. 3. Procedure for identifying a human node

Each pixel information can be inferred from the body component recognition. Define the following density estimate of the body component as:

$$f_c(\hat{x}) \propto \sum_{i=1}^{N} w_{ic} \exp\left(-\left\|\frac{\hat{x}-\hat{x}_i}{b_c}\right\|^2\right), \tag{8}$$

where $\hat{x}$ is the 3D spatial coordinates, N is the number of pixels, w_{ie} is the pixel weights, $\hat{x}_i$ represents the projection of the pixel x_i into world space, and b_ε represents the width of each component. $w_{\iota e}$ then balances the pixel inference probability and the spatial area probability:

$$w_{ic} = P(c|I, x_i) \cdot d_l(x_i)^2. \tag{9}$$

Such an approach improves the accuracy of joint predictions and also allows for depth-invariant density estimation.

2.2. *Dance movement feature extraction method*

The first set of features is computed from the 3D coordinate information of the key points of the human posture obtained from the dance video. These features utilize the information of the 25 skeletal joint points of the human body based on the angle descriptors formed by the adjacent skeletal parts defined by the adjacent skeletal joints.

The second set of motion features is computed from the time series of the 3D coordinate information of the key points of the human body posture.

The third set of motion features is also computed from the action timing sequence, and the total number of dance video frames is F. The instantaneous velocity of each joint i at each frame can be expressed as a velocity vector ve_i as shown in Eq. (10), and the velocity feature vector constructed from the 25 body average frame velocities $\bar{v}_i$ can be expressed as $Vels = (ve_1, ve_2, \ldots, ve_2)$:

$$ve_i = (v_i^1, v_i^2, \ldots\ldots, v_i^F). \tag{10}$$

2.3. Similarity measures for angular features

The Euclidean distance of the joint angles in the key frames of the dance practitioner's video of the to-be-tested movement and the corresponding standard frames of the reference dance movement is used as a measure of joint angle similarity.

$$a_i = 1 - \frac{\theta_i^{key} - \mu_i}{\mu_i}. \tag{11}$$

In this paper, K-Means algorithm is used to realize the prediction of key frames. The clustering center adopts a series of standard same dance action key frame frames extracted from the motion features beforehand, the distance between the clustering center and the sample points is the sum of the variance of the angles of all the corresponding joints between frames, and the calculation process is shown in Eq. (12):

$$Dt = \sum_{i=0}^{N-1} (\theta_i - \mu_i)^2 . \tag{12}$$

2.4. Similarity metrics for motion trajectory and velocity features

The DTW algorithm is a method created to measure the similarity between two time sequences, initially explored mainly in the field of speech recognition, and later extended to music, sports, etc [11]. DTW is suitable for similarity comparison between two time sequences of varying length. In the case of dance movements it is manifested in the movement trajectories of dancers in different videos with different joint velocities.

$$\begin{bmatrix} d(1,1) & d(1,2) & \ldots & d(1,n) \\ d(2,1) & & & \vdots \\ \vdots & & \ddots & \vdots \\ d(m,1) & d(m,2) & \ldots & d(m,n) \end{bmatrix}, \tag{13}$$

where

$$d(i,j) = \sqrt{(p_i - q_j)^2}. \tag{14}$$

The regularized path W is a continuous element in the distance matrix denoting the correspondence between P and Q as shown in Eq. (15):

$$W = \omega_1, \omega_2, \omega_3, \ldots, \omega_K. \tag{15}$$

3. Dual-stream convolutional dance action emotion recognition modeling

3.1. Dual-stream convolutional body movement emotion recognition based on dance videos

In this paper, we propose a dual-stream convolution-based emotion recognition method for dance movements, which can well characterize the whole body movement of the dancer through optical flow data.

The network proposed in this section contains two independent convolution channels, i.e., spatial flow convolution channel and temporal flow convolution channel, and the results are integrated with post fusion. The spatial flow channel is to characterize the spatial attributes of the human gestures in the video data. The temporal flow channel is to characterize the limb movement information of the person in the video data. In this paper, we set its input as a dense optical flow displacement field between stacked consecutive frames of video data. This input can explicitly characterize the pixel-level motion between the video frame data, so that the neural network does not need to implicitly estimate the motion features anymore, which makes the neural network more accurate in recognizing emotions:

$$\begin{cases} I_\tau(u,v,2k-1) = d^x_{\tau+k-1}(u,v) \\ I_\tau(u,v,2k) = d^y_{\tau+k-1}(u,v) \\ u=[l;w], \quad v=[l;h], \quad k=[l;L] \end{cases} \tag{16}$$

As shown in Eq. (16), the dense optical flow d, can be regarded as the displacement vector field of the video data t and $t+1$ in consecutive frames, and $d_t(u,v)$ represents the displacement vector of the point (u,v) in the tth frame of the video data. The displacement components d^x_t, d^y_t in the horizontal and vertical directions can be treated as two channels of the image, respectively, which is also very much in line with the characteristics of the convolutional network input.

However, later experimental results show that such a network structure fails to achieve good experimental results, therefore, this paper improves the two-stream convolution in Section 3.2.

3.2. Emotion recognition algorithm improvement

3.2.1. ConvLSTM. Recurrent neural network (RNN) is a kind of neural network with feedback mechanism that can support sequence-to-sequence input and output [9]. Video data has obvious sequence characteristics, and the dependency between successive frames of time-streamed data has a great impact on the success rate of emotion recognition in emotion recognition of dance movements based on video data.

Long Short-Term Memory (LSTM) recurrent neural network utilizes specific memory cells to store the information that needs to be remembered and the mechanism of specific gates determines when the information is updated [17]. The LSTM network updates the recursive formula as follows:

$$\begin{cases} i_t = \sigma(W_{xi}x_i + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i) \\ f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f) \\ c_t = f_t c_{t-1} + i, \tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \\ o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o) \\ i_t = o_t \tanh(c_t) \end{cases} \tag{17}$$

where w is the weight matrix of each gate unit, σ is the activation function, x_t is the feature vector of the new input, h_{t-1} is the hidden state of the previous time step, and b is the bias term. c_t is the memory cell, which is essentially a state accumulator that is updated to a certain extent for each input and output, i_t represents the input gate, which determines the extent to which the new input affects the new memory cell, f_t represents the forgetting gate, which determines the extent to which the old memory cell affects the new memory cell, and o_t is the output gate, from which the output is entered into the next LSTM cell as a hidden state. Each gate value is a vector of the same dimension as the memory cell. From Eq. (17), it can be seen that the gate values at moment t are jointly influenced by the memory unit at moment $t-1$, the hidden state, the new input, and the bias value.

