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

Research on Diversified Inheritance Paths under the Analysis of Artistic Characteristics and Popular Elements of Guanzhong Drum and Gong Dance Based on Image Processing Algorithm

Sitong Chen[1], Jian Yang[1,✉]

[1] *College of Literature and Media, Xi'an FanYi University, Xi'an, Shaanxi, 710105, China*

ABSTRACT

In recent years, with the rapid development of artificial intelligence, big data, machine learning and other technologies, human society is entering a more and more intelligent society, and the interaction between humans and machines becomes more and more common. In this paper, image processing operations are added on the basis of Kinect's original acquisition of gong dance images, which reduces the influence of external light, background and other factors, and makes the human capture efficiency increase dramatically, and a spatio-temporal graph is constructed on the basis of the continuous human posture key point data, which describes the distribution of the human posture key points in different dataset types. Aiming at the problems existing in the traditional spatio-temporal map convolutional network, a multi-dimensional attention mechanism is designed to guide the model to reasonably allocate the weight resources in three dimensions: space, time and channel, respectively. Experiments are conducted on NTU-RGB+D, Kinect skeleton and Taiji datasets, respectively, which show that the AGCN-STC proposed in this paper has better recognition performance on all three datasets, and the recognition accuracy is improved by 0.9 percentage points compared with AM-GCN. Two actors are used as samples for visual measurement and quantitative analysis to compare the differences between the performance gestures of the two ornaments. Finally, based on the results of the study, we propose a transmission path for the Guanzhong gong dance, which is a reference for the cultural transmission of the Guanzhong gong dance.

Keywords: Kinect sensor, spatio-temporal map convolution, attention mechanism, cultural inheritance

1. Introduction

Guanzhong region of Shaanxi has a long history and rich cultural heritage, it is the ancient capital of the pre-Qin emperors and has a rich cultural heritage. In terms of dance, the folk dances in Guanzhong

✉ Corresponding author.
E-mail address: sachi12345@163.com (J. Yang).
Received 20 February 2024; Revised 15 May 2024; Accepted 20 November 2024; Published Online 16 April 2025.
DOI: 10.61091/jcmcc127b-360

region of Shaanxi can also be characterized as diverse and varied [1]. Guanzhong gong dance has a strong cultural atmosphere and a rich mass base in individual villages and towns of the local folk, which is of great historical significance for the cultural form of reflecting the national character of the Guanzhong Plain, but its popularity is not high, and there are individual phenomena of "heat in the village" [2]. It is necessary to fully explore the cultural value of this important intangible cultural heritage, especially the artistic value and sports value, for the inheritance of folk dance culture [3-5].

Because of its non-physical and living characteristics, folk dance itself has the potential to realize continuous transmission through innovation and transformation [6]. In the digital era, the traditional oral transcription and text-printing dissemination methods have been difficult to meet the needs of its development, and digital technology, as an emerging means, provides new possibilities for the protection, inheritance and dissemination of folk dances [7-9]. Through the use of image processing technology to analyze the artistic characteristics, performance forms and aesthetic value of Guanzhong Drum and Drum Dance at a deep level, the essence of the dance movements and performance forms of Guanzhong Drum and Drum Dance can be extracted from it [10-12]. On the basis of this, a reasonable creation was carried out by integrating popular elements, aiming to make the Guanzhong Drum and Drum Dance vividly present in front of people's eyes, and to ensure that people can have a deeper understanding of Chinese values, traditional aesthetics, and national mentality while watching the Guanzhong Drum and Drum Dance [13-16]. The diversified analysis and rational creation of Guanzhong Drum and Gong Dance under digital technology support will be highly adapted to the aesthetics of the modern public, and bring them a good aesthetic experience while drawing people's attention to the Guanzhong Drum and Gong Dance, which will enable the folk Guanzhong Drum and Gong Dance to be inherited and carried forward [17-20].

The research carries out a series of processing of Kinect captured images by smoothing filtering, graying, edge detection, binarization, morphological changes and other methods to improve the detection rate and recognition rate of images. From the perspective of constructing the sequence of human gesture keypoints in the spatio-temporal dimension as an action spatio-temporal graph, to address the problems of spatio-temporal modeling imbalance and poor temporal feature extraction ability of spatio-temporal graph convolutional network, adaptive temporal convolution algorithms and channel topology models are introduced, and the temporal channel aggregation graph convolutional network is designed to aggregate the skeleton temporal features of different temporal scales, so as to enhance the temporal feature extraction ability of the spatio-temporal graph convolution. The cultural inheritance path is proposed through experimental validation and visualization analysis of artistic features on the dataset and performers.

2. Kinect-based image acquisition and processing of gong dance art

2.1. Kinect-based image capture of gong dance art

2.1.1. Kinect Software Architecture. The Kinect 2.0 sensor can be designed as software based on Microsoft's NUI library for Kinect for Windows, combined with user-developed algorithmic programs and open source libraries such as OpenCV. The Kinect For Windows architecture is shown in Figure 1.

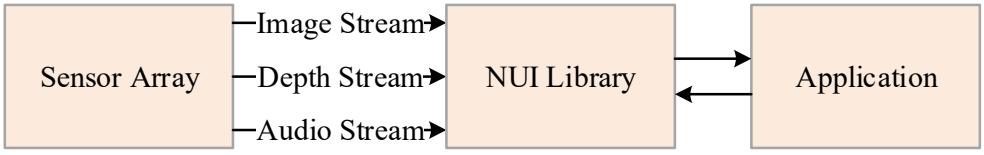


Fig. 1. Kinect For Windows architecture

The Kinect sensor transfers the image data, depth information and audio to the NUI library through the USB3-0 interface. The NUI library hides the complex hardware operations, and the user application

only needs to call the NUIAPI to access it, which includes several types of data such as color images, depth images, audio information and bone information. The flow of acquiring and processing the data source using the NUI library is shown in Figure 2.

Fig. 2. Data source acquisition and processing flow of Kinect 2.0 sensor

Sensor represents a specific hardware entity, an application can call one or more sensors (default is 1), and then call Source, which displays the metadata of the data source and provides a way for the reader to access it, and the sensor's depth frames, color frames, and audio sources will provide a data source for the user to call. The Reader provides access to frames, including the event mechanism (trigger model) and the polling mechanism "pull" model), a data source can have more than one Reader, the Reader can also be suspended. Frame is divided into FrameReferences and Frames, FrameReferences sends frame event parameters, including access to the actual frame using the method, the timestamp of the referenced frame and other specific information; Frames provide access to the frame data path (to establish a local copy of or direct access to the underlying buffer), including such as the frame format, width and height of frame containing frame metadata such as frame format, width, and height. In addition, Frames greatly reduces the time required to fetch frames, and users can obtain Data for different types of frames as needed.

