Deep Learning-Based Action Recognition Algorithm for Jiangsu Dan Opera Characters

JIABEI LI*, XIN YU
International College, Krirk University,
Bangkok 10220,
THAILAND
*Corresponding Author

Abstract: - With the rapid development of artificial intelligence technology, deep learning has achieved remarkable results in fields such as image recognition. This study takes Jiangsu Dan opera, an opera genre with unique regional cultural characteristics, as the object, and conducts an in-depth analysis of the artistic characteristics and character movements of Dan opera based on deep learning algorithms to clarify the importance and difficulties of movement recognition. By collecting a large amount of Dan opera image and video data, a dataset suitable for Dan opera character action recognition is constructed. Advanced deep learning models, such as deep learning ST-GCN network and OpenPose, a multi-person pose estimation algorithm, are used to train and optimize the dataset. The experimental results show that the proposed algorithm has high accuracy and robustness in Dan opera character movement recognition, can effectively identify different movement types, and provides a new technical means for the inheritance, protection, and innovative development of Dan opera.

Key-Words: - Deep learning, Opera, Dan opera, Action recognition, OpenPose, ST-GCN.

Received: July 19, 2024. Revised: February 21, 2025. Accepted: April 22, 2025. Published: August 12, 2025.

1 Introduction

As a unique performing art form, the highly condensed movement language and rich performance elements of opera have unique advantages for expressing characters' character, emotions, and thoughts. Through physical movements, facial expressions, voice intonation, and other means, opera actors vividly reproduce the inner world of dramatic characters, bringing the audience both visual and auditory experiences, [1]. At the same time, the art of opera is also an important part of the excellent traditional Chinese culture, and its unique cultural connotation and artistic value not only reflect the aesthetic pursuit of the Chinese people but also show the epitome of China's long history and profound culture, [2]. Therefore, the study of opera art not only helps to better inherit and develop this precious cultural heritage but also provides useful reference for other related fields.

With the continuous progress of computer vision technology, it has become possible to automate the analysis and understanding of opera performances using machine learning and other methods. As an important application of computer vision in the field of opera, opera character movement recognition can help us better understand and analyze the performance skills of opera actors, and provide new ideas for the inheritance and innovation of opera art. For example, through the automated analysis of opera movements, we can extract the characteristic movements of different schools and factions, which can provide references for the training of succeeding actors; at the same time, based on the feedback of movement recognition, it can also assist the choreographer in designing more vivid and interesting drama plots and movement designs, [3]. The current research work mainly focuses on action capture, action description, and action recognition, [4]. However, due to the complexity and diversity of the theatre movements themselves, the existing methods still have greater challenges in terms of accuracy and robustness, [5]. Therefore, how to design more robust and versatile action recognition algorithms while constructing high-quality opera action datasets remains a key issue to be solved.

As one of the important genres of Beijing opera, Danju opera has certain characteristics in its characterization and action performance. Dan's character often highlights his personality through exaggerated body movements and expressions, [6].

These distinctive opera actions are a big challenge to the traditional action recognition methods based on manual design features. The method based on deep learning ST-GCN network and a multi person pose estimation algorithm OpenPose is better at capturing these complex and diverse movement characteristics, and can more accurately identify the movements of the characters in Danju opera. On the one hand, the automatic action recognition technology based on deep learning can help record and analyze a large number of opera action data, and provide data support for the systematic research of opera art; On the other hand, this technology can also be applied to opera teaching and training, providing intelligent auxiliary means for the inheritance of opera art. In general, the value of in-depth learning in the protection and inheritance of traditional Chinese opera is multifaceted, which is worth our continued exploration and exploration.

2 Deep Learning-based Action Recognition Algorithm for Opera Characters

2.1 Algorithm Overall Framework Design

ST-GCN is a deep learning model for character action recognition in opera. The model adopts a network structure of spatio-temporal graph convolution, which can effectively capture the temporal and spatial features in character action sequences. The input of the network is the character action sequence and the output is the corresponding action category, [7]. The overall processing flow of ST-GCN is shown in Figure 1 (Appendix), where the input video is extracted using a pose estimation algorithm to obtain the skeleton sequence of the action, and then a graph-structured representation of the human body motion information is constructed based on the skeleton sequence information. Given the coordinates of the joints of the human body, they are connected as inputs to the ST-GCN network, which automatically extracts spatiotemporal features using multiple graph convolution modules, each consisting of a GCN and a TCN, and, finally, obtains the results of the action classification using a SoftMax classifier.

