Research on Automatic Dance Generation System Based on Deep Learning

Jia Lan

Zhengzhou Preschool Education College, Zhengzhou 450000, China

Correspondence should be addressed to Jia Lan; 2004005@zzpec.edu.cn

Received 19 April 2022; Revised 26 July 2022; Accepted 3 August 2022; Published 24 August 2022

Copyright © 2022 Jia Lan. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to model the mapping relationship between music and dance so that the model can automatically generate the corresponding dance based on the rhythm and style of music is the most important issue to be resolved in the automatic arrangement of music and dance. Due to the flexible movement of limbs and the change in camera angle, the visual information of various parts of the human body changes drastically in practical applications, posing significant difficulties in the generation of high-resolution target dance images. Based on deep learning, this paper proposes a two-way voluntary neural network (TWCNN) dance automatic generation system (DL). The long short-term memory (LSTM) network model is an action generation sub-network, which generates each frame of action based on the extracted high-level music features and the bottom-level action features from the previous moment. The recognition accuracy of the proposed algorithm is 12.4 percent and 6.8 percent better than that of the conventional depth CNN, according to the test results.

1. Introduction

With the advancement of computer animation, there is a growing demand for automatic choreography of dance animation. When traditional film and television animation technology generates a dance action sequence, a professional animator with extensive experience is required to adjust the action of the character skeleton frame by frame using 3D software. Due to the variety and complexity of dance movements, this process frequently requires a great deal of time and labor [1]. The next generation of motion capture technology can more effectively address the labor shortage. The animator only needs to bind the motion capture files to the character model, after which he can generate natural and realistic dance motion animation [2]. If we can learn the relationship between the motion capture file and the corresponding background music, and then automatically generate the corresponding action sequence based on the input music, it will be more efficient and less expensive to generate dance animation.

With the advancement of deep learning (DL) technology, an increasing number of research projects have been put into production, and remarkable progress has been made in every area. For example, literature [3] synthesizes the corresponding video by searching for the frame in which the mouth position matches the desired speech, allowing the speaker to say the phrases they did not initially say. Literature [4] employs DL technology to summarize unsupervised relocation motion data. Literature [5] presented a well-designed multi-view system for calibrating the personalized motion skeleton model and obtaining 3D joint estimation via training, rendering the image of the character subject performing new motion. The variational automatic encoder [6] proposed in reference [7] uses reparameterization to maximize the lower limit of data generation possibilities. This method’s other branch is the self-advancing model. The main idea is to calculate the product of the conditional distribution of pixels pixel by pixel as the image’s joint distribution of pixels [8]. By modifying the fine-grained portion of the image, people with clothing covering their entire bodies are generated. Literature [9] generates images with the condition of text description by adding text information to generators and discriminators, investigates the optimal position for using conditions, and generates images...
using key points or segmentation information. Nonetheless, diverse postures in other methods are consistent with any given perspective, making the adjustable expressive force inferior to our work. In the work of this system, posture skeleton information is utilized in a more transparent and adaptable manner; i.e., postures connected by key points are used to simulate various human shapes.

In this paper, the automatic dance generation system based on deep learning is described in detail, and the relationship between music and action, as well as the synthesis of music and dance, is the primary topic of investigation. This paper divides the automatic dance generation system based on deep learning into two modules: training and generation. In the training stage, for a particular type of dance, the corresponding relationship between music and action is trained using a rotation vector, and a corresponding relationship between music and action that meets the matching accuracy and operation speed is obtained. In the generation stage, the system searches action segments by traversing the action graph and synthesizes music segments and dance action segments.

This study proposes a system for the automatic generation of dances based on deep learning. It is possible to generate high-resolution and realistic dance videos using a well-designed network of generators and discriminators and a variety of loss functions as objective functions. By traversing the action graph, we can search through the action segments and choose the ones that correspond to the music segments. Following synthesis, the system simultaneously displays the input music segment sequence and the dance segment sequence that corresponds to it. The method for recognizing dance video motion is investigated in this paper. Based on an analysis of the traditional Inception V3 and 3D-CNN networks, it has been determined that the method proposed in this paper outperforms them in recognizing dance video motion.

