Feature data analysis of dance movements by motion capture

Chao Lu
School of Music, Handan University, Handan, Hebei, 056002, China
E-mail: *lcluc@hotmail.com*

Received 23 December 2024; accepted 23 February 2025; published online 6 May 2025
DOI https://doi.org/10.21595/jme.2025.24742

Abstract. Motion capture technology has been applied in more and more fields, but the research in the field of dance is relatively rare. In order to combine motion capture technology with dance research, better understand the characteristics of dance movements, and provide support for their digital analysis, this paper mainly studied the application of a motion capture technology called Kinect in the analysis of dance movement feature data. The skeleton data of different dance movements was first collected based on Kinect v2, and then the collected data was analyzed using a spatio-temporal graph convolutional network (ST-GCN). On the basis of the original ST-GCN, the multi-branch structure was adopted to realize co-occurrence feature learning, and the bone length feature and direction feature were introduced to further enrich the feature data. Experiments were carried out on the NTU RGB+D and dance datasets. It was found that the improved ST-GCN had better performance than other current motion classification approaches on the NTU RGB+D. The top-1 accuracy for cross-subject (CS) and cross-view (CV) was 92.4 % and 96.7 %, respectively, and the average accuracy of different dance movements for the dance dataset was 96.035. The findings confirm the effectiveness of the proposed approach in the analysis of dance movement feature data, and it can be applied in the actual research of dance movements.

Keywords: motion capture, dance movement, spatio-temporal graph convolution, classification effect, feature data.

1. Introduction

With the rapid advancement of computer technology, computer vision has also been applied in more and more fields such as sports training. With the assistance of computer technology, it is possible to perform more deeper research of human posture [1]. Motion capture technology has a wide range of application value [2], which can provide kinematics, dynamics, and other parameters for medical rehabilitation [3]. In sports training, it can assist coaches to realize the monitoring of athletes and obtain some action parameters for accurate technical guidance [4]. At present, the application of motion capture has become an issue of wide concern to researchers [5]. Giannakeris et al. [6] used multi-modal sensor data to solve the task of human action identification and found that it achieved effective recognition in medical platform services. Yuanze et al. [7] proposed an approach to motion recognition for healthcare quality control, developed an improved anchor-to-joint network, and conducted experiments on open source ITOP and healthcare resource estimation datasets. The approach achieved satisfactory real-time performance and accuracy. Balasubramanian et al. [8] established a human motion recognition architecture based on a wearable sensor network and used a convolutional neural network with long short-term memory (LSTM) to identify patient activities. They found that the method achieved 99.53 % accuracy. Xu et al. [9] put forward an attention-based multistage co-occurrence graph convolution-LSTM method based on 3D bone sequences to realize motion recognition and found that it achieved better performance than mainstream methods through experiments on datasets. In dance research, motion capture technology can provide support for the digital protection of national dances and provide technical guidance for dancers in actual training, which is of great research significance [10]. However, at present, the research of motion capture technology in the field of dance is

relatively rare. Therefore, based on the motion capture technology Kinect, this paper analyzed the feature data of dance movements and realized the classification of different dance movements. The combination of motion capture technology and dance movement analysis provides strong support for the informatization and digitalization of dance research.

2. The motion capture of dance movements

Many of the current motion capture technologies have relatively strict requirements for clothing and lighting, and the operation process is also complex. Kinect is an unmarked motion capture system [11] that can collect bone point motion data without wearing any sensor, which has advantages of low cost and easy implementation, so it has been more and more widely used in the field of motion analysis [12].

Kinect v2 used contained a variety of sensors such as high-definition cameras, which can realize real-time tracking of skeleton information. Table 1 displays its main parameters.

