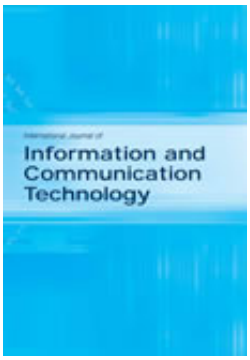
INDERSCIENCE PUBLISHERS
Linking academia, business and industry through research



International Journal of Information and Communication Technology

ISSN online: 1741-8070 - ISSN print: 1466-6642
https://www.inderscience.com/ijict

Digital dance generation and application based on hybrid density network

Qian Lu

Article History:

Received:	08 December 2024
Last revised:	18 December 2024
Accepted:	18 December 2024
Published online:	22 January 2025

Digital dance generation and application based on hybrid density network

Qian Lu

Conservatory of Music,
Huanggang Normal University,
Huanggang 438000, China
Email: 17798387177@163.com

Abstract: This article proposes a digital dance generation method based on mixture density network (MDN), aiming to effectively capture and generate complex dance action sequences. Firstly, we analysed the temporal dependencies and diverse features of dance movements, and designed a multimodal temporal generation framework using MDN and long short-term memory (LSTM) networks to capture dynamic correlations and pose changes between dance movements. This framework can generate action sequences that match the music style when inputting music or rhythm information, with high continuity, coordination, and naturalness. This paper assesses the generated dance motions by the model using a publicly available dance dataset, and verified the effectiveness of this method through subjective and objective quantitative indicators. The experimental results show that compared to traditional generative models, the MDN based model has improved the fluency, naturalness, and diversity of generated actions.

Keywords: deep learning; computer music choreography; feature extraction; action filtering.

Reference to this paper should be made as follows: Lu, Q. (2025) 'Digital dance generation and application based on hybrid density network', *Int. J. Information and Communication Technology*, Vol. 26, No. 2, pp.51–66.

Biographical notes: Qian Lu received her PhD from Silliman University in 2023. She is currently a lecturer at the Music College of Huanggang Normal University. Her research interests include dance teaching, AI dance choreography and AI digital animation.

1 Introduction

The complex mapping relationship between music and dance is manifested at the abstract algorithmic level as the synchronisation of the spatial patterns of body movements involved in dance and the time series patterns in music rhythm. Therefore, generating dance movements based on music involves the problem of cross modal transformation. For humans, cross modal imagination is the foundation of human brain creativity and an important feature that distinguishes the brain from computers (Henrickson, 2020). However, for computers, achieving this 'creativity' in computer vision (CV) systems is a major challenge, primarily because the heterogeneity differences between different modal data are difficult to measure (Baltrušaitis et al., 2018). Heterogeneity difference refers to

the similarity in content between different modal data, which is manifested in the synchronisation of dance action sequences and music beats in music based dance action generation tasks. Secondly, dance is a subjective art form, which makes it difficult to perform computational modelling and evaluation of dance movement generation tasks. At last, the dance's choreography should represent the content of the music that complements the dance rather than merely pursue its artistic value. Whereas artistic quality depends on the dance to have authenticity and variation, content of the music necessitates the dance to retain continuity with the music style. Therefore, developing dance motions based on music using computers is rather challenging, which has since been one of the main focus of study in the fields of computer vision and computer graphics.

Digital dance generation has progressively evolved as artificial intelligence and computer vision technology advance, one of the research hotspots. Digital content creation technology has been extensively used recently in domains such virtual reality (VR), augmented reality (AR), electronic games, and digital art, thereby imitating human behaviours and emotional expression to create extremely immersive and interactive digital experiences. Apart from producing several dance moves, dance generation technology can satisfy individual needs in social media, entertainment, and education as well as in industry. Nonetheless, the creation of dance movement sequences still presents several difficulties since dance movements have great continuity and complicated temporal relations and demand great coordination and authenticity of motions. Effective capturing of this complexity is challenging for conventional generating models.

Particularly in the application of generative adversarial networks (GANs) (Creswell et al., 2018) and variational autoencoders (VAEs) (Vahdat and Kautz, 2020), which have greatly enhanced the diversity and authenticity of generating tasks, the most recent developments in deep learning have given new technological means for dance generation. Nevertheless, the particular criteria for dance generation challenge conventional generation models since they demand models to effectively capture the continuity, fluency, and multimodal characteristics of dance movements. The present studies mostly concentrate on the creation of a single pattern, missing enough mining and modelling of complicated time series data and diverse action sequences, which limits the further enhancement of the generating impact.

Regarding application, dance generation techniques grounded on MDN offer great possibilities. First of all, this approach can give consumers real-time and highly interactive dance experiences in virtual reality (VR) and augmented reality (AR) settings fit for virtual dance performances, virtual idol stages, and other scenarios. Second, automated dance generation can lower creative expenses, increase the variety and originality of produced material, and give players individualised dancing interactive experiences in digital entertainment and game development. Using created dance motions to give novices demonstrations, help them in understanding and copying movements, and increase dance learning efficiency, this approach may also be used in dance instruction and auxiliary activities in the field of education.