The work in this paper is structured as follows. Section 1 is the introduction section, which describes the motivation, significance, and contribution of the work in this paper. Section 2 is the related work, which introduces and summarizes the previous work related to this paper. Section 3 is the methodology, which describes the details of the methodology. Section 4 presents the experimental results and analyzes the superiority of the method in this paper. Section 5 is the conclusion, which summarizes the work of this paper and discusses the shortcomings and future work of this paper.

2. Analysis of Related Technologies

DL theory is partially the result of the evolution of neural network theory. It could be considered a continuation of neural network. The benefit of DL is that its model is hierarchical, with a large number of parameters and sufficient capacity, so it can more accurately express the data’s characteristics [10, 11]. Aiming at the challenging problem of image and speech recognition, although manual design is required, a large number of training data can yield good results. Moreover, DL combines classifier and feature in a framework and uses training samples to learn and obtain features, which has the benefit of reducing manual design intervention, drastically reducing workload, and enhancing recognition efficiency.

Convolutional neural network (CNN) is the most widely used deep learning (DL) network in image processing. Figure 1 depicts the network’s structure, which consists primarily of convolution, pooling, and full joint operations [12].

2.1. Convolutional Layer. Convolution is a common mathematical operation in the field of information processing. The convolution operation method in discrete domain is shown in the following formula:

\[
(f * g)[n] = \sum_{m=-\infty}^{\infty} f(m)g(n-m).
\]

Convolution operation in discrete domain requires a reasonable choice of convolution kernel, which is usually a \(n^2\) matrix.

2.2. Pool Layer. Pooling layer refers to selecting a local area to replace the complete area in the features obtained by convolution, and pooling realizes the filtering and selection of features. Commonly used pooling operations include maximum pooling and mean pooling. Among them, the operation method of mean pooling is shown in Figure 2 [13].

For CNN, the introduction of pooling layer realizes the down-sampling of image information, which can effectively simplify the network structure and prevent over-fitting.

2.3. Full Connection Layer. At the end of CNN is the full connection layer, which synthesizes the features of the previous layer to get the classifier of the network. The connection diagram is shown in Figure 3.

The mode of operation of the fully connected layer is comparable to that of the conventional neural network with a single hidden layer, which connects the input layer to the hidden layer and the hidden layer to the output layer via weights and offsets. The calculation method is shown in the following formula:

\[
\begin{bmatrix}
A_1 \\
A_2 \\
A_3 \\
A_4
\end{bmatrix} =
\begin{bmatrix}
W_{11} & W_{12} & W_{13} \\
W_{21} & W_{22} & W_{23} \\
W_{31} & W_{32} & W_{33} \\
W_{41} & W_{42} & W_{43}
\end{bmatrix}
\begin{bmatrix}
X_1 \\
X_2 \\
X_3
\end{bmatrix} +
\begin{bmatrix}
b_1 \\
b_2 \\
b_3
\end{bmatrix}.
\]

2.4. Output Layer. The output layer transforms the output of the fully connected layer into the final output of the deep network with the help of nonlinear function. For the binary classification problem, the logistic function is usually selected [14]. In this paper, softmax cross entropy function [15] is selected, and its form is shown in the following formula:
where \( c \) stands for the category of classification, and when the output result is consistent with the actual category, \( y_c = 1 \).

3. Research Method

3.1. System Structure Design. The key to realize automatic dance generation system based on deep learning lies in generating pixels based on patch in the generation network, discriminating multi-scale images in the discrimination network, and designing the final loss function. The overall structure of this application is shown in Figure 4.

3.1.1. Interface Layer. The interface layer is primarily used for user interaction with the system, and all functions must be as straightforward as possible. The application interface consists primarily of various control and view modules, such as video button selection, video view display, dance button generation, and text prompt views. The application interface will present the user with the original dance video and the generated dance video, compare the synthetic effect, and generate the corresponding indicator description. A simple and generous interface design will result in a positive user experience [16].