Table 1. Kinect v2 parameters

	Parameter
Red, green, and blue (RGB) camera	Resolution: 1,920×1,080
	Frequency: 30 frames per second
Depth camera	Resolution: 512×424
	Frame rate: 30 frames per second
	Field of view angle: horizontal 70°, vertical 60°
	Detection range: 0.5 m-4.5 m
	Data interface: USB3.0

Kinect v2 can capture the three-dimensional coordinate information of 25 human bone articulation points. However, in the actual collection of dance movements, the collection effect of finger bone points is not good due to the influence of clothing and movements. Moreover, considering that finger movements have little influence on the overall change of dance movements, the finger bone data was excluded in the actual collection. Only 20 articulation points were retained, as shown in Table 2.

Table 2. 20 articulation points

1	Shoulder left	11	Wrist right
2	Shoulder right	12	Hand left
3	Shoulder center	13	Hand right
4	Spine	14	Knee left
5	Hip left	15	Knee right
6	Hip right	16	Ankle left
7	Hip center	17	Ankle right
8	Elbow left	18	Foot left
9	Elbow right	19	Foot right
10	Wrist left	20	Head

Data were collected from 50 healthy dancers, all of whom had no strenuous exercise 24 h before the experiment and were in good health. In the laboratory, the Kinect camera was 1-1.2 m away from the ground and 2-3 m away from the subjects. The subjects stood directly in front of the Kinect, and the collection site had a wide field of vision without occlusion, so as to ensure that Kinect could completely collect all bone points.

3. Feature data analysis based on spatio-temporal graph convolution

The spatio-temporal graph convolution network **(**ST-GCN) has good performance in data prediction [13], image classification [14], and other aspects. It can automatically learn the

spatio-temporal features of skeleton sequences and achieve end-to-end training. Compared with traditional GCN, ST-GCN has better performance. It takes the 3D coordinate sequence of human bone points as the input. After multi-layer spatio-temporal graph convolution operation, the feature processing is completed, and the classification of different actions is realized through the softmax layer.

First, spatio-temporal graph $G = (V, E)$ is established using skeleton sequence $G = (V, E)$, where V is the point of the skeleton sequence and E is the edge of the graph. In the time dimension, ST-GCN uses a temporal convolutional network (TCN) to complete the convolution operation, and its sampling range and function are:

$$B(v_{ti}) = \{v_{qj} | d(v_{tj}, v_{ti}) \leq D\}, \tag{1}$$

$$p(v_{ti}, v_{tj}) = v_{tj}, \tag{2}$$

where $d(v_{tj}, v_{ti})$ refers to the smallest path from v_{tj} to v_{ti}, $D = 1$, t refers to the frame in the motion video, i is the node in the skeleton sequence, v_{ti} is the joint vector of the i-th node in the t-th frame, and v_{tj} is the joint vector of the j-th node in the t-th frame.

The graph convolution formula in spatial dimension is:

$$f_{out}(v_{ti}) = \sum_{v_{tj} \in B(v_{ti})} \frac{1}{Z_{ti}(v_{tj})} f_{in}[p(v_{ti}, v_{tj})] \cdot w(v_{ti}, v_{tj}). \tag{3}$$

Based on the above equations, the graph convolution formula for each node can be obtained:

$$f_{out}(v_{ti}) = \sum_{v_{tj} \in B(v_{ti})} \frac{1}{Z_{ti}(v_{tj})} f_{in}(v_{tj}) \cdot w[l_{ti}(v_{tj})]. \tag{4}$$

However, ST-GCN is difficult to discover co-occurrence features in the processing of long-distance articulation points. In order to improve this point, this paper adds a co-occurrence feature learning branch on the basis of original ST-GCN to realize the learning of long-distance articulation co-occurrence features. The structure of the improved ST-GCN is displayed in Fig. 1.

In Fig. 1, a data with a shape of $N \times 3 \times T \times V \times M$ (N: the size of batch; 3: number of channels; T: skeleton sequence length; V: number of nodes; M: number of characters in the sequence) is entered. After being normalized by the BatchNorm layer, it is input to the nine-layer ST-GCN for feature processing. The output channels of the first three layers, the middle three, and the last three layers were 64, 128, and 256, respectively, and then the output is sent to the pooling layer for global pooling to obtain a 256-dimensional feature vector. After passing through the softmax layer, the classification is completed. Moreover, after passing through the third ST-GCN layer, a data of $N \times 64 \times T \times V \times M$ is output, and it enters the co-occurrence feature learning branch. After dimension swapping, a data with a shape of $N \times V \times T \times 64 \times M$ is obtained. After three-layer convolution and pooling, the classification results are obtained. The two branches share the parameters of the first three layers and jointly optimize the cross-entropy loss function:

$$L = -\sum_{i=1}^{N} y_i \log y'_i, \tag{5}$$

where y_i is the real value and y'_i is the predicted value.