2.1.1 Structure of Skeleton

First, we need to determine the nodes and edges of the graph. Taking human motion recognition as an example, nodes are usually human joint points, such as the wrist, elbow, shoulder, etc. The edge determines the connection relationship between joint points according to the human bone structure. These connections can be represented by an adjacency matrix. According to the given rules, the human skeleton space-time sequence is connected into a space-time map $G = (V, E)$. Where V represents the set of joint points in the spatio-temporal map, including the set of natural joint points of the human body in each frame, as well as the set of the same node in consecutive frames, marked $V = \{V_{ti} | t = 1, \ldots, T, i = 1, \ldots, N\}$. N represents the number of human joint points on a frame, T represents the number of frames of continuous video, and represents the ith node on the T frame; In the spatio-temporal graph, E represents the set of edges, which is composed of two subsets of spatial edges and temporal edges: the first part is the edge connected by human joints in each frame, representing the physical connection of joint points, which is recorded as$E_s = \{v_{ti} v_{tj} | (i, j) \in H\}$, where H is a group of naturally connected human joints, and the second part is the edge connected by the same node in the frame, which is recorded as$E_F = \{v_{ti} v_{(t+i)i}\}$. The eigenvector of the ith node on the T frame is represented by$F(v_{ti})$, which represents the first-order information, that is, the coordinate position and confidence of the relevant node.

2.1.2 Spatial Map Construction

After determining the spatial map, it is necessary to integrate the time series data. Suppose we have a video sequence, and each frame has the corresponding position information of human joint points. These frames are arranged in time order to form the time dimension of the spatio-temporal map. For example, a dataset containing 10 frames of human motion video with 20 joint points per frame can build a spatio-temporal map, in which the spatial dimension is 20 joint points and the time dimension is 10 frames. For the center node v_{ti} and its neighbor node v_{tj} in a frame T, the weight matrix W is introduced to realize the shared linear transformation of nodes, so as to obtain higher dimensional features, as shown in Equation (1).

$$\begin{aligned} z_{ti} &= W f_{in}(v_{ti}) \\ z_{tj} &= W f_{in}(v_{tj}) \end{aligned} \tag{1}$$

where $f_{in}(v_{ti}) \in R^C$ and $f_{in}(v_{tj}) \in R^C$ represent the input eigenvectors of node v_{ti} and node v_{tj} respectively, and then the two vectors in the above formula are spliced to obtain a new dimension vector, as shown in Equation (2).

$$a_{tij} = Leaky\mathrm{ReLU}(\vec{a}^T [z_{ti} || z_{tj}]), j \in B(v_{it}) \tag{2}$$

where e_{tij} represents the importance of node j to node i in frame T, a represents a single-layer feedforward neural network, and finally uses a leakyrelu activation function.

Finally, the input data is normalized to ensure that the value range of the data is within the same range. Convert their attention values into attention weight coefficients through softmax, as shown in Equation (3).

$$a_{tij} = \frac{\exp(LeakyReLU(\vec{a}^T[z_{ti}||z_{tj}]))}{\sum_{k\in B(v_{ti})}\exp(LeakyReLU(\vec{a}^T[z_{ti}||z_{tk}]))} \quad (3)$$

2.2 Data Acquisition and Pre-Processing

The core of the algorithm lies in the training of an effective deep-learning model, which relies on a large amount of high-quality training data. The OpenPose multi-person pose estimation algorithm, first utilizes the CPM method to detect the location of the body's key points, [8]. The detection results are derived by predicting a heat map of all human body key points, which enables the detection of a Gaussian peak at each human body key point, which is predicted to be a particular body key point by a representative network. Similarly, the same results can be performed for all other key points to obtain detection conclusions. Once the detection results are obtained, additional detection results for key points can be connected. The entire OpenPose process is roughly depicted in Figure 2 as follows: firstly, a normal image is input, and the parsed results can be obtained by two side-by-side branches: the part confidence map and the part association field, after which the key points are matched with each other by two-part graph matching.