However, the traditional long and short-term memory neural networks may lead to some spatial connections between local features of the data with spatial features being ignored.

In order to solve this problem, many scholars have proposed a convolutional long and short-term memory neural network (ConvLSTM), which can retain the advantages of the traditional long and short-term memory neural as well as the spatial features of the information such as images, videos, etc. [10]. The updating recursive formula of ConvLSTM is as follows:

$$\begin{cases} I_t = \sigma(W_{xi} * X_t + W_{hi} * H_{t-1} + W_{ci} * c_{t-1} + b_i) \\ F_t = \sigma(W_{xf} * X_t + W_{hf} * h_t + W_{cf} * c_{t-1} + b_f) \\ C_t = F_t \odot C_{t-1} + I_t \odot \tanh(W_{xc} * X_t + W_{hc} * H_{t-1} + b_c) \\ O_t = \sigma(W_{xo} * X_t + W_{ho} * H_{t-1} + W_{co} * C_t + b_o) \\ H_t = O_t \odot \tanh(C_t) \end{cases} \tag{18}$$

where $*$ denotes the convolution operation, $\odot$ denotes the Hadamard product, and $W_{x\sim}, W_{h\sim}$ denotes the two-dimensional convolution kernel. Gate unit I_t, F_t, O_t, memory cell C_t, and hidden state H_t, H_{t-1} are all three-dimensional tensors, and this structure can better preserve the local features of the data and the spatial relationship between the features compared to the traditional long and short-term memory neural networks.

3.2.2. Algorithm improvement based on attention-ConvLSTM. In order to preserve the local features in the data and the spatial relationships between the features while the dependencies between each successive frame of the optical flow data of the time stream in the dual-stream convolution are better utilized. In this paper, we use the convolutional long and short-term memory neural network ConvLSTM in the convolutional layer of the dual-stream convolutional time-stream channels, i.e., the input x of the long and short-term memory neural network is expanded into the input X_τ of the $N \times N \times D$, where $N \times N$ is the size of the feature-map in the current convolutional layer, and D is the number of channels of the convolutional layer. In addition, in order to make the neural network can pay better attention to the significant features related to emotion in the optical flow data, and make certain key frames which are important for emotion recognition better utilized, this paper also introduces the attention mechanism in ConvLSTM, and the structure of the convolutional long-short-term neural network based on the attention mechanism designed in this paper is shown in Figure 4.

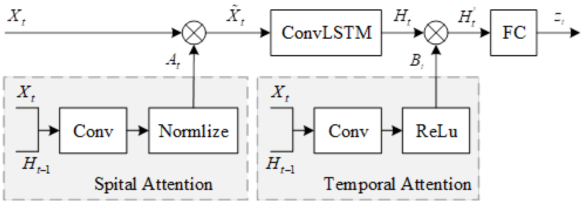


Fig. 4. Convolutional short-duration memory neural network based on attention mechanism

In this paper, in the convolutional layer of the neural network, the attention weights are obtained from the previous hidden state H_{t-1} and the current feature map X_t by convolution operation:

$$S_t = W_s * \tanh(W_{xa} * X_t + W_{ha} * H_{t-1} + b_a). \tag{19}$$

$$A_t^i = p(att_{ij}|X_t, H_{t-1}) = \frac{\exp(S_t^{ij})}{\sum_i \sum_j \exp(S_t^{ij})}. \tag{20}$$

$$\tilde{X}_t = A_t \odot X. \tag{21}$$

In Eq. (19), w is the convolution kernel, by replacing the inner product with the convolution, s_i is a two-dimensional attention score map, which can be normalized and calculated to obtain the attention weight map A_τ in Eq. (20), and the element A_τ^Ψ is the value of the attention map at position (i, j). For A_{ij} there exists an obvious constraint relationship, $\sum_i \sum_j A_i \approx 1$, in order to prevent the update of parameters is trapped in the local optimal solution, and lead to certain feature regions are completely ignored, this paper hopes that $\sum_{r=1}^T A_{\tau,j} \approx T$, T is the total number of frames of the optical flow information sampled in the video data, this topic through the Hardamad multiplication shown in Eq. (21) to obtain the feature map $\tilde{X}_t$ processed by attention.

In order to enable certain key frames to be more fully utilized, this paper also proposes an order-based temporal attention mechanism to provide different levels of attention to different frame data. Eq. (22) in B, represents the neural network's attention to each frame of optical flow information, and Relu is chosen as the activation function because of its good convergence performance:

$$B_t = \mathrm{R}elu(W_{xb} * X_t + W_{hb} * H_{t-1} + b_b). \tag{22}$$

$$H_t^{'} = B_t \odot H_t. \tag{23}$$

In order to prevent the optical flow data of a certain frame from influencing the final prediction result too much, and the problem of gradient vanishing during backpropagation, this study limits the size of $\sum_{r=1}^T \|B_t\|_2$ in Eq. (23), because the gradient is proportional to $1/B_\iota$ during backpropagation, and $H_\tau^{'}$ is the final output of ConvLSTM.

As shown in Eq. (24) in the optical flow part of the video level prediction is mainly based on the fully connected output z_t of each frame of optical flow data, where $o = (o_1, o_2, ..., o_C)^T$:

$$o = \sum_{t=1}^{T} z_t. \tag{24}$$

$$p(C_i|\chi) = \frac{e^{o_i}}{\sum_{j=1}^C e^{o_j}}, k = 1, ...C. \tag{25}$$

In Eq. (25), c is the number of all possible classifications, and $p(C_i|X)$ denotes the probability that the sentiment expressed by a given sequence frame X belongs to the ith class of sentiment.

Summarizing the above analysis, this topic proposes a new regular cross-entropy loss function as shown in Eq. (26):

$$L = -\sum_{i=1}^{C} y_i \log \hat{y}_i + \lambda_1 \sum_i^N \sum_j^N \left(1 - \frac{\sum_{i=1}^T A_{i,ij}}{T}\right)^2 + \frac{\lambda_2}{T} \sum_{t=1}^{T} ||B_t||_2 + \lambda_3 \sum_i \sum_j \theta_{i,j}^2. \tag{26}$$

$y = (y_1, ..., y_C)^T$ is the real label of the dataset, when the emotion expressed in the current video data belongs to class i emotion, and $y_i = 1, y_j = 0 (j \neq i), \hat{y}_i = p(C_i|X)$, θ denote all the model parameters, which are also restricted in order to prevent the neural network from overfitting.

