Research on Dance Movement Evaluation Method Based on Deep Learning Posture Estimation

Haoyu Tang
College of Computer Science and Cyber Security(Oxford Brookes College) Chengdu University of Techology
Chengdu, China
thy@stu.cdut.edu.cn

Yaohua Luo*
College of Computer Science and Cyber Security(Oxford Brookes College) Chengdu University of Techology
Chengdu, China
lyh@cdut.edu.cn

Jing Yang
College of Computer Science and Cyber Security(Oxford Brookes College) Chengdu University of Techology
Chengdu, China
yj@stu.cdut.edu.cn

2022 2nd International Conference on Big Data Engineering and Education (BDEE) | 978-1-6654-8184-7/22/$31.00 ©2022 IEEE | DOI: 10.1109/BDEE55929.2022.00036

***Abstract*—This study proposes a quantitative evaluation method of monocular video dance movements based on deep learning, and provides a scientific and quantitative dance movement evaluation system to help students and dance teachers quickly identify and judge their dance movement standards.In this paper, the neural network-based 3D human pose estimation method is used to realize the identification and collection of key points of video human pose and skeleton. Aiming at the problem that it is difficult for users to ensure the time consistency with the reference video in the actual video acquisition process, a video frame alignment method based on ant colony algorithm is proposed. Aiming at the problem of large coordinate displacement caused by the tester's height, fat and thinness, and different video shooting angles, the evaluation and analysis method of moving human body posture based on similarity matching between feature planes is used to solve the problem. The scores of the input movements, the user video and the benchmark video are output at the same time, and provided to dance learners, so that the students' "non-standard" dance movements can be identified in a timely and effective manner, and the evaluation results are fed back to the students in a friendly way, so that Students can correct dance movements in time to achieve better training results.**

Keywords—deep learning, dance assisted instruction, posture estimation, ant colony

I. INTRODUCTION

Dance education in universities is an important part of Chinese art education. Because dance teaching is a practical subject, we can use digital means and improve the evaluation system to stimulate the vitality of dance education and help the cultivation of dance education talents in universities[1, 2, 3]. In recent years, with the rapid development of computer technology and industry, computer virtual reality and visual technology have been widely used in all walks of life due to the development of computer graphics. Connecting computer virtual reality technology with dance teaching, and using digital means to assist dance teaching evaluation, bring new possibilities for further enhancing dance teaching efficiency and improving teaching effect.

1. According to the difference of human posture dimensions, human posture estimation tasks can be divided into two-dimensional human posture estimation and three-dimensional human posture estimation. The goal of 2D human pose estimation is to locate and recognize the key points of human body, and connect these key points according to the joint order to form a projection on the two-dimensional plane of the image, so as to obtain the human skeleton. The main task of 3D human pose estimation is to predict the three-dimensional coordinate position and angle of human joint points [4].

II. RELATED WORK

In terms of methods, three-dimensional human pose estimation methods can be divided into traditional methods and deep learning methods. The most obvious feature between the traditional 3D human posture estimation and the posture estimation based on deep learning is whether the learning method of multilayer neural network is used. Because of the different modeling methods, there are also great differences in estimation accuracy, computational complexity and so on[5]. Modeling is a very important aspect of 3D human pose estimation, and its purpose is to represent the key points and features extracted from the input data. In solving practical problems, the complexity of the environment in which the experimental individuals live increases the difficulty of model building, so it is very important to select appropriate and effective image features to simplify the process of model building[6]. Traditional methods mostly use the method based on human model to describe and infer human posture, and extract image posture features through algorithm. Therefore, there are relatively high requirements for the two dimensions of feature representation and the spatial position relationship of key points. Except for low-level features such as boundary and color, typical high-level features with stronger expression ability, such as scale invariant feature transformation and gradient histogram, can effectively compress the spatial dimension of features, Although they have advantages in time efficiency, they are still the traditional characteristics of manual design, with the following problems:

2. Some details of the image are lost. Traditional recognition methods will be seriously disturbed due to the limitations of occlusion and inherent geometric fuzziness, which seriously limits its scope of application.

3. Traditional methods require high quality of image and video data. Whether using multi camera or monocular camera, it's easy to be affected by acquisition cost, occlusion, illumination, environment and other factors.

978-1-6654-8184-7/22/$31.00 ©2022 IEEE
DOI 10.1109/BDEE55929.2022.00036

With the rapid rise of in-depth learning technology, computer vision technology based on in-depth learning is maturing. Video motion capture technology based on in-depth learning (human posture estimation technology) has been paid more attention due to its low cost and good estimation effect. It has been widely used in monitoring, behavior perception and virtual reality in recent years [7]. The in-depth learning model is relatively simple to operate and has a strong ability to represent features. It automatically extracts features from input information without manual feature extraction.