2.1.2. Artistic image capture of the gong dance. From the Kinect2.0 sensor software architecture, the depth data frames are obtained using NUILi-brary. First of all, to obtain the depth data source, we need to open a Kinect2. 0 sensor hardware entity, which is of type IKinectSensor according to NUILibrary, define the variable mykinectsensor and assign it using the GetDefaultKinectSensor function, and call the mykinectsensor→ope to get a Kinect 2.0 sensor entity. Subsequently, the depth data source needs to be manipulated. The depth data source type is IDepthFrameSource, so the variable depthsource is defined, and the depth data source can be obtained by calling the get_DepthFrameSource member function to assign a value to it. In addition, you need to get the resolution of the depth image, use the IFrameDescription type and get_FrameDescription function to get the depth image resolution. After setting up the depth data source depthsource and obtaining the resolution, you need to define a variable of type IDepthFrameReader to call Reader, and utilize the member function depthsource→ OpenReader of Source to open Reader. after successfully opening Reader, you can call After successfully opening the Reader, you can call the depthreader→AcquireLatestFrame member function to collect the latest frame of depth data. If the value returned by AcquireLatestFrame is "S_OK", it means that the depth data of the latest frame has been successfully acquired, and it will be saved in the depthframe variable, and the type of the depthframe is IDepthFrame. Then, process the acquired depth data frame. When the processing of this data frame is finished, call depthframe→Release to release the depth data frame. If necessary, call AcquireLatestFrame function again to get a new depth data frame for processing.

2.2. Drum dance art image detection

In this paper, using image processing technology, the collected color images are subjected to a series of processing, and finally the collected images of gong dance are transformed into binarized black and white images, which reduces the influence of environmental factors such as light and background in the process of image acquisition on the detection results.

Image detection The color images acquired by Kinect are processed by smoothing filter, grayscaling, edge detection, image binarization, morphological transformation, etc., respectively, to remove the interference of noise in the environment and enhance the image, reducing the interference of lighting,

background and other factors on the image. Finally, skeletal point detection is performed on the processed image to draw the human skeleton map.

2.3. Image processing for gong dance art

2.3.1. Image Smoothing Filtering. Interference noise is generally randomly generated, irregularly distributed, and irregular in size. In space, the gray scale of the noise pixels is uncorrelated and significantly different from the nearby pixels. Therefore, denoising and recovering the image is a very important element in image processing, and this denoising process is called image smoothing or filtering.

Usually, the common method of image smoothing is a linear filter, which weights and sums the image pixels with the following formula:

$$g(i,j)=\sum_{k,l} f(i+k,j+l)h(k,l) \tag{1}$$

where $h(k,l)$ is called the kernel. The kernel is a mathematical square array composed of numerical parameters.

Currently in image processing, there are many methods of image smoothing, the main methods are mean filtering, median filtering, Gaussian filtering.

1) Mean value filtering. The process of mean value filtering uses a template kernel operator, with which all pixels in the image are covered to obtain a weighted average, and finally the value is used to replace the target pixel value. The calculation formula is as follows:

$$g(x,y)=\frac{\sum f(x,y)}{m} \tag{2}$$

where $g(x,y)$ is the filtered pixel value, m is the total number of pixel points in the neighborhood containing the point, and $\sum f(x,y)$ is the sum of all pixel points in the neighborhood.

2) Gaussian Filtering. The Gaussian filtering template is calculated by Gaussian function as shown in Eq:

$$G(x,y)=\frac{1}{2\pi\sigma^2}e^{\frac{x^2+y^2}{2\sigma^2}} \tag{3}$$

where (x,y) is the pixel point coordinates and σ is the standard deviation.

3) Median Filtering. Median filtering usually takes the odd number of data in the window, sorts the data by size and takes the middle position data as the processing result, which is a kind of nonlinear filtering.

When processing an image, the gray value of the pixels in the image area is sorted, and the middle value is selected as the target pixel point output, and the calculation formula is shown below:

$$g(x,y)=median\{f(x-i,y-j)\} \tag{4}$$

where $g(x,y)$ type output pixel value, $f(x-i,y-j)$ is the input pixel value, and (i,j) is the pixel in the template window.

2.3.2. Image Grayscaling. Currently, the images captured by Kinect are color images, mainly RGB-type images, which contain a lot of information that will reduce the speed of image processing and will not be used when performing image analysis. Therefore, in order to remove useless information from the image and improve the speed of the algorithm, people often grayscale the image in image processing.

2.3.3. Edge detection. There are five commonly used detection operators for edge detection, namely Robert operator, Sobel operator, Prewitt operator, Krisch operator, Laplace operator. Sobel operator contains two kernels, one with maximum effect on the vertical edges in the image and the other with maximum effect on the horizontal edges. Assuming that the image being acted upon is M, it is derived in two directions separately:

1) In the horizontal direction, M is convolved with a kernel of odd size. The kernel size is 3, and G_x is computed as:

$$G_x = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix} * M \tag{5}$$

2) Vertically, the same result is computed as horizontally, when the kernel size is 3. G_y is computed:

$$G_y = \begin{bmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix} * M \tag{6}$$

Combining equations (5) and (6), the result of convolution of pixel points in the image is:

$$G = \sqrt{G_x^2 + G_y^2} \tag{7}$$

It can also be simplified:

$$G = |G_x| + |G_y| \tag{8}$$

By performing edge detection on the image, the main features in the image such as people and objects are detected, which reduces the amount of computation for data processing.

2.3.4. Image binarization and the OTSU algorithm. Image binarization is widely used in image processing and basically all image processing processes require the use of this technique. Image binarization is defined as shown in equation (9):

$$G_{out}(x,y) = \begin{cases} 0 & G_{in}(x,y) \le T \\ 255 & G_{in}(x,y) > T \end{cases} \tag{9}$$

Where $G_{in}(x,y)$ is the pixel gray value, T is the set threshold, and $G_{out}(x,y)$ is the output gray value. The setting of the threshold value affects the final display of the image, and the commonly used method for setting the threshold value is the OTSU method, i.e., the maximum interclass variance method, which can also be called the Otsu method. The OTSU method is currently considered to be the best method for selecting the threshold value in image segmentation [21].

Principle of OTSU method: if the total number of pixels in the image is N and the gray scale range is $[0,\ldots,L-1]$, for a point pixel (x,y), the number of pixels of the corresponding gray level i is n, and the probability of the occurrence of each gray level is as follows in equation (10).

$$p_i = \frac{n_i}{N} \quad i = 0,1,2,3,\ldots\ldots,L-1 \tag{10}$$

For p_i there are:

$$\sum_{i=0}^{L-1} p_i = 1 \tag{11}$$

Then the image is quickly divided into two regions C_0 and C_1 using thresholding T, whereupon the mean value of the image is:

$$u_T = \sum_{i=0}^{L-1} i p_i \tag{12}$$

The mean values of the regions C_0 and C_1 are u_0 and u_1, respectively, and are computed as shown in equation (13) below.

$$\begin{aligned} u_0 &= \sum_{i=0}^{T} i p_i / w_0 \\ u_1 &= \sum_{i=T+1}^{L-1} i p_i / w_1 \end{aligned} \tag{13}$$

In the formula, $w_0 = \sum_{i=0}^{T} p_i, w_1 = \sum_{i=T+1}^{L-1} p_i = 1 - w_0$.