The classification of the improved ST-GCN is determined jointly by two branches:

$$l_{out} = \lambda \cdot l_{gcn} + (1 - \lambda) \cdot l_{cooc}, \tag{6}$$

where l is the classification result of different branches and λ indicates the weight of different branches.

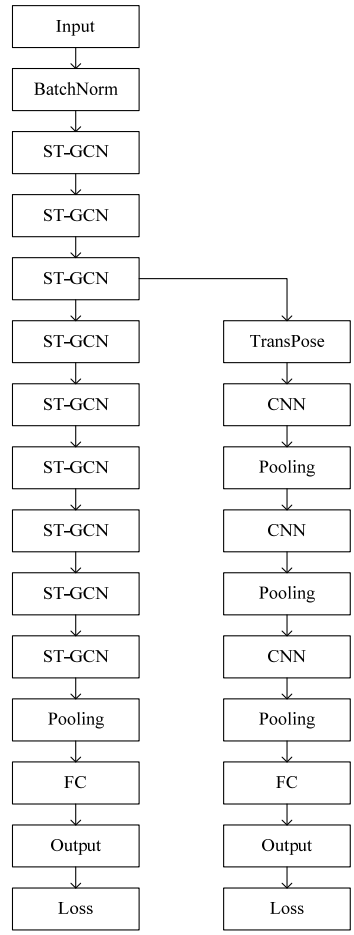

Fig. 1. Improved ST-GCN

In addition to the improvement of ST-GCN, the skeleton features are further analyzed in this paper. In the improved ST-GCN, only the coordinate information of the articulation point is used, and using more skeleton features can further enhance the classification ability of different dance movements. The point that is closer to the center of gravity of the skeleton is defined as source articulation point $v_1 = (x_1, y_1, z_1)$, and the point that is furthest away from the center of gravity of the skeleton is defined as articulation endpoint $v_2 = (x_2, y_2, z_2)$. The bone length and direction features are added in the classification to improve the classification effect:

(1) Bone length feature: $m = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2}$,

(2) Bone direction feature: $(\cos\alpha, \cos\beta, \cos\gamma) = \left(\frac{x_2 - x_1}{m}, \frac{y_2 - y_1}{m}, \frac{z_2 - z_1}{m}\right)$.

As shown in Fig. 2, after collecting the three-dimensional coordinates of the 20 bone joints through Kinect v2, the bone length features and bone classification features were calculated and input into the improved ST-GCN for learning. Finally, the movement classification result was obtained using the software classifier.

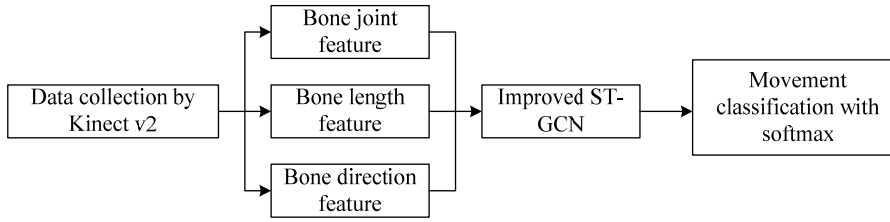

Fig. 2. The improved ST-GCN-based movement classification method

4. Results and analysis

4.1. Experimental setup

The experiment was conducted on a Windows 10 operating system with i5-13600KF processor, and the algorithm was implemented based on Python 3.7 and PyTorch 1.8. In the training stage, a stochastic gradient descent (SGD) optimizer was used, the initial learning rate was 0.1, the epoch was 80, the batch size was 8, and λ was 0.6. The experimental datasets are as follows.