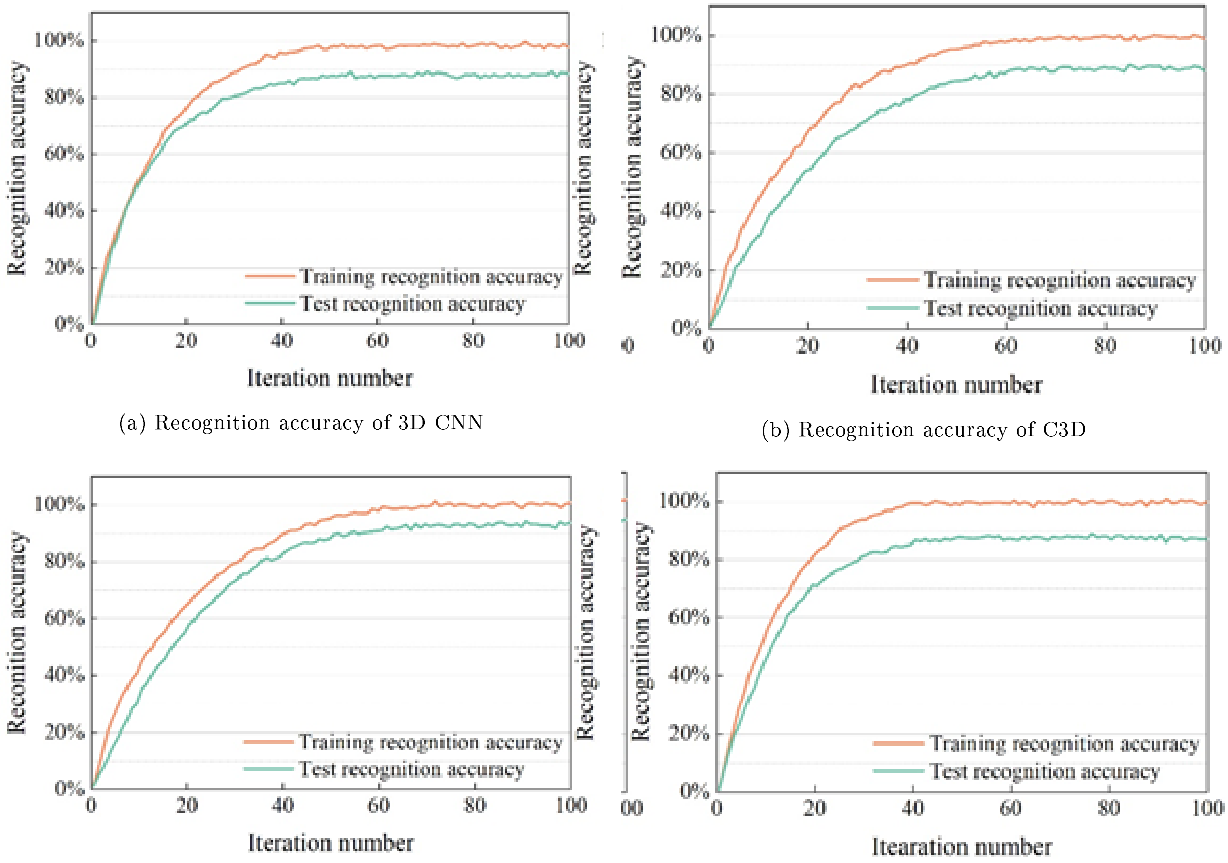
With the above improvements, the network model can not only make full use of the temporal and spatial characteristics in the video optical flow data, but also better capture the emotion-related body movement features and better utilize the key frames in the single-frame optical flow data.

4. Analysis of dance movement recognition and emotion expression mechanisms

4.1. Performance analysis of dance movement recognition

4.1.1. Action recognition performance. In order to verify the effectiveness of the dance action recognition model based on skeleton information proposed in this paper, the dance action recognition model in this paper is compared with other recognition models (3D CNN, C3D, PoseC3D). The experiments were conducted using the MSRAction3D dataset, which is a dance action dataset captured by the Kinect depth sensor and contains 20 classes of actions, totaling 625 action samples. The recognition effect of the dance action recognition model on the MSRAction3D dataset is shown in Figure 5, and Figures 5a to 5d show the recognition accuracies of 3D CNN, C3D, PoseC3D, and the dance action recognition model in this paper, respectively.

Observing Figure 5, it can be seen that the recognition accuracies of 3D CNN, C3D, and PoseC3D methods on MSRAction3D dance action dataset are 85.76%, 83.01%, and 82.66%, respectively, and the best results are achieved after 50, 60, and 70 iterations of these three action recognition methods, respectively. And the recognition accuracy of the dance action recognition model in this paper on the MSRAction3D dance action dataset is 88.34%, which is higher than the above three dance action recognition models, and it reaches the best recognition effect after only 40 iterations.