2.3.5. Image Morphology. Image morphology is commonly used to process binarized images for further improvement of image details. By extracting image shape regions and processing these regions, the image is made clearer and fuller in shape [22]. It is used to smooth ruggedness, fill holes, remove burrs, connect branches, etc. Morphological processing methods include: expansion operation, erosion operation, open operation, and closed operation.

1) Expansion and erosion. Expansion is mainly used to find the local maximum value of the image, and after expansion, the bright areas of the image become larger and the dark areas are reduced. The process of image expansion is realized by convolving the image with a given template, and the mathematical expression is as follows:

$$S = X \oplus B = \{x, y \mid B_{xy} \cap X \neq \varphi\} \tag{14}$$

where S is the output image, B is the expansion structure element which takes 0 or 1, and X denotes the binarized pixel set.

2) Open and closed operations. The open operation is usually used to deal with isolated dots, burrs, etc. in the image, which is a process of corrosion followed by expansion. After the open operation, the interference in the image can be significantly removed, in addition, the image area will not change significantly. The mathematical expression for the open operation is shown below:

$$S = X \cdot B = (X \otimes B) \oplus B \tag{15}$$

where S denotes the image after the open operation, B denotes the structural elements, each of which is 0 or 1, and X denotes the set of pixels after binarization.

3. Extraction of artistic movement characteristics of gong dance

3.1. Human Posture Map Sequence Construction

A human posture sequence is usually represented by the 2D or 3D spatial coordinates of each human joint for each video frame. An undirected spatio-temporal graph, which can be denoted as graph G, is reconstructed on a human gesture sequence that combines N human gesture joint points and T action time frames.

The graph G consists of a vertex set V and an edge set E, i.e., $G = (V, E)$, and the vertex-to-vertex relationship is represented by the adjacency matrix A, where $A \in R^{N \times N}$, N represents the number of nodes in a frame, and any two vertices in the adjacency matrix A $\{i, j\}$ if there is a spatially connected edge between the two vertices, then $A_{ij} = 1$, otherwise $A_{ij} = 0$. The input features of a human action recognition model usually have vertex features and edge features, described by X. Given a sequence of graphs GS with labels as shown in equation (16), G_n denotes the nth graph and y_n denotes the nth:

$$GS = \{(G_1, y_1), (G_2, y_2), \cdots, (G_{n-1}, y_{n-1}), (G_n, y_n)\} \tag{16}$$

Action Labeling, the action recognition task aims to learn a feature mapping relation F, the functional relation is shown in equation (17), which maps an input graph G to a set of labels, in which the most critical challenge is to extract different important action features from these input graphs.

$$y = F(G) \tag{17}$$

Combining a simple graph as a dynamic spatio-temporal graph over a sequence of human poses, where the vertex set $V = \{v_{ti} \mid t = 1, 2, \cdots T, i = 1, 2, \cdots, N\}$, which contains all human joints points in a sequence of human action samples poses, and v_{ti} denotes the i th human joint in the t th frame indicates that The GCN is modeled using this data, and the feature vector X includes the coordinate vectors of N joints on all T frames as well as the confidence of the key point prediction to construct the spatio-temporal graph on the human gesture sequence in two steps. Firstly, based on the connectivity of the human body structure, the joints within a frame are connected using edges as shown in Fig. 3, and then in consecutive time frames, each joint will be connected to the one that is the same as it, so that the connections in this setup can be defined naturally without the need to manually assign the body parts, and such an operation allows the feature extraction model to process input data with different styles.

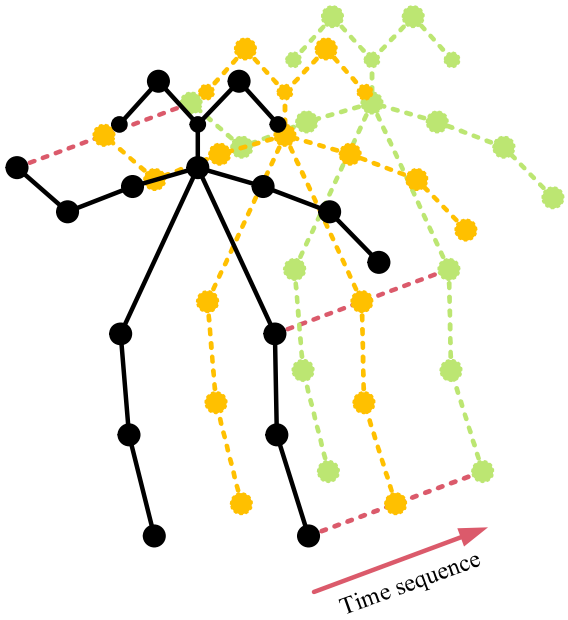


Fig. 3. Pose sequence spatio-temporal graph

3.2. Action feature extraction based on spatio-temporal graph convolution

3.2.1. ST-GCN Spatio-Temporal Graph Convolutional Network Modeling. The spatio-temporal graph convolution module mainly consists of 2 parts: spatial graph convolution and temporal convolution network. Considering spatial graph convolution, assuming that the number of keypoints in each frame of the input data is fixed, the number of edges formed by the natural connection of keypoints according to the human skeleton is constant, and if it is flattened into a two-dimensional grid, the spatial graph convolution kernel accomplishes the function similar to that of the traditional convolution kernel CNN function, which adopts the definition of the operation of traditional convolution kernel on the two-dimensional image, and for the given number of keypoints number of skeleton input data . Assuming that the size of the convolutional neural network convolution operator is $K \times K$ and the number of

channels is c, the output eigenvalue expression for a single channel at the spatial x position is shown in equation (18).

$$f_{out}(x)=\sum_{h=1w=1}^{K}\sum_{in}^{K}f_{in}(p(x,h,w))\cdot w(h,w) \tag{18}$$

Where: $p(x,h,w)$ is the sampling function, $p(x,w,h)=x+p^{'}(h,w)$, where h and w are the region indexes of the x position, and the sampling function p is used to enumerate neighboring nodes of x within the region position; and $w(h,w)$ is the weight function of the channel c.

In the set of neighboring nodes of a node, take a subset $B(v_{ti})=\{v_{tj}\mid d(v_{tj},v_{ti})\le D\}$ of the node with the same step size. B denotes the set of neighboring nodes of node v_i, and $d(v_{tj},v_{ti})$ denotes the minimum length from the two nodes between them, and in this paper we take the set of neighboring nodes whose distance from the joints is of length 1, i.e. $D=1$. Therefore the sampling function p can be rewritten as equation (19).

$$p(v_{ti},v_{tj})=v_{tj} \tag{19}$$

3.2.2. Adaptive data-driven graph convolution. In order to solve the problems of the fixed topology of the traditional spatio-temporal graph convolution network ST-GCN, this paper draws on 2SAGCN and changes the original predefined neighbor matrix to an adaptive neighbor matrix driven by data samples. The adaptive adjacency matrix can adaptively optimize the topology of the graph during the training process, modify the features of the adjacency matrix of the nodes, share the topology of the graph in multilayer graph convolution, and add the residual structure to connect different branches in order to ensure the stability of the original model.