By incorporating mixed density networks and temporal modelling approaches, this paper suggests a fresh and efficient generating method for digital dance, so offering new technological assistance for domains such virtual reality, digital entertainment, and education. We intend to improve the model's generating accuracy going forward, increase its flexibility in other dance forms, and investigate its expanded uses in other artistic disciplines.

Deep learning methods' ongoing development in the field of dance generation research has given fresh concepts for digital dance generation. Many researchers have recently produced notable findings in VAEs, GANs, and deep learning approaches grounded on sequence modelling.

First of all, generative adversarial networks find extensive use in dance generation. Wei and Mahmood (2020), for instance, proposed a GAN based generating technique whereby a generator creates dance sequences and a discriminator evaluates the authenticity of the dance. In a similar vein, Liu et al. (2014) created action sequences under music direction using Conditional GAN, therefore producing more harmony between the created dance and music. Concurrently, long variations of GAN, such CycleGAN and AttnGAN, have shown benefits in multimodal information fusion generation tasks and have been extensively used in the joint generating of music and dance (Ma et al., 2021).

Generating high-dimensional action data shows benefits from variational autoencoders (VAEs). Zheng et al. suggested a dance generation model based on VAE, for instance, which learns the possible structure of dance in hidden space so improving the variety of produced dances (Kritsis et al., 2022). To increase the coordination and temporal consistency of produced dance, VAE based generative models have also been applied for joint learning of music and dance (Chen et al., 2021). Based on VAE storing the temporal information of dance movements, Guo et al. (2021) study produces more natural and smoother created dance movements.

Temporal generating models are progressively being applied in dance generation. The production and modelling of dance motions extensively benefit from time series models including LSTM and gated recurrent unit (GRU). Using LSTM, Yuan and Pan (2022) generated smooth motion sequences and captured temporal correlations of dance motions. Furthermore preferred for effective processing of extended sequence creation is GRU because of its straightforward nature. Using GRU, Shailesh and Judy (2022) created dance sequences and confirmed the value of GRU in motion generating fluency.

To better reflect the diversity and probability of dance movements, probability distribution based generative models – such as mixture density networks (MDNs) – have lately been progressively included into the field of dance generation. For instance, Akber et al. (2023) suggested an MDN-based action generating technique that may create diverse dance movements. Furthermore included in the application of MDN in dance generation is the mix with other models such LSTM and Transformer, which helps the produced dance motions to have more delicacy and randomism (Myhrmann and Mabit, 2023).

Because the transformer model models extended sequences, it has also been used in dance generation. Yin et al. (2023) achieved effective matching of dance and music by using Transformer to create music and dance sequences and by means of a multi-level attention mechanism, therefore capturing complicated temporal information. Furthermore, in practice the dance generation model motivated by music has also produced decent outcomes. Based on audio input, Valle-Pérez et al. (2021) developed a Transformer dance generating technique that greatly increases the variety of produced dances.

Combining music and dance has drawn interest in the process of producing multimodal information. Combining dance action sequences with music aspects, Ferreira et al. (2021) attained multimodal generation of dance and music. Polignano et al. (2021)

investigated music driven action generation, in which the produced dance can show matching emotional styles by means of emotional data attached to the input music.

This work attempts to increase the diversity and authenticity of produced dances by combining mixed density networks with LSTM for dance creation, therefore leveraging multimodal inputs. These current investigations support this research both theoretically and technically and show that deep learning based digital dance creation has great possibilities in terms of model scalability and application potential.

This work presents a hybrid model based on MDN and LSTM, which captures the dynamic correlations between dance movements by building a multimodal temporal generating framework, in order to solve the aforesaid problems. MDN may provide varied alternative actions during the development of time series data, therefore addressing the issue that the conventional generating model does not adequately capture enough actions in a single mode and acting as a model for creating probability density distribution. Combining the temporal modelling capacity of LSTM, this framework is able to capture the continuity and correlation between dance motions during the generating process, therefore giving more naturalism and realism for dance creation. Furthermore, this approach permits dance generation depending on music or rhythm information, so improving the alignment of the produced action sequence with the musical style and hence improving the interaction between dance motions and background music.

The research objective of this article is to design a model that can automatically generate diverse, coordinated, and smooth dance action sequences based on input music or rhythm information. In the training phase, we used a publicly available dance dataset and captured the probability density distribution between different dance styles and movements through MDN, resulting in a dance action sequence with higher richness and coordination. Meanwhile, in order to enhance the effectiveness of generated dance, we conducted subjective and objective quantitative evaluations from the perspectives of motion fluency, naturalness, and diversity in the experiment. In terms of the precision and fluidity of produced dance movements, the experimental findings reveal that the suggested approach in this study is much better than conventional techniques, thereby improving the visual authenticity and emotional expression of the produced dance.

2 Relevant technologies

2.1 Musical features

Widely employed as characteristics for automatic speech and speaker recognition, MFCC are grounded in human auditory perception. Originally applied for several speech processing applications, MFCC has benefits in timbre representation according to research conducted in the field of MIR in machine learning.