(a) Recognition accuracy of 3D CNN

(b) Recognition accuracy of C3D

(c) Recognition accuracy of PoseC3D

(d) Recognition accuracy of our model

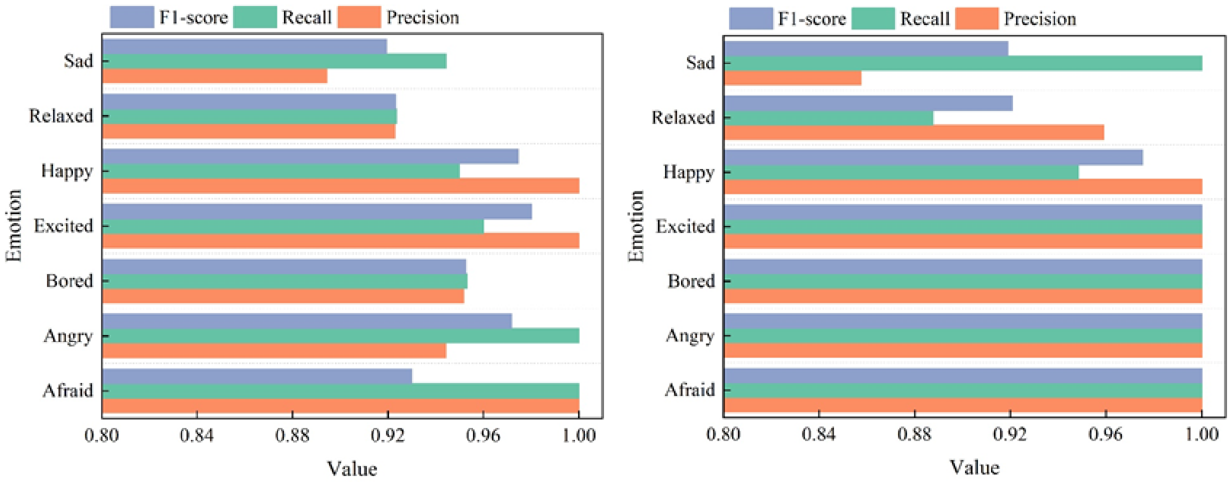
Fig. 5. Recognition results of dance movement recognition models

4.1.2. Emotion recognition performance. Currently, there is a lack of research related to emotion recognition based on dance movements, so in this subsection, decision tree algorithm, random forest algorithm, and deep learning model based on BLSTM were used to conduct a comparative experiment on emotion recognition based on dance movements on the same dataset (MSRAction3D), and the experimental results are shown in Table 1. As can be seen from Table 1, the recognition rate of all methods reaches more than 90%, while the Attention-ConvLSTM method in this paper has the highest recognition rate of 98.95%, which is higher than the other methods due to the consideration of the temporal features, and the recognition rate is improved by 1.90% compared with the ConvLSTM algorithm before optimization. Machine learning algorithms also achieve better recognition results, as can be seen in Table 1, the recognition rates of the Decision Tree algorithm and the Random Forest algorithm on the same dataset are 95.56% and 93.86%, respectively. The Attention model, on the other hand, performs relatively poorly with a recognition rate of 93.14%, which is lower than all other methods.

Table 1. Recognition rate of different models on the same data set

Method	Recognition rate
Attention-ConvLSTM	98.95%
ConvLSTM	97.05%
Attention	93.14%
Decision tree	95.56%
Random forest	93.86%

To further explore the recognition effect of the Attention-ConvLSTM emotion recognition algorithm in this paper for different dance movements, the recognition accuracy, recall and F1-score of ConvLSTM and Attention-ConvLSTM models for different emotion dance movements are shown in Figure 6, Figure 6a and Figure 6b show the recognition accuracy, recall and F1-score of ConvLSTM and Attention-ConvLSTM models for emotion recognition effect.



(a) Recognition results of ConvLSTM

(b) Recognition results of Attention ConvLSTM

Fig. 6. Recognition accuracy, recall rate and F1-score of different emotional dance movements

From the comparison of Figure 6a and Figure 6b, it can be seen that the precision, recall, and F1-score of the Attention-ConvLSTM model expressing the dance moves of Afraid, Angry, Bored, and

Excited are significantly improved with respect to that of the ConvLSTM model before optimization. However, the 3 measures of emotion recognition for dance moves expressing Happy did not change. The emotion recognition precision for the dance action expressing Relaxed increased by 3.62%, but the recall decreased to 88.76%, and the F1-score remained almost unchanged. The recall of emotion recognition for dance moves denoting Sad increased to 100%, but the precision decreased by 3.69% and the F-score remained the same. Combined with Table 1, it can be seen that the Attention-ConvLSTM model has overall better performance in most cases relative to the pre-optimization ConvLSTM method.

This subsection also uses seven dances with different emotions to test the performance of the Attention-ConvLSTM model, the ConvLSTM model, and the Attention model, respectively. The recognition rates of the three emotion recognition methods on the dance movements with different emotions are shown in Figure 7.

From Figure 7, it can be seen that the recognition rate of the Attention-ConvLSTM model is almost higher than that of the ConvLSTM model and the Attention model for each emotion, but the recognition rate is lower for the emotion representing Sad, and for the emotion representing Excited, the performance of all the three models is better, which is above 90%. It can be seen from the above experimental analysis that a higher recognition rate was obtained by adding the Attention model to the ConvLSTM model and combining the two. The lowest recognition rate was obtained using only the Attention model.

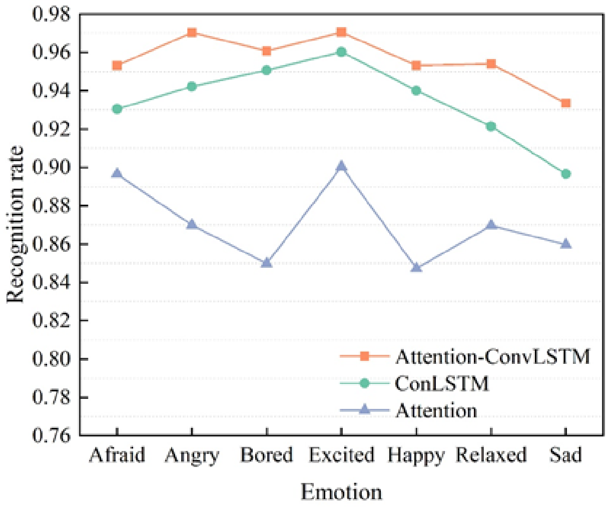


Fig. 7. Recognition rate of 7 different emotional dance movements

4.2. Dance movement emotion metrics

In this section of the experiment, 12 different body emotions were selected from the University of Cyprus' Dance Emotion Movement Database for different dancer sample data sets regarding afraid, angry, annoying, bored, excited, happy, miserable, pleased, relaxed, sad, satisfied, and tired. Among them, joy, anger, sadness, and fear are the 4 basic body movement emotions, and the others are derived emotions. Eight of these emotions were selected to verify the validity of the methodology of this paper. This section of the experiment is based on the similarity between different dance body movements to measure the expression of emotion, the lower the similarity, the more similar the two emotional body movements.