(1) NTU RGB+D dataset [15]: it was collected with Kinect v2, including 60 movements performed by 40 actors, totally 56,880 samples. It was divided using two criteria: (1) cross-subject (CS): the dataset was divided according to the character' identity, 40,320 clips of 20 actors were used as the training set, and 16,560 clips of another 20 actors were used as the verification set; (2) cross-view (CV): the dataset was divided according to cameras, 37,920 clips collected by two cameras were taken as the training set, and 18,960 clips collected by the other camera was taken as the verification set.

Fig. 3. The first movement in the dance called returning to the nest

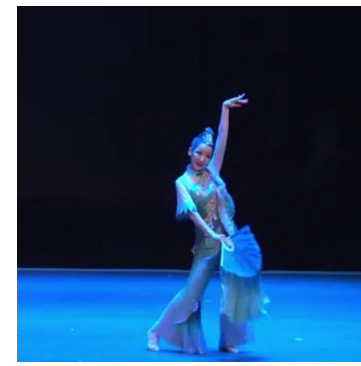

Fig. 4. The second movement in the dance called returning to the nest

Fig. 5. The third movement in the dance called returning to the nest

(2) Dance dataset: The method described in (1) was used to collect data from a dance called returning to the nest, which adopted Jingxing Lahua, a folk dance in Hebei Province, as the dance element. It told about the journey of magpie returning to the nest and presented the theme of animal protection. Three movements (Figs. 3-5) were selected for the classification study of dance movements. All subjects had a good command of these three movements and could perform them smoothly according to the music. Fifty subjects performed three dance movements in turn, and each movement performed 20 times. A total of 3,000 samples were obtained, 2,000 of which were taken as the training set and 1,000 as the verification set.

4.2. Result analysis

First, on the NTU RGB+D, the improved ST-GCN method was compared with some other current movement classification methods based on skeleton data. The top-1 accuracy of different approaches is presented in Table 3.

Table 3. Comparison using the NTU RGB+D

	CS	CV
Hierarchical recurrent neural network [16]	59.1 %	64.0 %
Deep LSTM [17]	60.7 %	67.3 %
ST-LSTM [18]	69.2 %	77.7 %
Spatio-temporal attention (STA)-LSTM [19]	73.4 %	81.2 %
TCN [20]	74.3 %	83.1 %
ST-GCN [21]	81.5 %	88.3 %
AS-GCN [22]	86.8 %	94.2 %
Attention-enhanced graph convolutional (AGC)-LSTM [18]	89.2 %	95.0 %
Directed graph neural network (DGNN) [23]	89.9 %	96.1 %
Multi-scale spatial temporal (MST)-GCN [24]	91.5 %	96.6 %
The improved ST-GCN	92.4 %	96.7 %

It can be found that the hierarchical recurrent neural network method had poor classification performance on the NTU RGB+D, with an accuracy of 59.1 % for the CS and 64.0 % for the CV. Compared with the hierarchical recurrent neural network method, the accuracy of several LSTM methods was improved to varying degrees, but the accuracy of the STA-LSTM method for the CV was more than 80 %, reaching 81.2 %. Compared with the previous RNN and LSTM methods, the methods based on GCN showed better performance in the classification of skeleton sequences, all of which achieved an accuracy more than 80 %. The accuracy of the improved ST-GCN method was 92.4 % for the CS and 96.7 % for the CV, which was 0.9 % and 0.1 % higher than the optimal GCN method, i.e., MST-GCN. These results verified the effectiveness of the improved ST-GCN method in movement classification.

Then, the proposed method was applied to the dance dataset to classify different dance movements, and the improvement effect of the ST-GCN method was compared.