The pre-emphasising processing equation looks like this:

$$y(n) = x(n) - \alpha x(n-1) \tag{1}$$

where n is the number of samples in each frame, $x(n)$ is the input signal, $y(n)$ is the output signal, and the filter coefficient α is usually 0.95, so that 95% of any sample is considered to come from previous samples.

The form of Hanming window $w(n)$ during the window addition process is as follows:

$$w(n) = (1-a) - a\cos\left(\frac{2\pi n}{N-1}\right) \tag{2}$$

where $0 \le n \le N-1$, a is proportional parameters.

The Mel frequency can be calculated using the audio frequency f using the following equation, in Hz:

$$F(Mel) = 2959 \times \log_{10}\left(1 + \frac{f}{700}\right) \tag{3}$$

Spectral Flux is the measurement of spectral changes between two consecutive frames and is calculated by the square difference of the normalised amplitudes of the spectra of two consecutive short-term windows. The equation is as follows:

$$SF_{(i,i-1)} = \sum_{k=1}^{N} \left(E_i(k) - E_{i-1}(k)\right)^2 \tag{4}$$

$$E_i(k) = \frac{X_i(k)}{\sum_{i=1}^{N} X_i(l)} \tag{5}$$

where $E_i(k)$ is the k-th normalised discrete Fourier transform (DFT) coefficient of the i-th frame. Spectral flux can be used to determine the timbre or initial detection of audio signals.

Compared to recorded audio, music has a unique rhythm in the temporal dimension. As a cognitive skill, beat induction (BI) allows us to hear the regular pulsations in music, perceive this regularity in music enables us to dance and create music together.

Ellis et al. used dynamic programming to achieve beat tracking, first setting the following equation as the objective function:

$$C\left(\{t_i\}\right) = \sum_{i=1}^{N} O\left(t_i\right) + \alpha \sum_{i=2}^{N} F\left(t_i - t_{i-1}, \tau_p\right) \tag{6}$$

$$F(\Delta t, \tau) = -\left(\log\frac{\Delta t}{\tau}\right) \tag{7}$$

where $\{t_i\}$ is the beat moment sequence found by the tracker, $O(t)$ is the 'fixed intensity envelope' derived from the audio, α is the weight that balances the two target terms, and $F(\Delta t, \tau)$ is a function that measures the difference between the beat interval Δt and the ideal beat interval τ_p defined by the target beat.

Based on the objective function, to recursively combine the best score time series to obtain the optimal score, the state transition function in dynamic programming is set as follows:

$$C^*(t) = O(t) + \max_{\tau=0\ldots t}\left\{\alpha F\left(t-\tau, \tau_p\right) + C^*(\tau)\right\} \tag{8}$$

2.2 Dance movement data acquisition

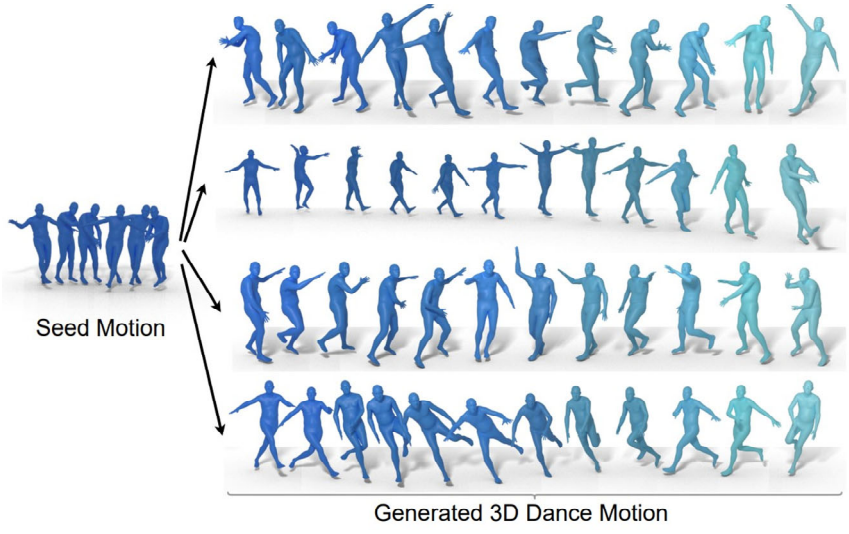
Motion capture (MoCap) is a technology that generates high-precision motion data by capturing the actual movements of actors. Its core lies in obtaining key point motion information of the human body through various sensors, converting it into formats such as three-dimensional coordinates or angles, and recording action details in a digital way. The common methods of motion capture are as follows:

1 Optical motion capture, which uses infrared cameras or high-speed cameras to capture the movement of markers on the actor's body.

2 Inertial motion capture captures the movement of body parts through inertial sensors (accelerometers, gyroscopes, etc.) without the need for optical equipment. The advantage of this method is convenience, but it may be slightly inferior to optical capture in terms of accuracy and stability.