In ST-GCN, actually the adjacency matrix A_k determines the topological form of the graph, and the mask matrix M_k denotes the strength of the connection between the root node and the neighbor nodes. In this paper, A_k and M_k are improved to realize the adaptive topology of the graph structure, and the theoretical calculation method of ACGN is shown in Equation (20).

$$f_{out}=\sum_{k}^{K_v}W_k f_{in}(B_k+\alpha C_k) \tag{20}$$

Introducing 2 different matrices, B_k and C_k, the adjacency matrix C_k, also known as the similarity matrix, is a localized graph associated with the data that expresses the similarity between two nodes in the graph. Let there are nodes v_i and nodes v_j on the skeleton structure, and the normalized Gaussian function representation of the connectivity relationship and the interaction strength between the two nodes is shown in Equation (21).

$$f(v_i,v_j)=\frac{e^{\theta(v_i)^T\phi(v_j)}}{\sum_{j=1}^{N}e^{\theta(v_i)^T\phi(v_j)}} \tag{21}$$

Where: N is the total number of nodes on the skeleton.

The interactive connectivity relationship between two nodes is computed using the dot product of vectors. Let the channel size of the input feature f_{in} of the network model be $C_{in}\times T\times N$, and use the embedding functions θ and ϕ in Eq. (21) to map the input feature map size to the size $C_e\times T\times N$. For the embedding function, two convolution kernels of size 1×1 are used in this paper to implement the embedding function θ and ϕ. The dimensions of the input feature maps are transformed, and the two convolution kernels output feature map dimensions of size $N\times C_eT$ and $C_eT\times N$, respectively, and the two output feature maps are subjected to matrix multiplication to eliminate the dimension C_eT, thus obtaining a similarity adjacency matrix C_k of dimension size $N\times N$, the value of the elements of which denotes the degree of similarity of its corresponding nodes v_i and v_j. The

normalization of the Gaussian embedding function in Eq. (21) is similar to the structural role of the SoftMax function, so in order to facilitate the implementation of the above function using the neural network layer, the SoftMax function can be used to compute the node's similarity matrix C_k, which is calculated as shown in Eq. (22).

$$C_k = SoftMax(f_{in}^T W_{\theta k}^{TK} W_{\phi k} f_{in}) \tag{22}$$

where: W_θ and W_ϕ denote the parameters of the embedding functions θ and ϕ respectively.

3.2.3. Multidimensional attention mechanisms. From the perspective of spatio-temporal channel, the spatial feature attention enhancement module, the temporal feature attention enhancement module and the channel feature attention enhancement module are established respectively. The model is guided to pay attention to the focused spatio-temporal channel features to reduce the influence of redundant features. The output feature maps of each module are multiplied with the input feature maps to enhance the attention on the basis of the original feature maps.

Spatial Feature Attention Enhancement Module: this module gives different attentional intensities to different key points in spatial perspective to enhance the feature representation of action-related joint features in training and learning in the model, which is calculated as shown in equation (23).

$$M_s = \sigma(g_s(\text{AvgPool}_t(X))) \tag{23}$$

Where: σ is a nonlinear transformation function, this section uses the Sigmoid activation function to accomplish this, the curve of the Sigmoid function is a kind of S-type growth curve, which can map the input variables between (0, 1), and its curve has the characteristics of smooth and easy to derive, which is more suitable for the adaptive graphical convolutional network with data-driven proposed in this paper, and it can strengthen the attention mechanism in the The role of resource reallocation. The s in g_s represents the spatial angle, and g_s denotes the extraction of features in the spatial angle.

Temporal Feature Attention Enhancement Module: the function of this module is similar to the SAM module, which guides the model to enhance the attention to local action time segments during training, while suppressing the influence of irrelevant time segments to improve the efficiency of training and learning, which is computed as shown in Eq. (24).

$$M_t = \sigma(g_t(\text{AvgPool}_s(X))) \tag{24}$$

Where: $M_t \in R^{1\times T\times 1}$ is the output of the temporal feature attention enhancement module, which is input to the next module after matrix dot-multiplication and summation with the output of the previous module; AvgPools denotes the averaging pooling operation of the input feature maps in the spatial dimension; g_t is the convolution operation of the pooled feature maps in the temporal dimension. Convolution operation is performed on the pooled feature maps in the time dimension; similar to the SAM module, this function is implemented using one-dimensional convolution.

Channel Feature Attention Enhancement Module:

The CAM module emphasizes that the data features of different channels have different semantic levels in judging the action categories of different data samples, and enhances the feature discriminative ability of the channels. The computation method of CAM is shown in Equation (25);

$$M_c = \sigma(g_{c2}(\delta(g_{c1}(AvgPool_{st}(X))))) \tag{25}$$

where $M_c \in R^{C\times 1\times 1}$ is the output feature map of the CAM module, which is output again after performing the dot-multiplication summation operation with the inputs of the module, and $AvgPool_{st}$ performs the average pooling operation in both spatial and temporal dimensions for the dimensional transformations of the input feature map, and the CAM module's g_{c1} and g_{c2} are different from the one-dimensional convolution of the above modules; in the CAM module, g_{c1} and g_{c2} are linear functions implemented using a fully-connected layer, and their main role is to perform the extraction

of the features along the channel dimensions; and the δ function is implemented using the ReLU activation function.

3.3. Experimentation and Analysis

3.3.1. Experimental environment. The experimental environment configuration is shown in Table 1.

Table 1. Experimental environment configuration table

Hardware and hardware environment	Parameter
Processor	Intel(R)Xeon(R)Silver4110CPU@2.10GHz
Memory	32GB
GPU	NVIDIA RTX 2080Ti
Operating system	Ubuntu18.04
Programming language	Python
Training framework	Pytorch1.11.0

3.3.2. Experimental data sets:

1) NTU RGB+D is a large-scale dataset for skeleton based human action recognition research, which is captured by three Kinect V2 cameras with different viewpoints at the same time. The dataset contains multiple different forms of RGB videos, depth map sequences, 3D skeleton data, and infrared sequences of each action sample in order to help the researchers to better understand the human actions.

2) Kinetics Skeleton 400: Kinetic is a large-scale, high-quality human movement video dataset containing 300,000 human movement videos divided into 400 categories.

3) Causeway dance movement dataset. Due to the specificity of the gong dance action, the object of study in this topic, in human action analysis, there is no large-scale gong dance action skeleton sequence dataset for experimental validation. Therefore, this topic utilizes the overall algorithm model (AGCN-STC) proposed in Chapter 3, which integrates adaptive data-driven graph convolution and multi-dimensional attention mechanism, to extract the skeleton sequence of gong dance movements, and the constructed gong dance movement dataset Taiji.