Table 4. Results on the dance dataset

	Movement 1	Movement 2	Movement 3	Average
ST-GCN	91.25 %	90.77 %	91.34 %	91.12 %
Multi-branch ST-GCN	93.45 %	92.56 %	93.77 %	93.26 %
Multi-branch ST-GCN+bone length and bone direction features	96.54 %	95.47 %	96.07 %	96.03 %

As shown in Table 4, the ST-GCN's classification accuracy for the movements selected was all above 90 %, and the average classification accuracy of different movements was 91.12 %. After the original ST-GCN method was improved into a multi-branch structure, the average value reached 93.26 %, showing an increase of 2.14 %. After skeleton features were enriched on this basis by introducing bone length and direction features, the classification accuracy of the

algorithm was further improved to 96.03 %, which was 2.77 % higher than that of the multi-branch ST-GCN method, and the classification accuracy of each movement was all above 95 %. These results verified the reliability of the improved method in analyzing dance movement feature data and its accuracy in classifying different movements in the dance called returning to the nest.

5. Conclusions

In this paper, the skeleton data of various dance movements were collected by Kinect motion capture technology, and an improved ST-GCN method was designed to realize the feature analysis of skeleton data. Through experiments on the NTU RGB+D and dance datasets, it was found that this method had a superior accuracy compared with the current movement classification methods, and its average accuracy for the dance dataset reached 96.03 %. The proposed approach can be further applied in practice to provide guidance for the digital preservation and scientific training of dance.

Acknowledgements

The authors have not disclosed any funding.

Data availability

The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

Conflict of interest

The authors declare that they have no conflict of interest.

References

[1] A. Sabater, L. Santos, J. Santos-Victor, A. Bernardino, L. Montesano, and A. C. Murillo, "One-shot action recognition in challenging therapy scenarios," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 2771–2779, Jun. 2021, https://doi.org/10.1109/cvprw53098.2021.00312

[2] L. G. Motti Ader, A. Keogh, K. Mcmanus, B. R. Greene, and B. Caulfield, "Human movement analysis: introduction to motion capture and applications for health," in *IEEE International Conference on Healthcare Informatics (ICHI)*, pp. 1–3, Nov. 2020, https://doi.org/10.1109/ichi48887.2020.9374307

[3] S. L. Raghu, R. T. Conners, C.-K. Kang, D. B. Landrum, and P. N. Whitehead, "Kinematic analysis of gait in an underwater treadmill using land-based Vicon T 40s motion capture cameras arranged externally," *Journal of Biomechanics*, Vol. 124, p. 110553, Jul. 2021, https://doi.org/10.1016/j.jbiomech.2021.110553

[4] A. Mazurkiewicz, "Biomechanics of rotational movement in off-ice figure skating jumps: applications to training," *Polish Journal of Sport and Tourism*, Vol. 28, No. 2, pp. 3–7, Jun. 2021, https://doi.org/10.2478/pjst-2021-0007

[5] G. Rodríguez-Vega, D. A. Rodríguez-Vega, X. P. Zaldívar-Colado, U. Zaldívar-Colado, and R. Castillo-Ortega, "A motion capture system for hand movement recognition," in *Lecture Notes in Networks and Systems*, Cham: Springer International Publishing, 2021, pp. 114–121, https://doi.org/10.1007/978-3-030-74614-8_13

[6] P. Giannakeris et al., "Fusion of multimodal sensor data for effective human action recognition in the service of medical platforms," in *Lecture Notes in Computer Science*, 2021, pp. 367–378.

[7] Y. Wang, H. Zhang, X. Wu, C. Kong, Y. Ju, and C. Zhao, "Lidar-based action-recognition algorithm for medical quality control," *Laser and Optoelectronics Progress*, Vol. 61, No. 12, pp. 306–314, 2024.

[8] K. Balasubramanian, A. V. Prabu, M. F. Shaik, R. A. Naik, and S. K. Suguna, "A hybrid deep learning for patient activity recognition (PAR): Real time body wearable sensor network from healthcare

monitoring system (HMS)," *Journal of Intelligent and Fuzzy Systems*, Vol. 44, No. 1, pp. 195–211, Jan. 2023, https://doi.org/10.3233/jifs-212958

[9] S. Xu et al., "Attention based multi-level co-occurrence graph convolutional LSTM for 3D action recognition," *IEEE Internet of Things Journal*, Vol. 8, No. 21, pp. 15990–16001, 2021.