3 Depth camera capture, using depth sensors such as Kinect to capture human body contours and postures, suitable for motion capture scenes that do not require marker points. These types of devices typically have lower accuracy, but lower cost and are easy to operate, making them suitable for rough capture of dance movements.

Figure 1 3D dance generation process (see online version for colours)



Motion capture technology may more easily recreate complicated motions and realistic physical interactions than conventional keyframe-based 3D model computer animation, therefore producing more realistic motion data and less effort in getting actions. For small-scale production, the cost of necessary software, tools, and staff may be too high; motion capture calls for certain hardware and software to record and process the generated data. Moreover, the gathered data is challenging to change once more. Should data error exist, the scene can only be re-shot.

More and more applications call for a lot of real-world human motion data as computer animation and robotics technologies improve; motion capture and manual production cannot satisfy this demand alone. Research into motion creation has so started

to take front stage. Two generalised action-generating algorithms can be distinguished: one is to learn the internal mapping and constraint relationships of action data by neural networks, so producing entirely new action sequences; another is to reuse and edit existing action fragments in the database, so combining them into new action sequences.

First category conventional dance synthesis algorithms based on music and action feature matching fit computer automatic choreography tasks. Action segments in the database provide the synthetic dance action sequences; hence, the variety of dance is constrained. Machine learning and deep learning techniques have started to be used in the field of action generating in order to create fresh action data. Though compared to the great learning capacity of neural networks, conventional machine learning methods have limited ability to capture data changes; HMM models, Gaussian processes, and dimensionality reduction techniques can all capture the intrinsic dependencies and potential correlations of action data. For action generation, this paper so decides to apply a sequence generating model grounded on deep learning.

The representation of dance movement data directly affects the input structure and feature extraction of the generative model. Common ways of representing actions include keypoint coordinates, joint angles, skeleton models, and motion feature vectors. The keypoint coordinate representation method describes human body posture through the positions of keypoints in three-dimensional space. Each keypoint corresponds to the three-dimensional coordinates of a body part (such as shoulders, elbows, knees, etc.). For dance generation, keypoint coordinates can fully record human posture, providing high detail resolution. Common representations include:

$$P = \left\{ \left(x_1, y_1, z_1 \right), \left(x_2, y_2, z_2 \right), ..., \left(x_n, y_n, z_n \right) \right\} \tag{9}$$

where n represents the number of keypoints, and (x_i, y_i, z_i) represents the three-dimensional coordinates of each keypoint.

The joint angle representation method represents human posture by describing the rotation angles of each joint. This method can directly reflect the skeletal structure of the human body, reduce data dimensions, and is suitable for action generation under physical constraints. Joint angles are usually represented by Euler angles or quaternions, and their calculation equation is:

$$\theta_{ij} = \arccos\left(\frac{v_i \cdot v_j}{|v_i||v_j|} \right) \tag{10}$$

where θ_{ij} is the angle between two adjacent bone vectors v_i and v_j.

The skeleton model representation considers the human body as a tree like structure composed of joints and bones, and describes the posture of the human body through the relative position of each joint. This model can reflect the hierarchical structure of human motion, making it easy to add physical and structural constraints during the generation process. Skeleton models are typically composed of nodes and edges, where nodes represent joint positions and edges represent bones, and are typically stored in a topological structure.

$$S = \left\{ \left(n_i, n_j \right) | \, i, \, j \in \, joints \right\} \tag{11}$$

This structure is suitable for multi joint constraints in generation to ensure the coordination of actions.

The motion feature vector method simplifies the representation of motion data by extracting feature values such as velocity, acceleration, etc. Feature vectors can reduce the dimensionality of input data and are suitable for feature driven dance generation.

Assuming a three-dimensional coordinate sequence $\{P_t\}$, its velocity and acceleration can be expressed as:

$$V_t = \frac{P_t - P_{t-1}}{\Delta t},\ A_t = \frac{V_t - V_{t-1}}{\Delta t} \tag{12}$$

This method provides important information about the motion characteristics for generating models, which helps to generate more natural actions.

2.3 Mixture density network

Fit for creating tasks with multimodal outputs, a mixed density network is a neural network able to forecast conditional probability distributions at the output layer. Learning these distribution parameters helps MDN to represent several ways for creating actions since the outcome is modelled as a weighted sum of several Gaussian distributions.

Minimising the sum of squares or cross error function on the input vector might help a network to produce approximative the conditional average of the goal data. Following the choice of a suitable encoding method, these average values reflect the posterior probability of every category in classification problems, hence guiding the network. However, the description of the target variable by the conditional mean is quite limited for the prediction (generation) problem of continuous variables, particularly with regard to multi value mapping, which frequently faces difficulties since the average of several correct target values may not always be the correct value. Separately modelling the conditional probabilities of the target data will help one to gain a full description of the data and solve the prediction problem of input vectors. Bishop put out a fresh network model combining mixed density models with conventional neural networks. The whole system is called a mixed density network; in theory, it can depict any conditional probability distribution.