3.3.3. Evaluation indicators. In order to evaluate the performance of model training, the features are classified using softmax layer, and the accuracy of classification is used as an evaluation metric to measure the performance of the model.

The accuracy rate, also known as the correct rate, is the number of total samples for which the correct result is predicted, and is calculated as:

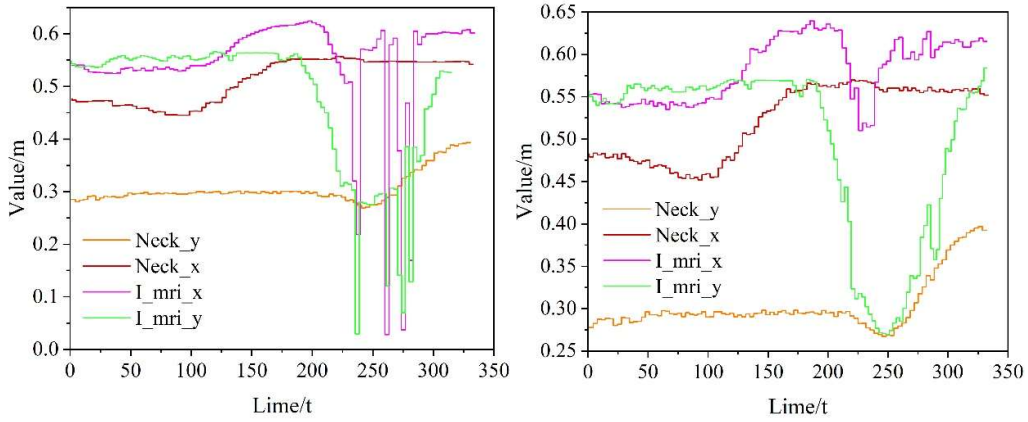
$$accuracy = \frac{TP + TN}{TP + FN + FP + FN} \tag{26}$$

Where TP is all the samples in which the positive examples in the prediction result are positive examples in actuality; TN is all the samples in which the positive examples in the prediction result are negative examples in actuality; FP is all the samples in which the positive examples in the prediction result are positive examples in actuality; and FN is all the samples in which the negative examples in the prediction result are negative examples in actuality.

TOP-1 Accuracy and TOP-5 Accuracy, TOP-1 Accuracy indicates the accuracy of the category with the first sample data volume in the sequence that matches the actual test; TOP-5 Accuracy indicates that the category with the top 5 sample data volume contains values for which the actual result is accurate. It is used for comprehensive evaluation of statistical sets with a large number of categories.

3.3.4. Data pre-processing. Since the skeletal data is the data estimated by the pose estimation algorithm, and the algorithm will have estimation error in some abnormal poses or severe occlusion.

This leads to a situation where the joint position coordinates are missing data is missing. These missing position data must be filled in to ensure that the feature vector maintains a fixed dimension in the subsequent feature extraction process. In the human frame extraction process, the fluctuation of the skeleton position information in the video stream is large, and the phenomenon of randomly losing the position data of a certain skeletal joint may occur in the continuous extraction process, especially the joints at the larger displacements. To address this problem, the missing joint data of the current frame is filled according to the gradient interpolation between the action frames before and after the missing joint. The calculation method for filling the missing skeletal joint data is shown in Equations (27) and (28), where n is the current frame position and m is the number of missing frames. The change of joint position after data filling is shown in Fig. 4, and the change of x and y direction after wrist data filling is in line with the trend of movement change.



(a) Data populate anterior bone joint data (b) Packed posterior bone joint data

Fig. 4. The skeleton is filled before and after the bone

$$x_{ti} = x_{(t-1)i} + n\frac{|x_{(t-1)i} - x_{(t+m)i}|}{m} \tag{27}$$

$$y_{ti} = y_{(t-1)i} + n\frac{|y_{(t-1)i} - y_{(t+m)i}|}{m} \tag{28}$$

The scale of the human body in different videos may be very different due to the differences in height and shooting distance. Therefore scale normalization of skeletal data is required, otherwise two similar actions may produce different results. Therefore the data is normalized through a batch normalization layer at the data input layer.

3.3.5. Experimental results and analysis:

1) Effectiveness of the attention mechanism. The experimental results on NTU-RGB+D (X-Sub) and Kinects skeleton (TOP-5) are shown in Table 2. From the experiments, it can be seen that the adaptive data-driven graph convolution algorithm facilitates action recognition. The recognition accuracy of AGCN-STC on NTU-RGB+D and Kinects skeleton is 90.88% and 61.38%, respectively, which is 0.9 percentage points higher compared to AM-GCN. The experimental results verify the effectiveness of the algorithmic model that incorporates adaptive data-driven graph convolution and multi-dimensional attention mechanism, the effectiveness of the algorithmic model.

Table 2. Recognition accuracy of different type of attention matrix

Method	NTU-RGB+D(%)	Kinects skeleton(%)
ST-GCN	83.28	54.58

ST-GCN(No:M)	82.08	54.08
AM-GCN	89.98	59.18
AM-GCN(No:A)	89.58	58.88
AM-GCN(No:B)	89.38	58.68
AM-GCN(No:C)	89.48	58.48
AGCN-STC	90.88	61.38

2) Number of parameters. Replacing the TCN network with the M-TCN network captures more temporal information while reducing the number of parameters of the model, making the model easier for subsequent deployment. The results of the experiments conducted on X-Sub in NTU-RGB+D are shown in Table 3. The experiments show that the number of parameters and recognition accuracy in the network increases with different attention mechanisms. The introduction of multi-scale time map convolution makes the amount of parameters decrease while the recognition accuracy of the network is improved to some extent.

Table 3. The parameter quantity and recognition accuracy of the multiscale time volume

Method	Parameter quantity(M)	Accuracy rate(%)
ST-GCN	4.1	82.3
AS-GCN	4.4	87.9
2S-AGCN	4.4	89.4
NSA-GCN	7.1	88.7
GNN(MS-TCN)	0.9	89.1
MS-G3D(T9)	4.5	87.1
MS-G3D(MS-TCN)	2.3	87.8
AM-GCN(T9)	4.3	88.5
AM-GCN(MS-TCN)	2.1	89.3
AGCN-STC(M-TCN)	1.8	90.5

3) Ablation experiments. In order to measure the performance of the proposed AGCN-STC network, experiments are conducted on three datasets, NTU-RGB+D, Kinect skeleton and Taiji, respectively, to compare with the existing methods of action recognition based on skeleton data, and the experimental results are shown in Tables 4, 5 and 6.2s-AGCN, NSA-GCN, Dynamic GCN, and MS-G3D use dual-stream data with skeletal joint point information and skeletal edge information in the experiments, and the experimental results show that the addition of skeletal edge information is beneficial to improve the recognition accuracy of the network. The algorithm model proposed in this paper only considers the skeletal joint point information, and the experiments on the three datasets show that the recognition accuracy of the AGCN-STC network is substantially improved.