[10] W. Zhen and L. Luan, "Physical world to virtual reality – motion capture technology in dance creation," in *Journal of Physics: Conference Series*, Vol. 1828, No. 1, p. 012097, Feb. 2021, https://doi.org/10.1088/1742-6596/1828/1/012097

[11] A. Bilesan, S. Komizunai, T. Tsujita, and A. Konno, "Improved 3D human motion capture using kinect skeleton and depth sensor," *Journal of Robotics and Mechatronics*, Vol. 33, No. 6, pp. 1408–1422, Dec. 2021, https://doi.org/10.20965/jrm.2021.p1408

[12] A. Balmik, M. Jha, and A. Nandy, "NAO robot teleoperation with human motion recognition," *Arabian Journal for Science and Engineering*, Vol. 47, No. 2, pp. 1137–1146, Sep. 2021, https://doi.org/10.1007/s13369-021-06051-2

[13] J. An et al., "IGAGCN: Information geometry and attention-based spatiotemporal graph convolutional networks for traffic flow prediction," *Neural Networks*, Vol. 143, pp. 355–367, Nov. 2021, https://doi.org/10.1016/j.neunet.2021.05.035

[14] B. Sun, H. Zhang, Z. Wu, Y. Zhang, and T. Li, "Adaptive spatiotemporal graph convolutional networks for motor imagery classification," *IEEE Signal Processing Letters*, Vol. 28, pp. 219–223, Jan. 2021, https://doi.org/10.1109/lsp.2021.3049683

[15] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "NTU RGB+D: A large scale dataset for 3D human activity analysis," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1010–1019, Jun. 2016, https://doi.org/10.1109/cvpr.2016.115

[16] Y. Du, W. Wang, and L. Wang, "Hierarchical recurrent neural network for skeleton based action recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1110–1118, Jun. 2015, https://doi.org/10.1109/cvpr.2015.7298714

[17] L. Sun, T. Su, C. Liu, and R. Wang, "Deep LSTM networks for online Chinese handwriting recognition," in *International Conference on Frontiers in Handwriting Recognition*, pp. 271–276, Oct. 2016, https://doi.org/10.1109/icfhr.2016.0059

[18] C. Si, W. Chen, W. Wang, L. Wang, and T. Tan, "An attention enhanced graph convolutional LSTM network for skeleton-based action recognition," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1227–1236, Jun. 2019, https://doi.org/10.1109/cvpr.2019.00132

[19] S. Song, C. Lan, J. Xing, W. Zeng, and J. Liu, "An end-to-end spatio-temporal attention model for human action recognition from skeleton data," in *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 31, No. 1, Feb. 2017, https://doi.org/10.1609/aaai.v31i1.11212

[20] T. S. Kim and A. Reiter, "Interpretable 3D human action analysis with temporal convolutional networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1623–1631, Jul. 2017, https://doi.org/10.1109/cvprw.2017.207

[21] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32, No. 1, Apr. 2018, https://doi.org/10.1609/aaai.v32i1.12328

[22] M. Li, S. Chen, X. Chen, Y. Zhang, Y. Wang, and Q. Tian, "Actional-structural graph convolutional networks for skeleton-based action recognition," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12018–12027, Jun. 2019, https://doi.org/10.1109/cvpr.2019.00371

[23] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Skeleton-based action recognition with directed graph neural networks," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7904–7913, Jun. 2019, https://doi.org/10.1109/cvpr.2019.00810

[24] Z. Chen, S. Li, B. Yang, Q. Li, and H. Liu, "Multi-scale spatial temporal graph convolutional network for skeleton-based action recognition," *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35, No. 2, pp. 1113–1122, May 2021, https://doi.org/10.1609/aaai.v35i2.16197

Chao Lu graduated from Jiangxi Normal University in 2016 and is now a dance teacher at the School of Music of Handan University. She is engaged in dance choreography and teaching and research of Chinese ethnic and folk-dance courses.