4.2.1. Affective state distance measures. Characteristics that affect the emotion of body movements are the tendency of the body's center of gravity, the relative position of bone points, and speed. The feature language is described based on the human center of gravity joint point Hip, according to the dancer's body structure characteristics of the description of the 9 kinds of feature language there are 3 kinds of characteristics can be quantified, respectively, speed, bone to distance and bone to the angle of pinch. As for the bone points, there are 17 main joints in the dancer's body structure, and the joints related to the expression of emotion are the arm, leg and head, and the experiments in this paper have found that the physical movements of the foot also have a certain influence on the expression of emotion, so the experiments in this section have chosen six bone points as the dancers' body structure, namely, Left leg, Left foot, Right leg, Right foot, Right arm, Left arm, Left arm, Right arm, Right arm, Right arm, Right arm, Right arm, Right arm, Left arm. Left leg, Left foot, Right foot, Right foot, Right arm, and Left arm are used as the measures of the dancers' emotional changes during the dancing process. In this section, the experiment combines different dancer samples, eight kinds of body movement emotions, three kinds of eigenvalues and six skeletal limbs, and takes the similarity between the emotion Excited and other seven kinds of emotions as an example to evaluate the similarity of body emotions of dance movements, and recognize and classify the emotions. The similarity of limb emotions between a large number of different movement samples is calculated, and the confusion matrix of the resultant average is shown in Table 2.

From Table 2, it can be seen that among the eight emotions, the distance between the emotions Excited themselves is 0, i.e., the emotions themselves are completely similar to each other. However, the similarity between the different dance movement segments emotions described by different feature languages are all different. For eigenvalue speed, the action limbs of emotion Excited and emotion Relaxed are basically the same, while the lowest similarity is between them and emotion Pleased. For Skeleton to Distance, the highest similarity between emotion Excited and Pleased was 8.439, in other words, the action limbs expressing these two emotions were very similar, whereas the limbs expressing them differed the most from emotions Miserable and Mixed. Similarly, the skeletal pair pinch angle, the shortest distance between emotion Excited and Relaxed is 0.503, i.e., the action limbs of emotion Excited and Relaxed are more similar, followed by emotion Pleased and Happy, and the lowest similarity with emotion Miserable and Sad.

Combining the results of the 3 different feature parameters, it can be seen that the emotions that are similar to emotion Excited are Relaxed, Pleased, and Happy, while they are least similar to emotions Miserable, Sad, and Mix. Therefore, in this paper, the emotions Excited, Relaxed, Pleased and Happy are categorized into the table Happy category of emotions {H}, while Miserable, Sad and Mix are categorized into the Sad category of emotions {S}.

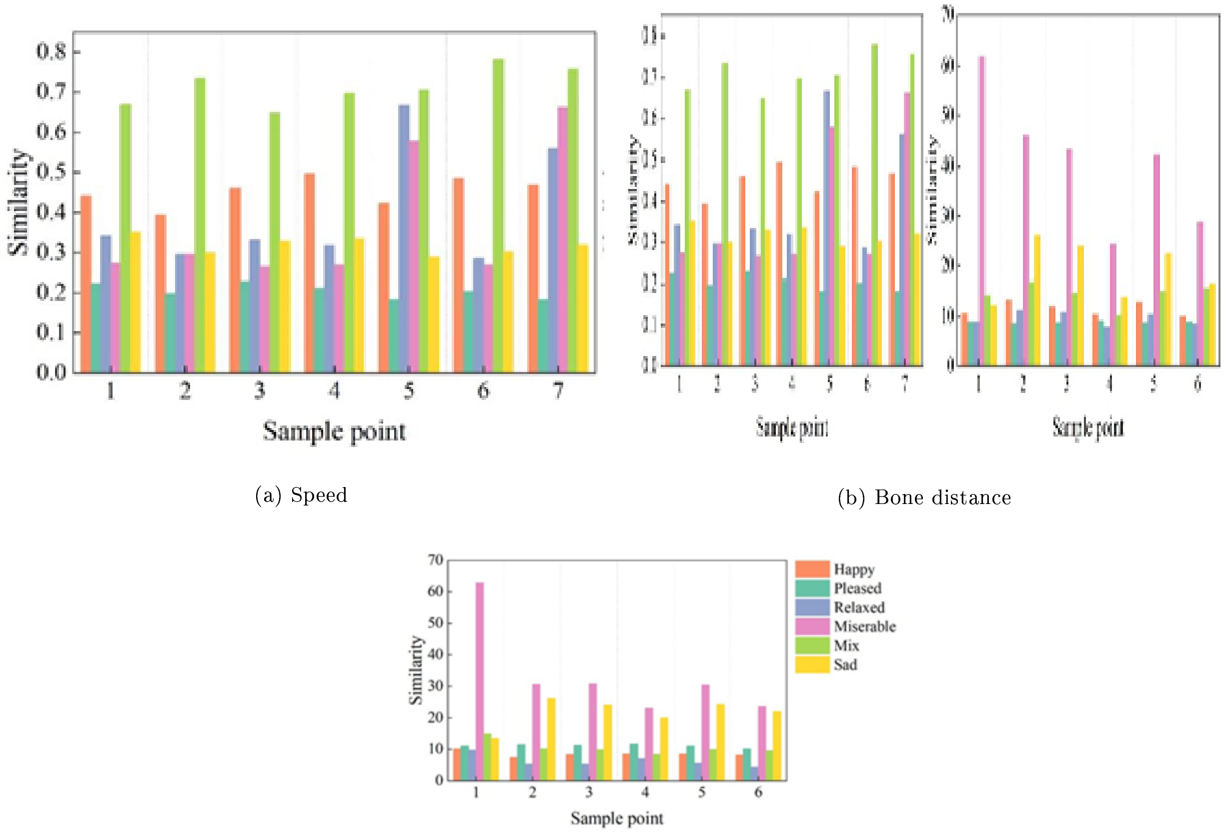
4.2.2. Effect of different feature parameters on sentiment similarity. The experiments in this chapter selected 3 feature parameters, namely, speed, bone-to-bone distance and bone-to-pinch angle, respectively, to measure the similarity between the emotions of different dance movements. However, considering the influence of the dancer's own body structure on emotion recognition, the 2 features of bone-to-distance and bone-to-pinch angle were also selected. From Section 4.2.1, it can be seen that there are effects of different feature parameters on the recognition results of emotion. Therefore, this section starts with the similarity metric for a more comprehensive analysis.