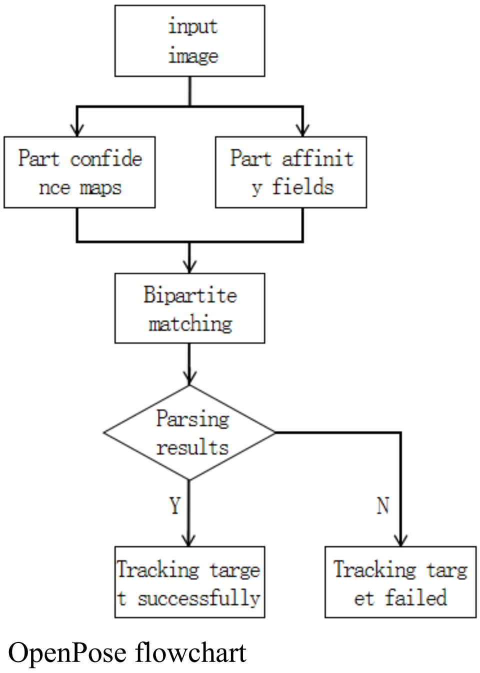

Fig. 2: OpenPose flowchart

OpenPose network structure can predict key points and partial affinity fields at the same time. The network structure is shown in Figure 3 (Appendix). First, the system uses the first ten layers of VGG19 to obtain the features of the picture and gets the feature map of the RGB image, marked as *F*, [9]. As the input of the network, the whole network is composed of multiple stages. For each stage, there are two branches, one of which can get the confidence set of key points, that is, branch 1 in the graph. Finally, the confidence graph set S, S=(S_1, S_2,..., S_j) is obtained, which represents the set of J key points of the human body, where SJ represents the probability of the jth part of the human body in the picture of h' × w'; The other branch predicts the connection between the key points, that is, Branch 2 in the figure, and finally gets a partial affinity field set L, L=(L_1, L_2,..., L_C) representing the C vector fields of the human body, that is, the logarithm of the C joints of the human body, where LC is the combined representation of the position and direction in an h' × w' × 2 picture.

2.2.1 Loss function

The initial stage of this network uses features from the first 10 layers of VGG as input data [10]. Subsequent stages then use the features F and the outputs St and Lt from the previous stage as input data, as shown in Equation (4).

$$S^t = \rho^t(F, S^{t-1}, L^{t-1}), \forall t \geq 2$$
$$L^t = \varphi^t(F, S^{t-1}, L^{t-1}), \forall t \geq 2 \quad (4)$$

where ρ and φ denote the CNN network at stage *t*. In this network architecture, the L_2 loss function is used as the optimization dependency for each branch. By iteratively optimizing the loss function, it can be ensured that the whole network eventually converges to a steady state, as in Equation (5):

$$f_s^t = \sum_{j=1}^{J}\sum_p W(p).||S_j^t(p) - S_j^*(p)||_2^2$$
$$f_L^t = \sum_{c=1}^{C}\sum_p W(p).||L_c^t(p) - L_c^*(p)||_2^2 \quad (5)$$

In this body part detection model, S_j^*represents the actual confidence map of the joints and L_c^* represents the real labeled body part affinity map. The binary mask *W(p)* = 0 for pixels with missing labeled information *p*. To obtain the final loss function for the whole network, the loss values of each stage need to be summed up as shown in Equation (6):

$$f = \sum_{t=1}^{T}(f_s^t + f_L^t) \quad (6)$$

2.2.2 Limb Part Affinity Fields

OpenPose uses joint confidence maps to detect human skeletal keypoints [11]. Given the location of a key point, generate a formula through Gaussian diffusion to represent the confidence map of the *j* key point of the *k* person.