MDN proposes to parameterise the distribution of numerous mixed components with neural network output. More especially, the whole network generates the probability density function of every dimension in the tensor rather than a single position tensor. The output data is used to define the weights α of each mixed component, as well as to parameterise the mean and variance σ of each mixed component. The weight α is normalised using the softmax function to ensure that they form an effective discrete distribution, while other outputs are processed using appropriate functions (such as exponential functions) to keep their values within a meaningful range. In RNN/LSTM networks, the mixed density model influences not only the current input but also limits the output distribution by the past input history.

A linear combination of several mixed components reflects the probability density of the target data; the probability of the target vector t given the output x is:

$$p(t \mid x) = \sum_{i=1}^{m} \left(\alpha_i(x) \varphi_i(t \mid x) \right) \tag{13}$$

where m is the total number of the mixed components and α_i is the mixing coefficient for every mixed component of the input x. Usually utilised as a Gaussian kernel function, function φ_i is the conditional density of the i-th kernel of the target vector t.

Therefore, written as a tensor, MDN has m (c + 2) output variables:

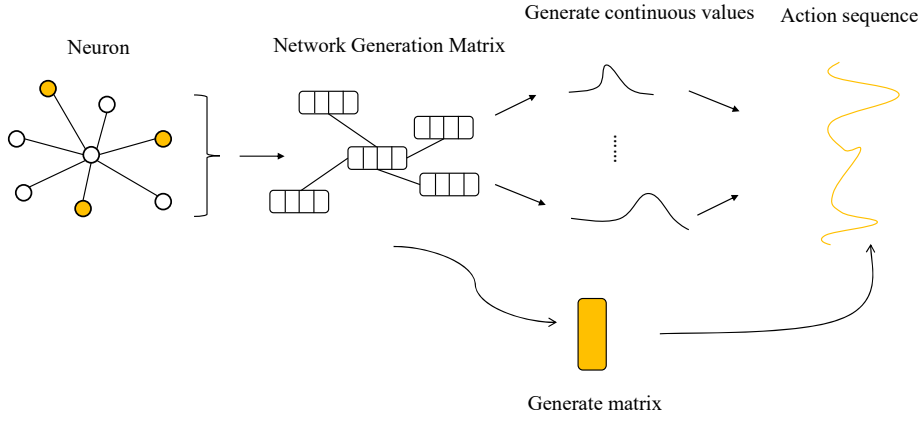
$$z = \left[z_1^{\alpha}, ..., z_m^{\alpha}, z_{m+1}^{\mu}, ..., z_{mc+m+1}^{\mu}, z_{mc+m+2}^{\sigma}, ..., z_{m(c+2)}^{\sigma} \right] \tag{14}$$

This includes all the parameters required to construct a hybrid model. The number of mixed components m is arbitrary and can be understood as the number of different choices that the network can make at each time point.

3 Method

The MDN-L algorithm proposed in this article is a digital dance generation method based on a hybrid density network and a LSTM network, aiming to generate dance action sequences that conform to the rhythm of music. This method introduces a multimodal temporal generation framework of MDN and LSTM, which can generate complex and natural dance movements driven by music. The algorithm design, model construction, and overall framework will be described in detail below.

Figure 2 MDN-L algorithm diagram (see online version for colours)



3.1 Algorithm design

The core of digital dance generation lies in capturing the temporal dependencies and diverse features of dance movements. In temporal modelling, dance movements have significant temporal characteristics, meaning that the previous movement state greatly influences the present condition. Therefore, the LSTM model is chosen to capture temporal dependencies in action sequences. LSTM can effectively solve the problem of gradient vanishing during the generation of long sequences, ensuring that the generated action sequences are smooth and natural. Dance movements have diversity and randomness, and a single generated result cannot meet the diverse needs of dance. To this

end, this article introduces a mixed density network (MDN) as the generation layer, which captures multimodal characteristics by outputting multiple Gaussian distributions through MDN. The output of MDN is a weighted Gaussian mixture distribution, it offers several action sample options during the production process therefore producing a variety of actions. Music features are employed as conditional inputs to drive the created dance motions to alter with music characteristics, therefore matching dance movements with rhythm information. This article selects features such as rhythm, beat, and musical emotion to ensure that the generated action sequence conforms to the music style.

3.2 Model building

3.2.1 Music feature extraction

Extensive feature extraction reflecting the commonalities of music and motions is essential while doing computer music choreography to choose dance routines that complement the provided target music. To first match action segments and guarantee that the style, speed, and other features of each action segment in the final overall choreography are unified and more aesthetically pleasing, this article analyses the general characteristics of music and investigates the impact and control of its overall features on dance movements.

One can classify music features into low-level and high-level ones somewhat broadly. Among the low-level characteristics are amplitude envelope, spectral characteristics, short-term energy, short-term power spectral density, etc. The high-level elements of the music consist in its emotions, styles, etc. The present widely used music emotion style classification algorithms use machine learning algorithms to acquire the mapping link between music low-level features and emotional styles since the difficulties in quantifying and evaluating the high-level aspects of music. Stated differently, low-level qualities allow one to characterise the high-level aspects of music. Consequently, this part mostly examines the general low-level aspects of music.