Table 4. Different methods of identification accuracy on NTU-RGB+D

Method	X-Sub(%)	X-View(%)
Deep LSTM	62.7	69.4
ARRN-LSTM	82.7	89.9
ST-GCN	83.4	89.2
AS-GCN	88.7	96.2
2s-AGCN	90.2	97.2
Dynamic GCN	92.6	97.0
GNN	91.5	97.0
MS-G3D	92.5	97.5
AGCN-STC	90.8	96.9

Table 5. The recognition accuracy of different methods in Kinect skeleton

Method	TOP-1(%)	TOP-5(%)
Deep LSTM	18.18	37.08
ARRN-LSTM	22.08	41.78
ST-GCN	32.48	54.58
AS-GCN	36.58	58.28
2s-AGCN	37.88	60.48
Dynamic GCN	38.88	61.88
GNN	39.68	63.08
MS-G3D	34.88	56.98
AGCN-STC	39.78	62.68

Table 6. The identification accuracy of different methods on Taiji

Method	TOP-1(%)	TOP-5(%)
ST-GCN	34.18	56.48
AS-GCN	37.58	59.18
2S-AGCN	38.98	61.88
MS-G3D	41.38	65.08
AGCN-STC	39.28	63.28

4. Analysis of the artistic characteristics and popular elements of the gong dance

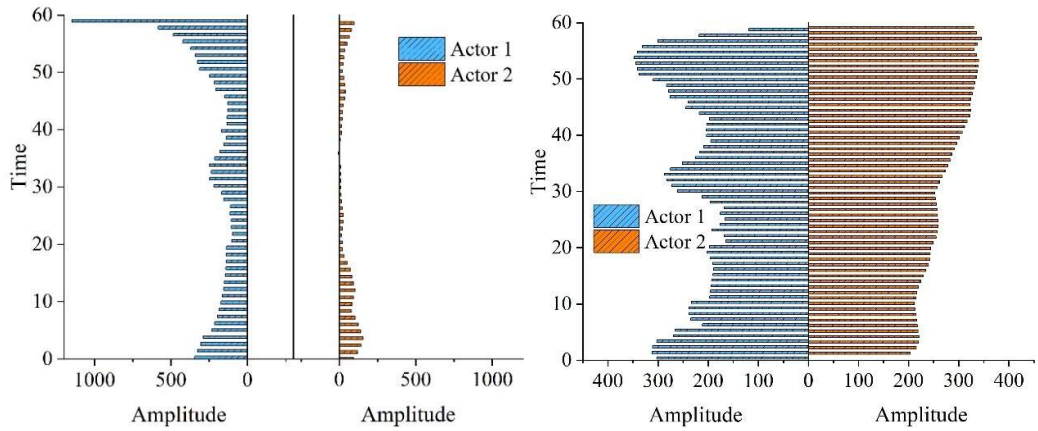
4.1. Characterization of gong dance gestures

4.1.1. Body gestures express hypocritical characteristics. By calculating the variance (σ) data in the X-axis and Y-axis columns of the table, the fluctuation of the relevant data can be obtained: the larger the value of the variance, the greater the fluctuation of the data; the smaller the value of the variance, the smaller the fluctuation of the data (later on, the limb dynamics motion curves and the data form to obtain the operating procedures are the same as this). According to the above operation procedure, the movement amplitude value of dance shoulder amplitude σ in the dance of two actors is obtained as shown in Fig. 5 and the data of shoulder amplitude range value as shown in Table 7.

From the values of σ and amplitude range values in the table and figure, it can be seen that the horizontal amplitude of the dance shoulders of Actor 1 σ is 46.044, with an amplitude range of 176.00626, and the vertical amplitude of the shoulders σ is 45.274, with an amplitude range of 144.0746.The change in the amplitude of the shoulders of the actor is both horizontally and vertical, show a relatively stable trend, so the actor's inner emotions are relatively flat. The horizontal amplitude of actor 2's shoulders σ is 141.172, with an amplitude range of 1065.78536; the vertical amplitude of the shoulders σ is 55.875, with an amplitude range of 284.39574. The changes in the amplitude of the actor's shoulders in the horizontal and vertical directions are more intense, which shows that the actor's heart is rich in emotion, and that he is restrained at the same time, and that the Guanzhong Drum and Gong Dance has been performed by the actor's heart. At the same time, it presents the crazy side of Guanzhong Drum and Gong Dance, which is more expressive.

Table 7. Actor shoulder amplitude motion data table

1-x(min)	1-x(max)	Amplitude range	2-x(min)	2 x(max)	Amplitude range
86.19134	1151.9767	1065.78536	-10.85786	165.1484	176.00626
1-y(min)	1-y(max)	Amplitude range	2-y(min)	2--y(max)	Amplitude range
63.66146	348.0572	284.39574	206.7684	345.843	144.0746



(a) Lateral amplitude (b) Longitudinal amplitude

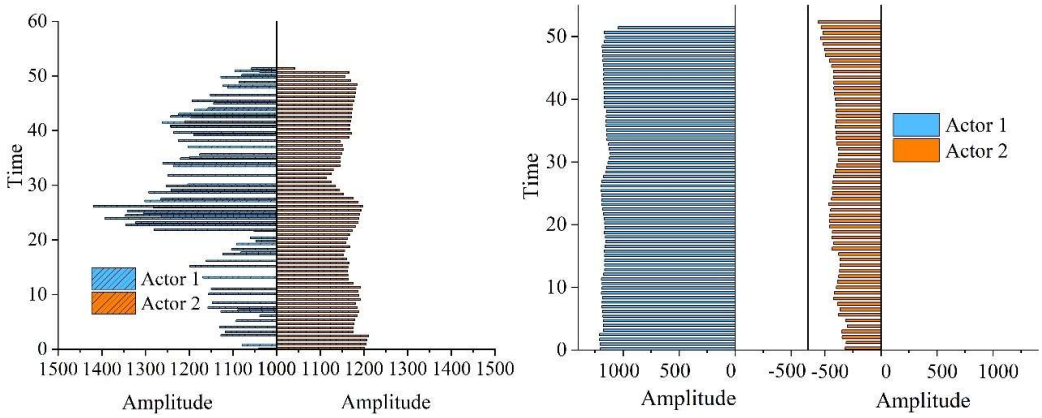
Fig. 5. Actor motion amplitude

4.1.2. "Static" and "dynamic" gestures combine to convey grief. In order to show the realism of the actor's dance mood, the body movements of the two actors have different degrees of dynamic embodiment, through the above operation program on obtaining the value of the amplitude of the dynamic movement of the singers' limbs, to obtain the value of the amplitude of the movement of the two actors' body swaying σ shown in Fig. 6 and the amplitude range of the value of the data table as shown in Table 8.