Based on the emotion similarity in Section 4.2.1, in order to analyze the influence of different feature parameters on emotion recognition, this section again averages the similarity metric for the same emotion for multiple samples, and the results are shown in Figure 8, with 8a to 8c being the

Table 2. Similarity between movements and emotions

Speed								
	Excited	Happy	Miserable	Mix	Pleased	Relaxed	Sad	Tired
Excited	0	0.452	0.425	0.471	0.531	0.394	0.510	0.462
Happy	0.452	0	0.448	0.402	0.448	0.401	0.427	0.449
Miserable	0.425	0.448	0	0.724	0.964	0.448	0.795	0.657
Mix	0.471	0.402	0.724	0	0.516	0.391	0.455	0.470
Pleased	0.531	0.448	0.964	0.516	0	0.372	0.389	0.410
Relaxed	0.394	0.401	0.448	0.391	0.372	0	0.704	0.629
Sad	0.510	0.427	0.795	0.455	0.389	0.704	0	0.456
Tired	0.462	0.449	0.657	0.470	0.410	0.629	0.456	0
Bone distance								
	Excited	Happy	Miserable	Mix	Pleased	Relaxed	Sad	Tired
Excited	0	10.050	63.074	15.426	8.439	11.522	13.089	12.088
Happy	10.050	0	25.465	11.562	16.852	10.845	13.585	11.746
Miserable	63.074	25.465	0	34.450	83.748	52.645	60.456	29.452
Mix	15.426	11.562	34.450	0	16.352	11.026	9.185	8.253
Pleased	8.439	16.852	83.748	16.352	0	10.035	17.166	16.875
Relaxed	11.522	10.845	52.645	11.026	10.035	0	14.653	11.344
Sad	13.089	13.585	60.456	9.185	17.166	14.653	0	12.066
Tired	12.088	11.746	29.452	8.252	16.875	11.344	12.066	0
Bone angle								
	Excited	Happy	Miserable	Mix	Pleased	Relaxed	Sad	Tired
Excited	0	0.604	0.705	0.633	0.568	0.503	0.673	0.618
Happy	0.604	0	0.698	0.564	0.549	0.502	0.563	0.615
Miserable	0.705	0.698	0	0.685	1.225	0.975	0.978	0.875
Mix	0.633	0.564	0.685	0	0.804	0.562	0.789	0.560
Pleased	0.568	0.549	1.225	0.804	0	0.481	0.594	0.587
Relaxed	0.503	0.502	0.975	0.562	0.481	0	1.063	0.976
Sad	0.673	0.563	0.978	0.789	0.594	1.063	0	0.567
Tired	0.618	0.615	0.875	0.560	0.587	0.976	0.567	0

results of emotion similarity for speed, bone-to-distance, and bone-to-intercept angle, respectively. From Figure 8, it can be seen that for the feature parameter Speed, the similarity of emotions Relaxed, Sad and Miserable is very close, so these three emotions can be easily confused with each other, while the difference between emotions Mixed, Happy and Pleased is more obvious. For the feature parameter Skeleton Pair Distance, the similarity of Emotion Pleased, Relaxed and Happy is basically the same, and it can be seen from Section 4.2.1 that these 3 emotions are of the same type of emotion, while Emotion Miserable, Sad and Mix belong to the Sad category of emotion, and it can be seen from Figure 8a that it is very easy to differentiate between the similarity of these 3 emotions and that of the Happy category of emotion. The similarity of the feature parameter Skeletal Pair Clamping Angle, Emotion Mix with Pleased, Relaxed and Happy emotions is basically the same, i.e., Emotion Mix can be easily confused with the Happy emotion. In conclusion, analyzing from the perspective of similarity of emotions, the feature parameter Skeletal Pair Distance has a better effect on distinguishing different types of emotions, while the similarity of emotions calculated based on the other 2 feature parameters, there are action emotions that are confused with each other.