3.1.2. Logic Layer. Logic layer is often the core of the whole application system, including the realization of core functions and the connection between upper and lower layers. The main functions include reading local video or collecting video data from the current camera, and collecting data is an essential function for generating models.

3.1.3. Data Layer. The data layer is primarily used to store and retrieve video data, and also includes storage and retrieval of model and generation video data. Due to the fact that this system is an offline system, the model is too large and the training time is too long, making online training and generation impossible at this time [15]. The system must therefore store the pretrained character action model and generate the dance video of the corresponding character.
based on the skeleton model’s input. The generated video of dance will be saved in the specified directory.

As depicted in Figure 5, the system can be roughly divided into four parts based on the implementation flow of dance generation: input data, data processing, dance generation, and video output.

Input data refer primarily to local video reading or system camera calls to collect character action video. In the data processing section, the input video is divided into images and then fed into the openpose program to obtain the body and hand key points; to generate the dance module, it is necessary to input the detected key point data into the model in order to obtain the generated figure dance images. In the final output phase, the generated dance images must be converted into user-viewable videos and displayed on the system’s user interface.

3.2. Feature Analysis and Extraction. The dynamic characteristics of dance movement refer to the movement speed, acceleration, and movement direction of each joint point. Static features are used to describe the posture features of human characters. The static features extracted in this paper include movement spacing, movement intensity, arm shape, footstep footprint, and balance [17, 18]. The calculation method of each feature is described in detail below.

The movement speed \( m_{\text{velocity}} \) of joint points refers to the displacement of human joint points in unit time, and its calculation method is as follows:

\[
m_{\text{velocity}} = \frac{f_i(j + 1) - f_i(j)}{\Delta t}.
\]  

\( f_i(j) \) represents the position of joint point \( i \) at frame \( j \), \( f_i(j + 1) - f_i(j) \) represents the position variation of joint point \( i \) at frame \( j \), and \( \Delta t \) is the time length of frame \( j \).

Action acceleration, similar to the definition of action speed, is the change in velocity of each joint point of the human body over a specified period of time, and acceleration can represent the acting force on the joint. This paper contends that there is a correlation between the rate of action and the intensity of music. Accelerating action is accompanied by fast-paced music and vice versa. The calculation formula for acceleration is as follows:

\[
m_{\text{acceleration}} = \frac{v_i(j + 1) - v_i(j)}{\Delta t}.
\]  

\( v_i(j) \) represents the speed of the joint point \( i \) at the \( j \) th frame.

Action spacing is a comprehensive description of human motion amplitude, with motion amplitude referring to the amount of space spanned by human joint movements [19]. In this paper, the action distance is computed based on the moving distance of each node position in the action segment, whereas the moving distance of joint points is determined by calculating the distance between each joint point and its central position in the action segment. The centroid of joint points represents the average value of joint points within the action segment. Action spacing can be determined using the following formula:

\[
C_i = \frac{1}{N} \left( \sum_{j=1}^{N} f_i(j) \right),
\]

\[
m_{\text{span}} = \sum_{j=1}^{n} \left| f_i(j) - C_i \right|,
\]  

\( C_i \) is the center position of all positions of the \( i \)-th joint point in the action segment, \( N \) represents the number of frames in the action segment, and \( n \) represents the number of joint points used in the human model.

Dense action is when a joint moves a great deal in a relatively small area over a period of time; sparse action is when the node moves a small amount over a relatively large area [20]. Therefore, the ratio of action speed to action spacing can be used to express action density.

\[
m_{\text{density}} = \frac{m_{\text{velocity}}}{m_{\text{span}}} = \frac{\sum_{j=1}^{N} f_i(j + 1) - f_i(j)}{\sum_{j=1}^{n} \left| f_i(j) - C_i \right|}.
\]  

According to the human body model’s representation method, there are three components of arm posture characteristics in the local rectangular coordinate system: horizontal, vertical, and vector direction. In the vertical direction, it describes the degree of upward or downward movement of the arm. In the horizontal direction, it describes the degree of outward expansion and inward contraction of the arm.