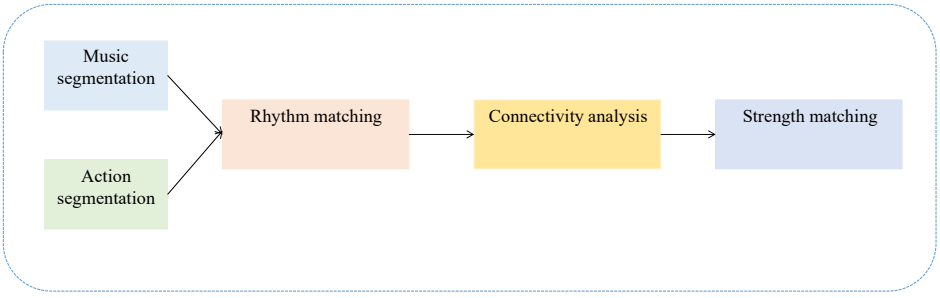
We extracted the waveform and Mel spectrum of music as basic music features, and also extracted beat, note chromaticity, and note onset intensity as advanced music features. Among them, waveform and Mel spectrum are basic acoustic features, beat refers to the total number of notes in each bar of music, reflecting the temporal information of the music, and note chromaticity projects the spectrum into 12 different intervals, representing 12 different semitones within a time interval. The initial intensity of a note is used to define the rhythm of music.

3.2.2 Action feature extraction and matching

By extracting action features, the model can understand and reproduce the dancer's posture, action sequence, and temporal relationships, ensuring that the generated dance movements are natural, smooth, and artistic.

This paper mostly depends on the matching degree of rhythm and intensity elements when matching music and action sections. Four steps define the feature matching algorithm: rhythm feature matching, connectivity analysis, constructing action graphs to extract connected action sequences, and intensity matching. The algorithm flow is shown in Figure 3.

Figure 3 Calculation process of music and action matching degree (see online version for colours)



The rhythm matching between music and action mainly tests the synchronisation degree of rhythm points between each action segment in the action database and the given music segment. The degree of rhythm synchronisation is measured by the number of rhythm points that match the music and action segments. The higher the degree of matching, the stronger the temporal correspondence between the rhythm points of the two. Furthermore allowed during matching is a limited proportion of scaling to fully use action data and get better matching results. This article uses a scaling ratio applicable for any values with a step size of 0.05 within the range of 0.9 to 1.1. The equation to find rhythm matching degree is:

$$s = \max_{s,f_0} \sum_{f=1}^{L} \frac{F_R^{Mc}(f) \cdot F_R^{Mo}(s \cdot f + f_0)}{F_R^{Mc}(f) + F_R^{Mo}(s \cdot f + f_0)} \tag{15}$$

where L is the length of M_i and N_i, s is the scaling factor, and f_0 is the translation amount.

In order to ensure that the synthesised actions look more natural and harmonious, it is necessary to perform connectivity analysis on the adjacent actions obtained from matching. When calculating the distance between two frames of actions, it is not only necessary to ensure that the individual two frames are complete, but also to take a window of length k frames and calculate the sum of the distances between k pairs of frames within the window:

$$D(f_i, f_j) = \sum_{t=0}^{k-1} diff(f_{i-k+1+l}, f_{j+l}) \tag{16}$$

where $diff(f_i, f_j)$ is the sum of the distances between the corresponding joint positions of two frames of actions.

Based on the results of connectivity evaluation, it can be determined whether adjacent candidate action segments that match the music segment can be connected. In order to obtain a complete and connectable action sequence, this paper uses a graph based depth first algorithm to traverse the rhythm matched action segments. Consider action fragments as nodes in the action graph, and if two action fragments are connectable, establish a directed edge between these two nodes.

Perform strength matching on the target music sequence and each candidate connected action sequence. The action sequence with the highest strength similarity is the

one with the highest matching degree, and it is used as the synthesised dance action sequence of the target music.

When analysing intensity similarity, the first step is to obtain intensity histograms of music and action segments, and measure the degree of intensity matching based on the similarity of the histograms. This article uses the Bhattacharyya coefficient to measure strength similarity, and the strength matching equation for music and action segments is:

$$D = \sum_{f=1}^{L_{MC}} \sqrt{\frac{F_1^{Mc}(f)}{\sum_{k=1}^{L_{MC}} F_1^{Mc}(k)} \cdot \frac{F_1^{Mo}(f)}{\sum_{k=1}^{L_{MO}} F_1^{Mo}(k)}} \tag{17}$$

where L_{MC} is the length of M_i, and L_{MO} is the length of N_i.

4 Experiment

4.1 Data set

The AIST++dataset provides significant advantages in the richness and accuracy of dance movement data and music features. It not only includes multiple dance styles, but also encompasses various music genres, and all motion data is captured with high precision in 3D using professional equipment. In addition, the AIST++dataset also provides synchronised music data and corresponding feature annotations, making the dataset widely applicable to tasks such as generative AI, action generation, and dance style transfer. AIST++contains 3D dance data for 10 music genres, each containing multiple dance segments. It contains multiple types of music, ensuring the harmony between dance movements and music. Music data is stored in MP3 format, accompanied by detailed music beats and rhythm annotations. The AIST++ dance video scene is shown in Figure 4.