As can be seen from the values of σ and the values of amplitude ranges in the table and the figure, the horizontal amplitude of the body of Actor 1 σ is 79.157 with an amplitude range of 411.874, and the vertical amplitude of the body of Actor 1 σ is 54.375 with an amplitude range of 298.9178, while the horizontal amplitude of the body of Actor 2 σ is 22.892, with an amplitude range of 135.173; and the longitudinal amplitude of the body σ is 17.154, with an amplitude range of 79.8573. It can be seen that Actor 1's body in the physical space has a higher value of maximum movement than that of Actor 2, whose inner psychological emotions are expressed through the external activities, and also demonstrates that the actor has a high degree of mastery of the dance and lyrics. It also shows that the actors have a high degree of grasp of the content of the dance and the lyrics of the song, so as to achieve a real, natural and rich inner experience of the character state.

Table 8. The range of the actor's range

1-x(min)	1-x(max)	Amplitude range	2-x(min)	2 x(max)	Amplitude range
1009.218	1421.092	411.874	1089.98	1225.153	135.173
1-y(min)	1-y(max)	Amplitude range	2-y(min)	2--y(max)	Amplitude range
-566.9178	-268.4672	298.9178	-857.374	-777.5167	79.8573



(a) Lateral amplitude (b) Longitudinal amplitude

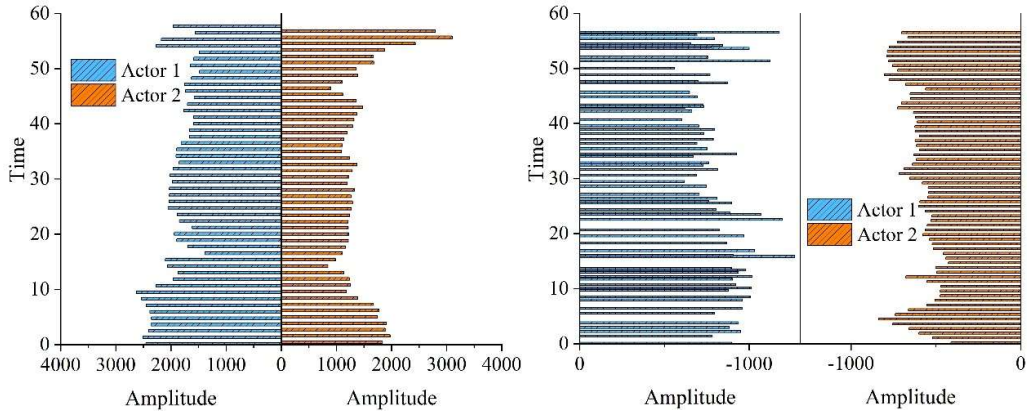
Fig. 6. Actor motion amplitude

4.1.3. Voice and body coexist in an exuberant gesture. Through the above operational procedures on the acquisition of the dynamic movement curves of the actor's limbs, the movement amplitude values of the two singing versions of the body swaying σ are obtained as shown in Fig. 7 and the data table of amplitude range values as shown in Table 9.

From the σ and range values of the actor's body amplitude changes in the table and figure, it can be seen that in Actor 1, the σ of the horizontal body amplitude is 324.654, and the range of the amplitude is 1,387.363, and that the σ of the longitudinal body amplitude is 145.112, and the range of the amplitude is 719.348, while the σ of the horizontal body amplitude of Actor 2's body is 409.124, with a range of 3301.795, and the vertical amplitude of Actor 2's body is 106.073, with a range of 487.9785. By way of comparison, it can be seen that Actor 1's change in body amplitude is more prominent vertically, whereas the change in Actor 2's body amplitude is more drastic horizontally. The performance of Actor 1's body amplitude change is more prominent in the vertical direction, while the change of body amplitude after Actor 2's night is more intense in the horizontal direction. In both versions, the music and body movements, as well as the specific gestures, are bound together in a synchronized relationship, with the sound wandering around the performers' bodies, making the characters, language, and plot vivid, while the gestures guide the audience's attention and make the state of the Guanzhong Drum and Drum Dance “externalized” through the language of the body.

Table 9. The range of the actor's range

1-x(min)	1-x(max)	Amplitude range	2-x(min)	2 x(max)	Amplitude range
1300.11	2687.473	1387.363	719.348	4021.143	3301.795
1-y(min)	1-y(max)	Amplitude range	2-y(min)	2--y(max)	Amplitude range
-1297.072	-567.7375	729.3345	-898.7739	-410.7954	487.9785



(a) Lateral amplitude (b) Longitudinal amplitude

Fig. 7. Actor motion amplitude

4.2. Incorporation of Romantic Music Elements in the Art of Drum Dance

The basic tempo of this gong dance is = 41.9, and the piece is divided into two parts. As shown in Figure 8, the second part of the piece (bars 39-79) has two large undulations compared to the first part (bars 1-38). Referring again to the full score of this gong dance lyric tune12, the second part is marked by significantly more expression, tempo, and other markings, suggesting an increasing tension in the tone

of this passage. Predictably, in this era of Romanticism, the expression terms and tempo markings left by the actors in the gong dance leitmotifs were no longer sufficient for the detailing required for the performance, and these markings were only an approximate basis for judging tempo.

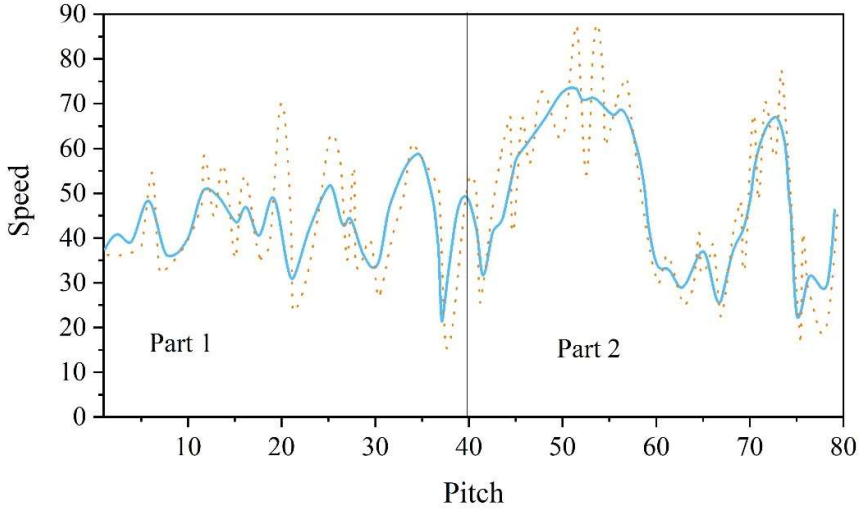


Fig. 8. The speed of the melody

4.3. Multiple Inheritance Paths of Causeway Dance

1) Establishment of a demonstration area for the living heritage of the Guanzhong gong dance. The establishment of the Guanzhong Drum and Drum Dance Inheritance Demonstration Area is to make it "active" as a whole through scientific, systematic and holistic protection based on the living inheritance characteristics of the Guanzhong Drum and Drum Dance, and to place it in its original mode of production and way of life as far as possible under the well-built ecological, humanistic and natural environments, and to make it "active" as a whole under the infiltration of the strong folk customs and folk styles. The overall "activity" of the dance is protected. As the inheritance of non-heritage dance is a form of fluid change and development in the world, that is, what we call "living nature", so it has a strong dependence on the survival of the social environment, natural environment, cultural ecology. At the present stage of China's non-heritage dance research and development, from the protective inheritance of intangible culture, towards the protection of the development and dissemination of innovation direction change.