In this paper, the number of steps in an action segment is defined as the footstep track characteristic. Stunning dance movements are accompanied by high-frequency footstep movement, whereas simple and rigorous dance is accompanied by low-frequency footstep movement.

In the vertical direction, it can be assumed that the foot does not move on the ground if the foot’s position does not exceed a certain limit range and the foot’s position does not change within a certain time period [21]. Therefore, the characteristics of the footstep track can be determined by calculating the amount of time that the footstep is off the ground throughout the entire action segment. The method of calculation is as follows:

\[
m_{\text{footsteps}} = \sum_{j=1}^{N} \left( S_n(f_i(j + 1) - S_n(j)) \right),
\]

\[
S_n(j) = \begin{cases} 
1, & \text{if } f_i(j) > \text{thread}, \\
0, & \text{other.}
\end{cases}
\]

\( S_i(j) \) calculates whether the footstep moves, 1 indicates movement, and 0 indicates movement. \( \text{thread} \) is the threshold set by the system for judging whether the footsteps move or not.

The posture of human body can be well balanced and stable, or unbalanced and unstable. The balance characteristic describes the stability of human posture, through which we can describe the movement characteristics of dance conveniently. The mapping position of human body center on the ground is represented by \( C \), a rectangle is formed with left and right heels as the measurement
standard, and the center position of the rectangle is marked as \( R \). This paper measures the balance of human body by calculating the distance between the mapping position and the center position. The calculation method of human body balance is as follows:

\[
\text{balance} = |C - R|.
\]  

(9)

3.3. Automatic Generation Method of Dance Action

3.3.1. Dance Action Recognition Based on DL Network. The method of feature extraction in spatial domain is consistent with the method of image information extraction, and the features in temporal domain are identified by optical flow (OF). OF reflects the change track of pixels after the object’s motion state changes in space, which is widely used in motion detection. The acquisition method is as follows:

At time \( t \), for the spatial coordinate position point \( O(x, y) \), the pixel brightness of this point is \( I(x, y, t) \). In \( dt \) time, the point moves to a new position \( (x + dx, y + dy) \) in the next frame. At this time, because the time is very short, the luminance of this point has the relationship [22, 23] as shown in the following formula:

\[
I(x, y, t + dt) = I(x + dx, y + dy, t + dt).
\]  

(10)

The Taylor expansion formula is shown in the following formula:

\[
\frac{\partial I}{\partial x} dx + \frac{\partial I}{\partial y} dy + \frac{\partial I}{\partial t} dt = 0.
\]  

(11)

At this time, the OF equation of this point can be obtained as shown in the following formula:

\[
\frac{\partial I}{\partial x} V_x + \frac{\partial I}{\partial y} V_y + \frac{\partial I}{\partial t} = 0.
\]  

(12)

In the above formula, \( V_x \) and \( V_y \) are OF vectors. From this differential equation, pyramid algorithm is needed to solve the OF vector.

For a pixel area OF \( 3 \times 3 \) size, there are 9 of tracks, which can be expressed as the following formula in the form of matrix:

\[
Av = b.
\]  

(13)

The variables are shown in the following formula:

\[
A = \begin{bmatrix}
I_x(q_1) & I_y(q_1) \\
I_x(q_2) & I_y(q_2) \\
\vdots & \vdots \\
I_x(q_9) & I_y(q_9)
\end{bmatrix},
\]

\[
v = \begin{bmatrix}
V_x \\
V_y
\end{bmatrix},
\]

\[
b = \begin{bmatrix}
-I_x(q_1) \\
-I_x(q_2) \\
\vdots \\
-I_x(q_9)
\end{bmatrix}.
\]  

(14)