III. METHODOLOGY

A. Three Dimensional (3D) Human Posture Estimation Method

In this study, videoPose3D framework is used as the main framework of three-dimensional human posture estimation[8]. It is a method based on video frame sequence. The context information provided by adjacent frames in the video can provide more information, better predict the posture of the current frame, and reduce the occurrence of multi solution problems.

VideoPose3D model uses temporal revolutionary network to process sequence information and output 3D pose. The essence of TCN is convolution in the time domain[9]. Its biggest advantage over RNN is that it can process multiple sequences in parallel, and the computational complexity of TCN is low, with less model parameters. In VideoPose3D, cavity convolution is used to expand the receptive field of TCN. The specific network structure is similar to that of simple baseline3D[10], that is, the full convolution network with residual connection is adopted[11]. In addition, VideoPose3D also includes a semi supervised training method. The main idea is to add a trajectory prediction model to predict the absolute coordinates of the root joint, and project the absolute 3D pose in the camera coordinate system back to the 2D plane, thereby introducing re-projection loss. Semi supervised method can better improve the accuracy when 3D label is limited.Keep your text and graphic files separate until after the text has been formatted and styled. Do not use hard tabs, and limit use of hard returns to only one return at the end of a paragraph. Do not add any kind of pagination anywhere in the paper. Do not number text heads-the template will do that for you.

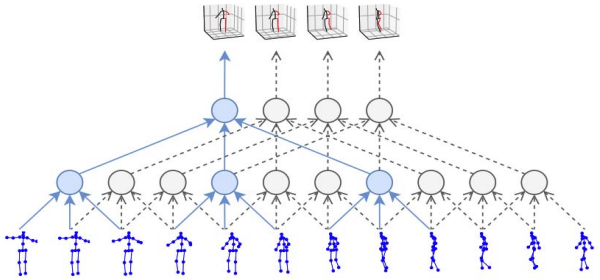

Fig. 1. VideoPose3D model framework[8]

This research is supported by Higher education talent training quality and teaching reform project of Chengdu University of Technology(JG2130225)

Algorithm Process

1. In order to reduce the amount of calculation, select the corresponding bone feature point data within 1 minute from the beginning of the two videos as the input data.

2. Initialize relevant super parameters, such as ant number m, pheromone factor α, heuristic function factor β, pheromone volatilization factor ρ, pheromone constant Q, maximum number of iterations t, threshold of characteristic attitude difference thr, etc.

3. Construct solution space. The corresponding feature point sequence of the reference video is divided into m parts. that is, each ant is responsible for finding the minimum feature attitude difference value and its corresponding time offset between its own interval and the corresponding feature point sequence of the video to be evaluated.

4. Update pheromone. Each ant is given a random time offset to calculate the characteristic pose difference value, and the pheromone concentration of the path is updated according to the magnitude of the characteristic pose difference value. In the next iteration, the ant uses roulette method to select the selection path (time offset). Different from the traditional ant colony algorithm, if it hits the attempted time offset, it randomly selects a time offset in the neighborhood of the time offset to perform the calculation. The size of the neighborhood is inversely proportional to the characteristic attitude difference corresponding to the central time offset. Run to convergence. If the ant fails to find a time offset after the specified number of iterations, so that the corresponding characteristic attitude difference value is less than the threshold, the corresponding time interval of the ant is discarded. During iteration, the calculation process of each ant is independent of each other, so the calculation of each ant can be allocated to multiple CPUs to complete in parallel.

5. Calculate the optimal solution. Collect the optimal time offset obtained by each ant, bring the above time offset into all the time intervals that have not been discarded, calculate the characteristic posture difference, and select the time offset corresponding to the minimum characteristic posture difference value as the optimal alignment time difference.