Figure 4 AIST++ dance video scene (see online version for colours)

4.2 Evaluation

Using structural similarity index (SSIM) and Fréchet inception distance (FID) as evaluation metrics can help assess the quality and diversity of generated 3D dance movements. SSIM is a metric used to measure the similarity between images, initially used for image quality assessment, but can also be used for similarity evaluation in sequences generated from 3D actions through feature images of skeleton keyframes. We project the skeleton actions in the generated dance sequence onto a two-dimensional plane (such as front view, side view), and calculate the SSIM score between each frame's generated action and the real action to measure the structural similarity between the generated action and the real data. Convert the generated skeleton action sequence into a 2D skeleton image sequence and perform SSIM calculation with the real skeleton image sequence. The SSIM calculation equation is as follows:

$$SSIM(x, y) = \frac{\left(2\mu_x\mu_y + C_1\right)\left(2\sigma_{xy} + C_2\right)}{\left(\mu_x^2 + \mu_y^2 + C_1\right)\left(\sigma_x^2 + \sigma_y^2 + C_2\right)} \tag{18}$$

where μ_x and μ_y are the average values of two images, σ_x^2 and σ_y^2 are the variances, σ_{xy} is the covariance, and C_1 and C_2 are constants used for stable calculations.

Often used for evaluating generative models, FID is a metric used to compare the distribution of produced data with that of real data. FID is used in dance generation tasks to measure the diversity and quality of generated action sequences, and is particularly suitable for establishing a global quality evaluation between the generated skeleton sequence and the real skeleton sequence. Firstly, the 3D skeleton data is used to extract high-dimensional feature vectors through a pre trained feature extraction network (such as Inception network), and the feature distributions (mean and covariance) of the generated skeleton and the real skeleton are calculated. Then calculate the Fréchet distance between the generated data and the real data, using the following equation:

$$FID = \left\|\mu_g - \mu_r\right\|^2 + Tr(\left(\Sigma_g + \Sigma_r - 2\left(\Sigma_g\Sigma_r\right)^{\frac{1}{2}}\right) \tag{19}$$

where μ_g and μ_r generate the mean vectors of the feature distributions of the data and the real data, respectively, while Σ_g and Σ_r are the covariance matrices of the generated data and the real data, respectively.

4.3 Experimental results and analysis

There is no standard answer for the implementation of dance, so subjective evaluation is an important reference for its quality assessment. This section first compares the visualisation results of dance sequences predicted using the method in this chapter with the baseline method, and then presents the visualisation results of the approach in this chapter.

To validate the effectiveness of the proposed generative model, we examined it using three state-of- the-art techniques and compared Zhuang et al.'s (2022) model, DanceNet (Li et al., 2020), and Deep Dance (Wang and Ton, 2022), as shown in Table 1.

Table 1 Comparison and analysis results of models

Model	*FID value(average)*	*SSIM value (average)*
Zhuang et al. (2022)	69.82	0.89
DeepNet	68.43	0.90
DeepDance	67.56	0.92
Ours	46.28	0.94

Figure 5 Generated 3D dance movements (see online version for colours)

Figure 5 shows the experimental findings of this technique, which show that the MDN-L model generates high-quality and varied dancing photos and videos outperforming other approaches. The FID of the proposed model is significantly lower, about 30% lower than other models, indicating an improvement in the quality and variability of the generated images. A lower FID score indicates higher authenticity and diversity since the created dance motions more precisely depict the distribution of real dance data.

Furthermore, the suggested model's greatest SSIM rating emphasises its benefit in maintaining the intricate form of the original dance moves. The produced film highlights the model's capacity to generate motions that closely reflect natural dance flows since the fundamental elements of dance – such as posture accuracy, movement fluidity, and overall spatial configuration – are better retained in this created video.

Furthermore, the suggested model generates dancing motions that are tightly linked with the beat and emotional tone of the music, therefore capturing the dynamic variations in rhythm and intensity. This synchronising guarantees that the 3D dance video created by the model not only looks like actual dancing motions but also coordinates with the music accompaniment, therefore improving the immersive quality of the produced performance.

5 Conclusions

Aiming to produce high-quality dance action sequences that fit music rhythm and emotion, in this work we offer a digital dance generating approach based on MDN and long short term memory (LSTM). Particularly obtaining better performance in FID and SSIM, the experimental results show that the proposed method greatly exceeds conventional generative models in terms of fluency, naturalness, and diversity in generating dance movements. This suggests that the multimodal temporal generating framework combining MDN and LSTM can efficiently capture temporal dependencies and pose changes between dance movements, and produce dynamic dance videos that fit the music style, so improving the structural consistency and diversity of the produced results.