(16) can be solved by using the following formula:

\[
v = (A^T A)^{-1} A^T b,
\]

\[
\begin{bmatrix}
V_x \\
V_y
\end{bmatrix} = \left[\sum_t I_x(q_t)^3 \sum_t I_x(q_t) I_y(q_t) \sum_t I_x(q_t) I_y(q_t) \sum_t I_y(q_t) \right]^{-1} \left[\sum_t I_x(q_t) I_x(q_t) - \sum_t I_y(q_t) I_y(q_t) \right].
\]  

(16)

3.3.2. Action Generation Sub-Network. The action generation sub-network mainly refers to the ERD model proposed in article [24], which is composed of a nonlinear encoder, a two-layer long short-term memory (LSTM) network, and a nonlinear decoder, and can generate the next action according to the input action prediction.

The input of the action generation sub-network is a 120-dimensional feature vector formed by connecting the action
feature $M_{t}^{G,1}$ of the previous moment and the coded music feature $A_{t}^{d,1}$ of the current moment.

Firstly, the feature vector is input to an encoder composed of two fully connected layers, in which the first fully connected layer is followed by a ReLU activation layer. Then, the coded music action features are input into the LSTM with a two-layer hidden unit size of 1024. Finally, the output features of LSTM are transmitted to the decoder, which is composed of three fully connected layers and outputs 60-dimensional motion feature vector $M_{t}^{G}$ at the current time.

3.3.3. Network Training. Most motion generation models choose the rotation vector as the motion feature [6, 25], because compared with Euler angle, the rotation vector has no problems of rotation axis sequence and universal lock, and is more concise than quaternion form, and there is no restriction that the output vector needs to be a unit vector.

For music and dance data, we can use rotation vector to express its motion characteristics, and the dimension of each frame is $60 \times l$. In order to make the input action data conform to a uniform distribution, this paper uses z score standardization method to normalize each dimension of the input data to an average value of 0 and a variance of $l$ as shown in the following formula:

$$\hat{M}_{i}^{F} = \frac{(M_{i}^{F} - \mu_{M})}{\sigma_{M}}.$$  \hspace{1cm} (17)

$M_{i}^{F}$ is the motion feature vector of the $i$th frame, $\mu_{M}$ is the vector composed of the average value of each dimension motion feature, $\sigma_{M}$ is the vector composed of the variance of each dimension motion feature, and $\hat{M}_{i}$ is the normalized motion feature vector.

During training, action samples and music samples must be divided into windows based on the duration of the training and window data must be randomly selected to form batches. Attention must be paid to the one-to-one correspondence between action samples and music samples, so that when window data are randomly selected, they retain their original correspondence.

4. Experimental Analysis

4.1. Experimental Design. In order to assess the validity of the model, dance video action data sets are utilized in experiments. In this data set, there are 101 categories of dance movements, the video frames are all 25 fps, the resolutions are all 320*240, and the video duration ranges from 2.31 to 67.24 seconds.

In order to measure the recognition accuracy of dance movements, this paper uses the commonly used evaluation indexes $F_1$ and MSE in DL. These two indexes are defined as shown in the following formula:

$$F_1 = \frac{2 \times P \times R}{P + R},$$ \hspace{1cm} (18)

$$MSE = \frac{\sum_{i=1}^{N} (\text{target} - \text{output})^2}{N}.$$ \hspace{1cm} (19)

4.2. Application Test of Dance Generation. According to the application’s demand analysis and structural design, this section describes in detail the testing of each functional module during the testing phase. For each test case, we must repeatedly test each function multiple times, and all tests must pass for the case to be deemed successful. Specific test results are detailed in Table 1.

4.3. Validity of Correspondence. Choose the optimal correspondence based on the accuracy of the learned correspondence between music and action during the training phase. In the automatic generation phase of dance movements, the validity of the corresponding relationship is verified. In the phase of automatic generation of dance movements aimed at particular types of dances, such as social dance, the experimental data set’s corresponding relationship between music and movements is used to generate dance movements. Using tap dance and tap dance music as examples, Figure 6 illustrates the corresponding relationship between music and action and its accuracy (where $T_1$ is energy root mean square, $T_2$ is action distance, $T_3$ is footstep track, $T_4$ is fundamental frequency mean, $T_5$ is left ankle joint, $T_6$ is right ankle joint, $T_7$ is left knee joint, and $T_8$ is right knee joint).