Fig. 2. Algorithm process of video alignment method based on ant colony algorithm

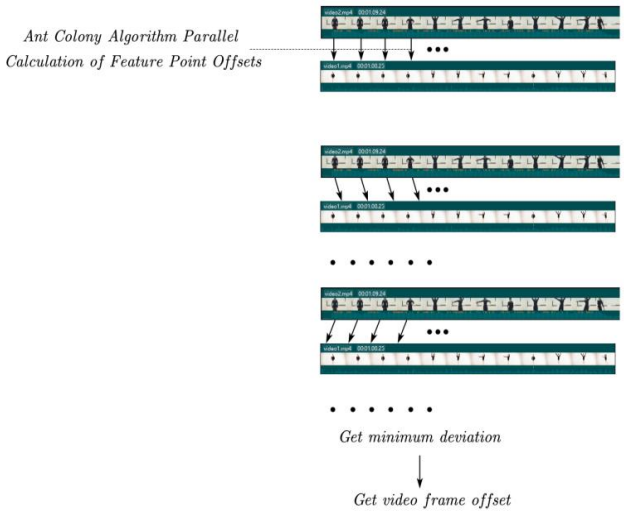

Fig. 3. Video frame offset calculation process

B. Video Alignment Method Based on Ant Colony Algorithm

In the Dance Evaluation System, how to align the video recorded by the tester with the base video of the standard dance posture is a big challenge. This paper presents a time alignment

method based on ant colony algorithm and feature points. The implementation is as follows.

Audio cannot be used to locate video frames for videos that do not record background music or have noisy background sound. Therefore, an alignment method based on parallel optimization of ant colony algorithm is proposed. This method uses the feature point sequence obtained from human posture estimation as input data. After adding a time offset to a feature point sequence, the feature posture difference is calculated frame by frame for the feature point collection. After summing up the feature posture difference, the corresponding value of the feature posture difference is obtained. If a time offset is found so that the characteristic posture difference value corresponding to the offset is the smallest of all the offsets, then the offset can be considered as the time alignment offset of the video, so the video alignment problem based on skeletal key point information can be considered as an optimization problem. To improve the efficiency of the algorithm, an ant colony algorithm[12] based on parallel optimization is used to speed up the algorithm process. Ant colony algorithm is a heuristic algorithm, which is a probability algorithm used to find the optimal path in the graph. It's inspired by ants finding their way through food. Ants release a substance called "pheromone" during their walk to identify their own path. When looking for food, the ant colony chooses its walking direction based on the concentration of the "pheromone" and eventually reaches where the food is located. Pheromones evaporate over time. When solving the video alignment problem, we will assign a larger "pheromone" to the time offset with smaller characteristic posture difference values, and the next round of ants will prefer to find a better solution near the time offset with smaller characteristic posture difference values.

C. *Dance Posture Evaluation Method Based on Feature Plane Matching*

After obtaining the key feature points of the human skeleton through the human posture estimation method, it is necessary to evaluate the similarity between the key feature points of the tester's skeleton and the key feature points of the dancer in the benchmark video on the same alignment time frame. In the motion trajectory of feature points, the motion data curves of each marked point are independent of each other. The comparison method based on Euclidean distance is to compare the data of two actions, and the corresponding distance value will be obtained in the comparison process of each action sequence. According to the comparison of the preset threshold value, the percentage of data matching degree can be obtained. However, the amount of calculation of this method is too large. At the same time, taking the displacement of feature points as a reference standard has little significance in the process of dance posture comparison, and can not truly reflect the accuracy of the comparison. This experiment uses a similarity calculation method based on feature plane matching[13]. The similarity between the tester and the standard action can be attributed to the similarity of object geometry. Three feature points are known to determine a feature plane, and the skeleton of human posture is mainly composed of seven feature planes, as shown in the shadow of Fig. 5.

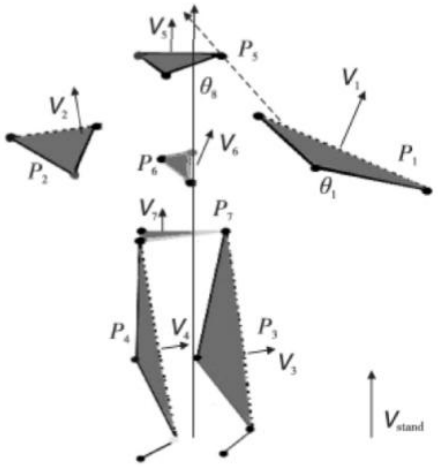

Fig. 4. Feature surface composed of feature keys[13]

The vector angle relationship between joint points is used to evaluate the difference between dancer's action and standard action. The specific calculation method is as follows:

1. Calculate cosine similarity. According to the principle of ergonomics, the human body takes the spine as the main axis, the spine as the z-axis of the space rectangular coordinate system, and the x-axis and Y-axis of the horizontal plane as the ground plane. The standard analysis of human motion can be simplified as the comparison of the similarity of edge vectors of the same plane and the comparison of the similarity of normal vectors between planes. The cosine similarity is used as the similarity function. By measuring the cosine value of the angle between the inner product of two vectors in the space, the similarity between them can be measured. Compared with Euclidean distance measurement, cosine similarity measurement pays more attention to the difference in the direction between two vectors. The formula is as follows:

$$similarity(\theta_i) = \frac{\sum_{i=1}^{n} A_t \times B_t}{\sqrt{\sum_{t=1}^{n}(A_i)^2} \times \sqrt{\sum_{t=1}^{n}(B_i)^2}} \quad (1)$$

Where: θ_i is the joint angle, A_t . B_t is the edge vector of the corresponding feature plane. The calculated cosine value is [0,1]. If the value is close to 1, it indicates that the action of the dancer to be tested is consistent with the standard action, and the dance is standardized. If the value is close to 0, it indicates that the deviation between the dancer's action to be tested and the standard action is too large.

2. Calculate the correlation parameters of key actions. Cosine similarity can not only measure the difference in the direction between vectors, but also measure the similarity and difference between angles. There are individual differences between human bodies, such as height, weight, arm length and other problems, but the proportion of human bodies is certain. Therefore, through the angle similarity, we can also measure whether the motion amplitude of the limb meets the standard. The calculation formula is as follows:

$$corr(A,B) = 1 - (\frac{arccos(similarity(\theta_i))}{\pi}) \quad (2)$$

3. Calculation of characteristic attitude difference. For two video frames of each synchronization time, the correlation coefficient corr of each part of its action is calculated; Secondly, the correlation coefficient error is calculated in the form of relative error, and its calculation formula is as follows:

$$\Delta corr_i = \frac{|corr_i - corr_j|}{corr_i}\% \quad (3)$$

Where: A is the time point of the synchronous video frame; B is the relative coefficient of dance movement standard; C is the relative coefficient of the dance action to be tested. The basis for judging the action standard is the relative error of calculating the correlation coefficient, and the condition of error convergence is:

$$\Delta corr_i \leq C \quad (4)$$

Where: $C$is the selected error threshold, which can be modified according to the teacher's requirements for dance movements.

IV. Experiment

This experiment platform is a PC with AMD R5 5600X CPU, 48GB memory, NVIDIA RTX 3070 GPU× 2, and use Ubuntu 20.04 operating system. Using Python 3.7 language as the main development environment, using JavaScript and web technology to achieve the visualization part, to improve the efficiency of parallel development, using Go lang as the implementation language of ant colony algorithm alignment method based on parallel optimization. The dance database built contains five dance action segments (each dance segment contains one baseline video and multiple video to be evaluated), each of which lasts from 1 to 3 minutes. The subjects of the experiment were selected as college students majoring in dance. The specific experimental steps are as follows:

A. Dance Video Collection

First ask the dance teacher to record the standard dance movements using a mobile phone (fixed position), with a video frame rate of 24 frames per second. The subjects were then asked to imitate the dance teacher's actions, perform the same dance tracks, and take pictures with their mobile phones. Use this method to record several different dance segments.

Fig. 5. Standard videos and videos to be evaluated

B. Three-dimensional Human Pose Estimation

Human skeletal features were extracted using the mmpose framework [14] and the VideoPose3D model. In this paper, using VideoPose3D model trained by Human 3.6M dataset as a three-dimensional human posture estimation model [15], the video input model is used to extract skeletal features, and the human skeletal key points for each video frame in the video are obtained. Record three-dimensional feature point coordinates.

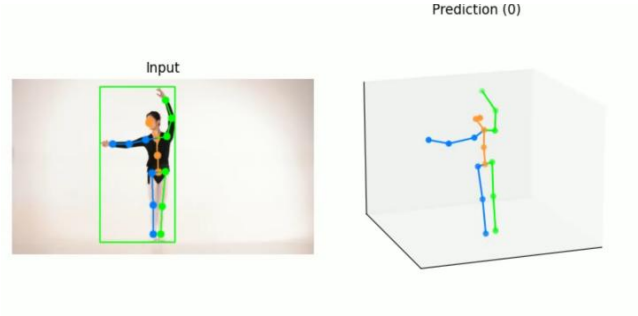

Fig. 6. 3-D human posture estimation results

C. Video Frame Alignment

Using the video frame alignment method based on ant colony algorithm, the time deviation between the base video and the video to be evaluated is calculated to align the base video with the video to be evaluated.