Although this method has achieved certain results in generating dance action sequences, there are still some shortcomings and directions for future improvement. The dance dataset used in this study has certain limitations in terms of style and variety. In the future, richer dance styles and diverse dance action datasets can be introduced to further enhance the model's generalisation ability, making it applicable to multiple dance types and action styles. The current model has a large computational load and is difficult to achieve real-time dance generation. Through model structure optimisation and lightweight design, future exploration can improve generation efficiency and achieve real-time dance action generation, which can be applied to interactive scenarios such as virtual reality and real-time dance accompaniment.

Acknowledgements

This work is supported by the Huangmei Opera Art Research Center, Hubei Key Research Base of Humanities and Social Sciences (No. 201807703).

References

Akber, S.M.A., Kazmi, S.N., Mohsin, S.M. et al. (2023) 'Deep learning-based motion style transfer tools, techniques and future challenges', *Sensors*, Vol. 23, No. 5, p.2597.

Baltrušaitis, T., Ahuja, C. and Morency, L-P. (2018) 'Multimodal machine learning: a survey and taxonomy', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 41, No. 2, pp.423–443.

Chen, K., Tan, Z., Lei, J. et al. (2021) 'Choreomaster: choreography-oriented music-driven dance synthesis', *ACM Transactions on Graphics (TOG)*, Vol. 40, No. 4, pp.1–13.

Creswell, A., White, T., Dumoulin, V. et al. (2018) 'Generative adversarial networks: an overview', *IEEE Signal Processing Magazine*, Vol. 35, No. 1, pp.53–65.

Ferreira, J.P., Coutinho, T.M., Gomes, T.L. et al. (2021) 'Learning to dance: a graph convolutional adversarial network to generate realistic dance motions from audio', *Computers & Graphics*, Vol. 94, No. 1, pp.11–21.

Guo, X., Zhao, Y. and Li, J. (2021) 'DanceIt: music-inspired dancing video synthesis', *IEEE Transactions on Image Processing*, Vol. 30, No. 1, pp.5559–5572.

Henrickson, L. (2020) 'The Artist in the Machine: The World of AI-Powered Creativity by Arthur I. Miller', *Configurations*, Vol. 28, No. 3, pp.398–400.

Kritsis, K., Gkiokas, A., Pikrakis, A. et al. (2022) 'Danceconv: dance motion generation with convolutional networks', *IEEE Access*, Vol. 10, No. 3, pp.44982–45000.

Li, X., Wang, L., Wang, M. et al. (2020) 'DANCE-NET: density-aware convolution networks with context encoding for airborne LiDAR point cloud classification', *ISPRS Journal of Photogrammetry and Remote Sensing*, Vol. 166, No.1, pp.128–139.

Liu, P., Deng, W., Li, H. et al. (2024) 'MusicFace: music-driven expressive singing face synthesis', *Computational Visual Media*, Vol. 10, No. 1, pp.119–136.

Ma, L., Huang, K., Wei, D. et al. (2021) 'FDA-GAN: flow-based dual attention GAN for human pose transfer', *IEEE Transactions on Multimedia*, Vol. 25, No. 9, pp.930–941.

Myhrmann, M.S. and Mabit, S.E. (2023) 'Estimating city-wide hourly bicycle flow using a hybrid LSTM MDN', *Transportation Research Part A: Policy and Practice*, Vol. 176, No.9, p.103783.

Polignano, M., Narducci, F., de Gemmis, M. et al. (2021) 'Towards emotion-aware recommender systems: an affective coherence model based on emotion-driven behaviors', *Expert Systems with Applications*, Vol. 170, No. 4, p.114382.

Shailesh, S. and Judy, M. (2022) 'Understanding dance semantics using spatio-temporal features coupled GRU networks', *Entertainment Computing*, Vol. 42, No. 7, p.100484.

Vahdat, A. and Kautz, J. (2020) 'NVAE: a deep hierarchical variational autoencoder', *Advances in Neural Information Processing Systems*, Vol. 33, No. 1, pp.19667–19679.

Valle-Pérez, G., Henter, G.E., Beskow, J. et al. (2021) 'Transflower: probabilistic autoregressive dance generation with multimodal attention', *ACM Transactions on Graphics (TOG)*, Vol. 40, No. 6, pp.1–14.

Wang, S. and Tong, S. (2022) 'Analysis of high-level dance movements under deep learning and internet of things', *The Journal of Supercomputing*, Vol. 78, No. 12, pp.14294–14316.

Wei, R. and Mahmood, A. (2020) 'Recent advances in variational autoencoders with representation learning for biomedical informatics: a survey', *IEEE Access*, Vol. 9, No. 1, pp.4939–4956.

Yin, W., Yin, H., Baraka, K. et al. (2023) 'Multimodal dance style transfer', *Machine Vision and Applications*, Vol. 34, No. 4, p.48.

Yuan, X. and Pan, P. (2022) 'Research on the evaluation model of dance movement recognition and automatic generation based on long short-term memory', *Mathematical Problems in Engineering*, Vol. 2022, No. 1, p.6405903.

Zhuang, W., Wang, C., Chai, J. et al. (2022) 'Music2dance: Dancenet for music-driven dance generation', *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, Vol. 18, No. 2, pp.1–21.