(a) Speed

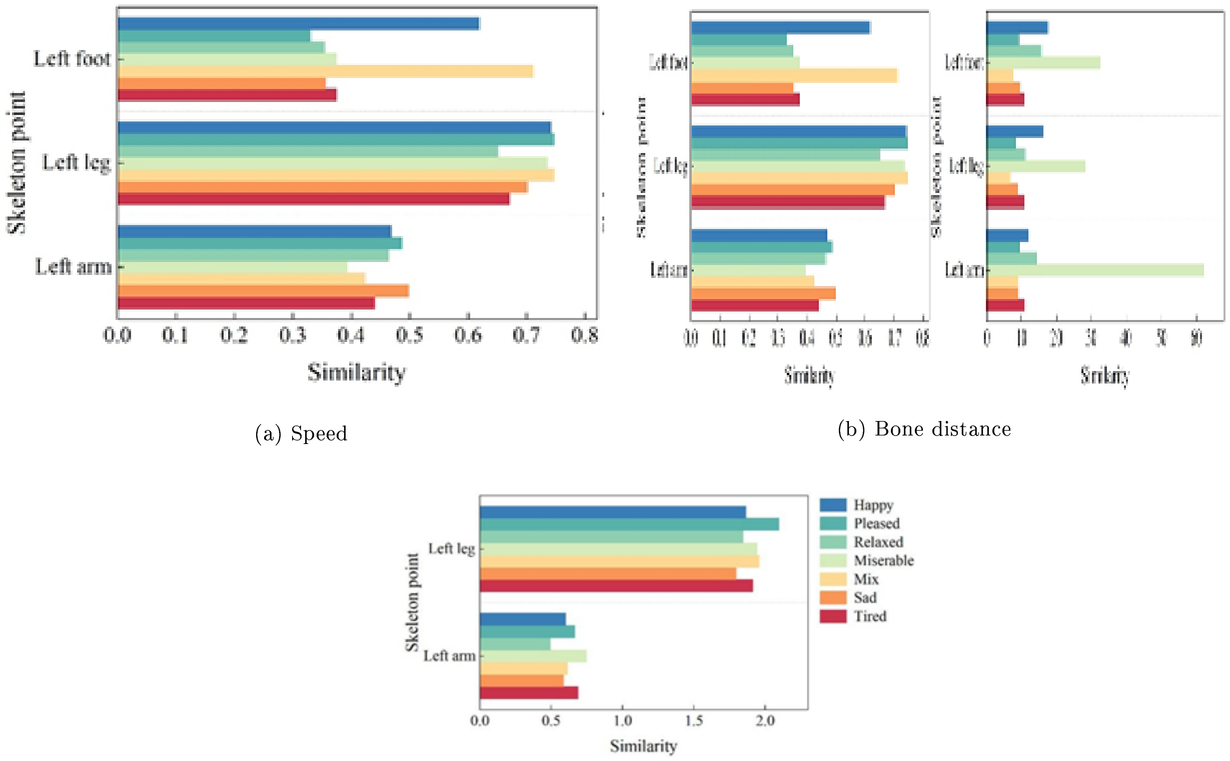
(b) Bone distance

(c) Bone angle

Fig. 8. Emotional similarity measurement results

4.2.3. Effect of different skeletal points on emotional similarity. Section 4.2.1 mentions that there are 17 main active skeletal points in the dancer's body and that the physical expression of different dance emotions is different for different dance movements. Therefore, this section analyzes the emotional similarity of the same emotion in different limb parts, verifies and illustrates what are the main skeletal (joint) points that affect the physical emotion of dance movements. In this section, three kinds of feature parameters were selected to measure the similarity between the joint points Left arm, Left leg and Left foot and the emotion Excited respectively, and the emotional similarity of different skeletal points is shown in Figure 9, which is used to indicate the magnitude of action tension and expressiveness of emotional expression, and 9a to 9c are the emotional similarity of speed, bone-to-distance, and bone-to-pinch angle respectively Results.

As can be seen from Figure 9, the similarity between the analysis and emotional Excited from the perspective of speed, it can be seen from Figure 9a that the bone leg has a greater magnitude of action tension compared to the bones arm and foot, i.e., the leg bone point has a greater impact on emotional expression. For the characteristic parameter bone-to-distance, Figure 9b, the emotional similarity of Left arm, Left leg and Left foot when the dancer expresses the emotion Miserable is 62.37, 28.31, and 32.48, respectively.Whether it's the upper half of the limb (Left arm) or the lower half of the limb (Left leg and Left foot), the two kinds of emotion's movement limb tensions are different though, but the lower half of the body has a smaller movement amplitude than the upper itself. Secondly, for the emotion Tired, the movement expressiveness of the lower half of the limb (10.68, 10.68) was greater than the movement amplitude of the upper half of the limb (10.95). When interpreting the emotion Mix, the expressiveness of limb FOOT (7.67) was greater, while the other emotional trends did not differ significantly across limb parts. Thus, limb ARM, LEG and FOOT all have a small effect on emotional expression. Skeletal pair pinch angle in the expression of different movement emotions with emotion Excited, Figure 9c, the amplitude of the upper body (arm) limb movement is significantly smaller than the lower body (leg) movement amplitude, that is, the limb leg emotion expression has a greater impact in the expression of different dance limb movement emotions.



(a) Speed

(b) Bone distance

(c) Bone angle

Fig. 9. Emotional similarity of different skeleton points

4.3. Emotional expression mechanism construction

Body language, also known as gesture language or action language, is a non-verbal way of communication that conveys information, expresses emotions and exchanges thoughts through the body's gestures, movements and rhythms. In the art of dance, body language plays a crucial role and is the main means of expressing emotions and shaping artistic images. Emotional expression is through a series of body language, movements, gestures, facial expressions and music in conjunction with each other, together building a colorful dance world, so that the audience can feel the profound emotions conveyed by the dance. Based on the recognition and measurement of the emotion of the dance movements in the previous section, the mechanism of dance emotion expression is constructed according to the changes of the movements of each part of the body.

1) Realize emotional expression with the help of head movement changes. When expressing emotions with the help of head movement changes, performers should pay attention to the coordination between head movements and body movements, avoid over-exaggeration or false pretense, and really understand and feel the connotation of the dance through in-depth understanding, and make the dance smooth and natural, real and believable, and convey emotions through the cooperation between the head and the body.

2) Realize emotional expression with the help of hand movement changes. Among the dance performance movements, hand movements are extremely rich, such as the different meanings represented by different hand shapes are also different. Performers should pay attention to the changes in hand movements and other parts of the limbs with the changes in movement, in order to show the different roles of emotion.

3) Emotional expression with the help of leg movement changes. In the process of practicing leg movements, the dance performers should grasp the specific meanings represented by different leg movements, and the skillful use of different leg movements will make the image of the character more distinct.

4) Realize emotional expression with the help of waist movement changes. Dance performers should pay attention to the practice of waist strength, and realize the coordination of the overall body movements through such practice. Through the practice of waist movement, the performer will be more flexible in mobilizing other parts of the body at the same time, to achieve the effective expression of artistic emotion.

5) Emotional expression with the help of torso movement changes. The torso plays a linking role for other parts of the body. Grasping the movement of the torso will help the dancers to realize enough performance tension, thus conveying the emotional connotation of the work more profoundly. Changes in torso movements can reflect the dancer's inner emotional fluctuations in a very subtle and effective way. The rhythm and strength of the torso movements are also important factors in expressing emotions. Fast and powerful movements may express anger, excitement or tension, while slow and soft movements may convey tranquility, calmness or melancholy.

5. Conclusion

In order to study the emotion conveying mechanism in dance, this paper constructs a dance action recognition model based on skeleton information to extract the dance action features in dance videos and recognize them, based on which the Attention-ConvLSTM classifier is introduced to measure the emotion conveyed by dance actions.

Compared with other action recognition models, the dance action recognition model based on skeleton information in this paper has the highest accuracy on the same dataset, reaching 88.34%, and has the least number of iterations, reaching the best result after only 40 iterations. The Attention-ConvLSTM-based dance movement emotion recognition model in this paper has the best results for the seven emotions on the MSRAction3D dataset, with a recognition rate of 98.95%.

Of the 8 emotions on the dancer sample dataset, the emotion Excited has a distance of 0 on its own. Among the eigenvalue velocities, the lowest similarity between it and the emotion Pleased is

0.531. Among the bone pair distances, the highest similarity between Excited and Pleased is 8.439. Among the bone pair pinch angles, the distance between Excited and Relaxed is the Shortest is 0.503.

For the feature parameter Velocity, the emotions Relaxed, Sad and Miserable are easily confused with each other. For the feature parameter Skeletal Pair Distance, the similarity between Excited, Relaxed and Happy is almost the same. For the feature parameter Skeleton Pair Angle, the emotions Mixed are easily confused with the Happy category of emotions.