The correspondence accuracy is low when there are few matching features between music and action, as depicted in Figure 6. The correspondence accuracy will gradually improve as the number of matching feature pairs increases, while matching feature pairs with low accuracy will decrease the correspondence accuracy between music and action.

This paper compares the dance synthesis accuracy between the bottom features matching and the high-level statistical features matching in order to evaluate the overall performance of music and action matching for the bottom features. By comparing synthetic dance to real dance using the dance synthesis accuracy formula, the accuracy of dance synthesis can be calculated.

Based on the experimental data presented in Figure 7, it is evident that, compared to the high-level statistical features, constructing the relationship between music and action using the lower-level features yields greater accuracy. High-level statistical features are unsuitable for all music and dance matching calculations.

4.4. Compared with Existing Methods. In order to better measure the recognition effect of the model in this paper on dance movements in video, two deep convolutional networks which have been widely used in industry are introduced: Inception V3 and 3D-CNN network. Figure 8 shows the performance comparison results of three networks.
It can be seen from the test results in Figure 8 that among the three networks, the worst $F_1$ index of the network is 0.681 of initiation v3, the middle one is 3D-CNN network, and the best one is TWCNN, which is 13.2% higher than that of initiation V3. MSE and $F_1$ are two negatively correlated indexes, and the data in the second column of Figure 8 better verify this data trend and prove the validity of the test results. According to the accuracy of dance action recognition in the third column of Figure 8, the two-way CNN algorithm proposed in this paper has been improved by 12.4% and 6.8% for Inception V3 and 3D-CNN networks, respectively.

5. Conclusion and Discussion

The traditional dance automatic generation system selects the most appropriate dance action segments from the constructed dance action database based on the input music characteristics and cannot generate dance action segments that are not present in the action database. This study proposes an automatic dance generation system based on deep learning. Using a well-designed network of generators and discriminators and a variety of loss functions as objective functions, it is possible to generate high-resolution and realistic dance videos. By traversing the action graph, we are able to search through the action segments and select those that correspond to the music segments. Following synthesis, the system simultaneously displays the input music segment sequence and the corresponding dance segment sequence. In this paper, the method for recognizing dance video motion is investigated. Through analysis of the traditional deep convolutional network, it has been determined that the method proposed in this paper has superior performance in the recognition of dance video motion than the traditional Inception V3 and 3D-CNN networks.

Table 1: System function test case.

<table>
<thead>
<tr>
<th>Test content</th>
<th>Expected result</th>
<th>Test result</th>
</tr>
</thead>
<tbody>
<tr>
<td>Local video read-in</td>
<td>When the user clicks the Read Local Video button, the system will pop up the file selection interface, and the system will display the first frame of the video after the user selects the video.</td>
<td>Verification by</td>
</tr>
<tr>
<td>Online video entry</td>
<td>When the user clicks the record button, the system will call the available camera and display the picture being recorded in real time.</td>
<td>Verification by</td>
</tr>
<tr>
<td>Perform dance generation</td>
<td>When the user clicks the start button, the system will display the currently synthesized picture in real time</td>
<td>Verification by</td>
</tr>
</tbody>
</table>

Table 1: System function test case.

**Mathematical Problems in Engineering**

**Figure 6: Accuracy of different mapping sets.**

**Figure 7: Comparison of the accuracy of correspondence between bottom features and high-level statistical features.**

**Figure 8: Comparison results of network performance of different models.**
optimize the structure of the time-domain convolution network and enhance the algorithm’s performance. In the future, we will introduce more advanced deep learning techniques to predict different dance movements [23, 26].

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The author declares that he has no conflicts of interest.

References