D. Dance Match Evaluation

Two sets of skeletal feature sets after alignment are input into the evaluation module, which is based on the similarity calculation method of feature plane matching. The similarity score of human body parts (head, trunk, left and right arms, left and right legs) corresponding to each video frame is calculated, and the overall similarity score under the current video frame is calculated by weighted average. Computations are performed on each video frame to obtain the corresponding similarity score for each video frame. A sample set is constructed by counting the scores of multiple testers at each frame, and the normal distribution probability density function of the sample set is calculated. All the scores are converted into scores that satisfy the normal distribution within the range of [0,100]. Using the results of the teacher's manual evaluation, adjust the super parameters to make the model evaluation close to the real evaluation results.

E. Evaluation Result Display

Based on the time difference between the two videos calculated earlier, a video clipping program is written based on the FFmpeg, which aligns the base video with the video to be evaluated, plays the base video and the video to be evaluated simultaneously, and displays the similarity score corresponding to the current frame in real time. Human posture estimation may result in strong jitter due to video shaking, occlusion of the subject or a brief departure from the field of view of the lens, resulting in abnormal scores in a short period of time. To avoid the influence of the above on the result of dance pose evaluation and reduce the disturbance of abnormal frame score on the total score, the time windowed method is used to smooth each frame score.

Fig. 7. 3-D Human Posture Estimation Results

V. CONCLUSION

In this paper, a monocular visual 3D human posture estimation technology based on in-depth learning is used to track and detect dancing movements in milliseconds. Video data is used only and no sensor is needed, which greatly reduces the cost of using the system while ensuring the accuracy of feature point acquisition is available. A feature alignment method based on ant colony algorithm is proposed to achieve video alignment when there is a "video misalignment" between the collected video and the base video start time. To solve the problem of coordinate displacement offset caused by different height and weight of trainees, the feature point vectors are standardized based on the similarity matching between feature planes to achieve uniform score. This evaluation system can accurately analyze the dancer's posture, effectively reduce the burden of dance teachers, improve the quality of teaching, and play an important role in the digital teaching and development of dance discipline.

REFERENCES

[1] Juanli Huang, Fengfu Ye. Learning Evaluation of Dance Course for Music Education Major [J]. Higher Education Journal(Chinese), 2020 (07): 55-57.

[2] Taozi Liu. Research on quantitative teaching evaluation of dance courses in Secondary Vocational Colleges in Hunan Province [D]. Central China Normal University, 2018.

[3] Maotuo Hua. The Construction of Multi-teaching Evaluation System of Dance Discipline [J]. Sichuan Drama(Chinese), 2020 (11): 159-161.

[4] Faming Wang, Jianwei Li, Sixi Chen.A survey of three-dimensional human pose estimation[J]. Computer Engineering and Applications(Chinese), 2021,57(10):26-38.

[5] Lun Zhang. Research on Algorithms of Image-based 3D Human Pose Estimation[C]. University of Electronic Science and Technology of China, 2022.

[6] ZhenHua Tang. Research on 3d human posture estimation algorithm based on deep learning in monocular image[C]. Shenzhen University, 2020.

[7] Liu Yujie, Zhang Minjie, Li Zongmin, Li Hua. Lightweight Human Posture Estimation Based on Global Posture Perception [J/OL]. Journal of Graphics(Chinese): 1-11 [2022-04-10].

[8] Pavllo D, Feichtenhofer C, Grangier D, et al. 3d human pose estimation in video with temporal convolutions and semi-supervised training[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 7753-7762.

[9] Lea C, Vidal R, Reiter A, et al. Temporal convolutional networks: A unified approach to action segmentation[C]//European Conference on Computer Vision. Springer, Cham, 2016: 47-54.

[10] Julieta Martinez, Rayat Hossain, Javier Romero, James J. Little; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2640-2649

[11] Szegedy C, Ioffe S, Vanhoucke V, et al. Inception-v4, inception-resnet and the impact of residual connections on learning[C]//Thirty-first AAAI conference on artificial intelligence. 2017.

[12] Dorigo M, Birattari M, Stutzle T. Ant colony optimization[J]. IEEE computational intelligence magazine, 2006, 1(4): 28-39.

[13] Li Han, Lucheng Wang, Meichao Zhang, et al. A real-time pose analysis method of dance based on feature vector matching[J].Application Research of Computers(Chinese):33(12): 3892-3896, 2016.

[14] Contributors M M P. Openmmlab pose estimation toolbox and benchmark[J]. 2020.

[15] Ionescu C, Papava D, Olaru V, et al. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments[J]. IEEE transactions on pattern analysis and machine intelligence, 2013, 36(7): 1325-1339.