The leg bone points have a greater influence on the expression of emotion. When expressing the emotion Miserable, the emotional similarity of Left arm, Left leg and Left foot were 62.37, 28.31 and 32.48, respectively, and the movement amplitude of the lower half of the body was smaller than that of the upper itself. When expressing the emotion Tired, the movement of the lower half of the limb is more expressive than than the upper half. When interpreting the emotion Mix, the expressive power of the limb FOOT (7.67) was greater. The dance emotion expression mechanism is constructed through the gesture transformation and cooperation of the head, hand, leg, waist, torso and other parts of the body.

References

[1] A. Aristidou, Q. Zeng, E. Stavrakis, K. Yin, D. Cohen-Or, Y. Chrysanthou, and B. Chen. Emotion control of unstructured dance movements. In *Proceedings of the ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, pages 1–10, 2017. https://doi.org/10.1145/3099564.3099566.

[2] M. Arsith and D. A. Popa Tanase. Nonverbal communication through dance. *Acta Universitatis Danubius. Communicatio*, 12(1):53–63, 2018.

[3] N. F. Bernardi, A. Bellemare-Pepin, and I. Peretz. Enhancement of pleasure during spontaneous dance. *Frontiers in Human Neuroscience*, 11:572, 2017. https://doi.org/10.3389/fnhum.2017.00572.

[4] B. Bläsing and E. Zimmermann. Dance is more than meets the eye—how can dance performance be made accessible for a non-sighted audience? *Frontiers in Psychology*, 12:643848, 2021. https://doi.org/10.3389/fpsyg.2021.643848.

[5] T. G. Borowski. How dance promotes the development of social and emotional competence. *Arts Education Policy Review*, 124(3):157–170, 2023. https://doi.org/10.1080/10632913.2021.1961109.

[6] D. Chen et al. Basic characteristics and body aesthetics analysis of modern dance. *Frontiers in Art Research*, 6(4), 2024. 10.25236/FAR.2024.060408.

[7] J. F. Christensen, L. Bruhn, E.-M. Schmidt, N. Bahmanian, S. H. Yazdi, F. Farahi, L. Sancho-Escanero, and W. Menninghaus. A 5-emotions stimuli set for emotion perception research with full-body dance movements. *Scientific Reports*, 13(1):8757, 2023. https://doi.org/10.1038/s41598-023-33656-4.

[8] D. Davenport. Dance is academic. *Journal of Dance Education*, 17(1):34–36, 2017. https://doi.org/10.1080/15290824.2016.1177642.

[9] B. A. Demiss and W. A. Elsaigh. Application of novel hybrid deep learning architectures combining convolutional neural networks (cnn) and recurrent neural networks (rnn): construction duration estimates prediction considering preconstruction uncertainties. *Engineering Research Express*, 6(3):032102, 2024. 10.1088/2631-8695/ad6ca7.

[10] U. Gawande, K. Hajari, Y. Golhar, and P. Fulzele. A novel gray wolf optimization-based key frame extraction method for video classification using convlstm. *Neural Computing and Applications*, 36(32):20355–20385, 2024. https://doi.org/10.1007/s00521-024-10266-3.

[11] M. Han. Systematic financial risk detection based on dtw dynamic algorithm and sensor network. *Measurement: Sensors*, 34:101257, 2024. https://doi.org/10.1016/j.measen.2024.101257.

[12] L. Keevallik. Vocalizations in dance classes teach body knowledge. *Linguistics Vanguard*, 7(s4):20200098, 2021. https://doi.org/10.1515/lingvan-2020-0098.

[13] M. Li. Analysis of dance art performance in dance education. *Art and Performance Letters*, 4(9), 2023. https://dx.doi.org/10.23977/artpl.2023.040915.

[14] Y. Lu. Analysis of body and emotion in dance performance. In *2021 Conference on Art and Design: Inheritance and Innovation (ADII 2021)*, pages 46–50. Atlantis Press, 2022. https://doi.org/10.2991/assehr.k.220205.008.

[15] J. Rae. Drawing the language of dance. *PAJ: A Journal of Performance and Art*, 40(2):42–45, 2018. https://doi.org/10.1162/pajj_a_00421.

[16] S. A. Savov. Dance narratology (sight, sound, motion and emotion). *TICS*:1332, 2017. 10.24308/iass-2014-139.

[17] K. Shejul, R. Harikrishnan, and H. Gupta. The improved integrated exponential smoothing based cnn-lstm algorithm to forecast the day ahead electricity price. *MethodsX*, 13:102923, 2024. https://doi.org/10.1016/j.mex.2024.102923.

[18] N. Sherman. Dancers and soldiers sharing the dance floor: emotional expression in dance. In *Social Aesthetics and Moral Judgment*, pages 121–138. Routledge, 2018.

[19] R. Smith and F. Pollick. The role of dance experience, visual processing strategies, and quantitative movement features in recognition of emotion from whole-body movements. In *Dance Data, Cognition, and Multimodal Communication*, pages 274–294. Routledge, 2022.

[20] M. G. Stutesman and T. R. Goldstein. Mechanisms for affect communication from dance: a mixed methods study. *The Journal of Creative Behavior*, 58(1):28–46, 2024. https://doi.org/10.1002/jocb.622.

[21] Q. Sun and X. Wu. A deep learning-based approach for emotional analysis of sports dance. *PeerJ Computer Science*, 9:e1441, 2023. https://doi.org/10.7717/peerj-cs.1441.

[22] M. Susino. Emotional expression, perception, and induction in music and dance: considering ecologically valid intentions. *The Journal of Creative Behavior*, 57(3):409–418, 2023. https://doi.org/10.1002/jocb.587.

[23] T. Tahsin, K. M. Mumenin, H. Akter, J. J. Tiang, and A.-A. Nahid. Machine learning-based stroke patient rehabilitation stage classification using kinect data. *Applied Sciences*, 14(15):6700, 2024. https://doi.org/10.3390/app14156700.

[24] M. Toppen. Using dance to promote sel skills. *Edutopia. George Lucas Educational Foundation, August*, 12, 2019.

[25] E. Van Dyck, B. Burger, and K. Orlandatou. The communication of emotions in dance. In *The Routledge Companion to Embodied Music Interaction*, pages 122–130. Routledge, 2017.

[26] X. Zhao. Emotion analysis and expression algorithm of dance action based on machine learning. *Journal of Electrical Systems*, 20(6s):1468–1481, 2024. http://dx.doi.org/10.52783/jes.3066.