2) Establishing the main person responsible for the inheritance of Guanzhong Drum and Drum Dance. Because of the evaluation and recognition mechanism of cultural organizations and institutions for the inheritors, there is a certain difference and lag between the acceptance and inheritance of Guanzhong Drum and Drum Dance by the general public and the self-conscious behavior of the public. Compared with other intangible cultural heritages, which have a complete system of inheritors, masterpiece lists, and inheritance institutions, the establishment of an inheritance protection system for the Guanzhong Drum and Drum Dance is more urgent. In the current situation, the cultural department should focus on cultivating the Guanzhong Drum and Drum Dance inheritors and establishing inheritance groups and organizations, setting up inheritance models, and increasing the cultivation of folk inheritors, improving their cultural literacy and artistic cognitive level through professional learning and training, and providing them with practical and powerful support at both the material and spiritual levels.

5. Conclusion

In this paper, based on the image acquisition of gong dance art, the images are processed. Secondly, for the influence of redundant features of irrelevant nodes, a multidimensional attention mechanism is

proposed to direct the model to focus on important joint, temporal, and channel features in the skeleton sequence, and to reduce the influence of spatio-temporal features of redundant nodes. Validation is carried out through comparative experiments, in which the AGCN-STC algorithm model proposed in this paper is experimented on NTU RGB+D, Kinect skeleton 400 and Taiji datasets, and the results show that the recognition accuracies of AGCN-STC on NTU-RGB+D, Kinects skeleton are 90.88%, 90.88%, and 61.38%, which is better than other models. Then comparing the visualization analysis and quantitative analysis of two gong dance performers, the performers' dances and words and tones can reflect the performers' inner emotions, as well as the perfect integration and embodiment of the contemporary popular romanticism.

Funding

This work was supported by Shaanxi Province Art Science Planning and Research Project, the project name is "Research on the Artistic Characteristics and Diversified Inheritance Path of "Gong Inspiration in Guanzhong" Based on Modern Popular Elements" (project number: SYE2024002).

References

[1] Liu, S. (2020). The Chinese dance: a mirror of cultural representations. Research in Dance Education, 21(2), 153-168.

[2] Tan, B. (2021). Thoughts on the Inheritance of Shaanxi Folk Dance in Universities. The Educational Review, USA, 5(7).

[3] Zdravkova-Djeparoska, S. (2018). FOLKLORE, DANCE IN THE CONTEXT OF MODELING IDEOLOGICAL MESSAGES. Res Humanitariae, 24.

[4] Jiang, L. (2020). Folk art inheritance and its value in art design education. Frontiers in Educational Research, 3(5).

[5] Gao, H. (2017, July). Research on the Cultural Connotation of Yangge in Northern Shaanxi. In 2017 3rd International Conference on Economics, Social Science, Arts, Education and Management Engineering (ESSAEME 2017). Atlantis Press.

[6] Yang, H. (2024). The Fusion of Traditional and Modern Dance: Embracing Cultural Diversity. Frontiers in Art Research, 6(1).

[7] Xie, L. (2023). Rural Folk Dance Movement Recognition Based on an Improved MCM‐SVM Model in Wireless Sensing Environment. Journal of Sensors, 2023(1), 9213689.

[8] Miao, Z., Wang, W., Xie, J., Ma, L., & Hu, N. (2025). Research on original environment folk dance movement evaluation based on spatio-temporal graph convolutional networks. Signal, Image and Video Processing, 19(3), 266.

[9] Li, J., Wu, S. R., Zhang, X., Luo, T. J., Li, R., Zhao, Y., ... & Peng, H. (2023). Cross-subject aesthetic preference recognition of Chinese dance posture using EEG. Cognitive Neurodynamics, 17(2), 311-329.

[10] Grammalidis, N., Kico, I., & Liarokapis, F. (2020). Analysis of Human Motion based on AI technologies: Applications for safeguarding folk dance performances. In Strategic Innovative Marketing and Tourism: 8th ICSIMAT, Northern Aegean, Greece, 2019 (pp. 321-329). Cham: Springer International Publishing.

[11] Protopapadakis, E., Grammatikopoulou, A., Doulamis, A., & Grammalidis, N. (2017). Folk dance pattern recognition over depth images acquired via kinect sensor. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 42, 587-593.

[12] Zhang, N. (2025). A deep learning-based method for identifying differences in folk dance movements.

Journal of Computational Methods in Sciences and Engineering, 14727978241312996.

[13] Pang, Y., & Niu, Y. (2023). Dance video motion recognition based on computer vision and image processing. Applied Artificial Intelligence, 37(1), 2226962.

[14] Yang, T., & Ismail, M. J. (2023, September). Research and Analysis on the Aesthetics of Folk Dance Based on the Background of Big Data. In 2023 4th International Conference on Big Data and Informatization Education (ICBDIE 2023) (pp. 507-514). Atlantis Press.

[15] Kico, I., & Liarokapis, F. (2020). Investigating the learning process of folk dances using mobile augmented reality. Applied Sciences, 10(2), 599.

[16] Liang, Y., & Pan, F. (2024). Interactive experience design of traditional dance in new media era based on action detection. Computer-Aided Design and Applications, 21(S7), 241-255.

[17] Gao, X., & Wang, Y. (2020). Optimized integration of traditional folk culture based on DSOM-FCM. Personal and Ubiquitous Computing, 24(2), 273-286.

[18] Rallis, I., Voulodimos, A., Bakalos, N., Protopapadakis, E., Doulamis, N., & Doulamis, A. (2020). Machine learning for intangible cultural heritage: a review of techniques on dance analysis. Visual Computing for Cultural Heritage, 103-119.

[19] Shi, Y., & Han, S. (2025). Multimedia interactive creative dance choreography integrating intelligent chaotic art algorithms. Journal of Computational Methods in Sciences and Engineering, 14727978251318055.

[20] Zhang, L. (2023). Intelligent action recognition and dance movement optimization based on multi-threshold image segmentation. International Journal of System Assurance Engineering and Management, 1-12.

[21] Guohua Zhai,Yabin Liang,Zhisen Tan & Sirui Wang. (2024). Development of an iterative Otsu method for vision-based structural displacement measurement under low-light conditions. Measurement,226,114182-.

[22] Akhil Masurkar,Rohin Daruwala & Varsha Turkar. (2024). Multi-Frequency Analysis of Fully Polarimetric Speckle Filter Based on Image Morphology for Land Use Classification. Journal of the Indian Society of Remote Sensing,52(12),1-25.