The loss function of the whole network is obtained by summing the losses of each stage of the network as shown in Equation (7):

$$S^*_{j,k}(p) = \exp(-\frac{||p-x_{j,k}||_2^2}{\sigma^2} \qquad (7)$$

where *p* is the pixel coordinate, $x_{j,k}$ is the *j* site keypoint location for the kth person, and *σ* controls the spread of the peak. In the above equation, the closer *p* is to *x* the larger the value of S becomes. Ultimately, the highest confidence level of the score is obtained by taking the maximum value. As shown in Equation (8):

$$S^*_j(p) = \max_k S^*_{j,k}(p) \qquad (8)$$

For the detected body region points, how to connect them into a complete body form without knowing how many there are is a key challenge. To solve this problem, OpenPose proposes the concept of Part Affinity Fields (PAF) for human body. This method not only retains the position information but also contains the orientation information in the candidate region. The specific formula is as follows: if any point *p* is located on the limb *c* of the *k* person, the value of $L^*_{c,k}(p)$ is *v*; if *p* is not on that limb, the value is 0. See Equation (9).

$$L^*_{c,k}(p) = \begin{cases} v, if\ p\ on\ limb\ c,k \\ 0, otherwise \end{cases} \qquad (9)$$

where *v* is the unit vector from site j_1 to site j_2 of the person, and the formula expression for *v* is given in Equation (10):

$$v = \frac{(x_{j2,k}-x_{j1,k})}{||x_{j2,k}-x_{j1,k}||_2} \qquad (10)$$

The basis for determining whether *p* is on the limb is that the positions from point *p* to key points j_1 and j_2 are within a certain threshold. The following formula needs to be satisfied to ensure that point *p* is on the limb, as shown in Equation (11).

$$0 \le v.(p - x_{j1,k}) \le l_{c,k} \text{ and } | v_\perp.(p - x_{j1,k}) | \le \sigma_l$$
$$l_{c,k} = \| x_{j2,k} - x_{j1,k} \|_2 \qquad (11)$$

Among them, $v_\perp$ represents the unit vector of *v* in the vertical direction. $L_{c,k}$ is the pixel length of limb *c*, and σ_1 is the pixel width of limb *c*. Finally, merge all people's parts into one image, take the average vector of *k* people at point *p*, and finally obtain the keypoint affinity field of limb *c*, as shown in Equation (12).

$$L^*_c(p) = \frac{1}{n_c(p)} \sum_K L^*_{c,k}(p) \, (12)$$

Among them, $n_c(p)$ represents the non-zero vector number of all *k* individuals at point *p*.

2.3 Construction of Dan Opera Homemade Dataset

Because there are many kinds of Dan opera movements, but all kinds of movements are basically composed of basic standard movements, and these basic movements are crucial for Dan opera movement recognition, so this paper collects more than 200 videos of Jiangsu Dan opera performances from different theatres, covering more than 30 common movement categories, such as lifting the leg up high, turning around, kneeling down, and dancing the sleeve robe, and so on. Through manual annotation, a dataset containing more than 80,000 action clips was constructed. A solid data base was provided for subsequent algorithm training and evaluation.

In the data preprocessing process, the video size of the Dan drama dataset is adjusted to a resolution of 340×256, the video frequency is set to 30FPS, and then the OpenPose algorithm is used to obtain 18 joint points on each frame, [12]. A two-dimensional coordinate and each joint point confidence are obtained, and then the human skeleton key point coordinates and the confidence of each key point is saved as a json file, which corresponds to the format used for graph convolutional networks. By analyzing the canonical formats of the Kinetics dataset and NTU-RGB+D dataset.

2.4 Feature Extraction and Action Classification

With a well-trained deep learning model in place, it can be utilized to perform feature extraction and action classification on new action sequence data. Specifically, the input action sequence data is fed into the model's forward computation process, from which high-level semantic features are extracted. These features can well describe the spatio-temporal characteristics of the actions, such as the amplitude, speed, and rhythm of the actions. Immediately after that, these features are fed into the final fully connected layer of the model, and the probability distributions of various types of actions are output using the softmax function, [13]. Figure 4 shows the construction process of the dan drama dataset, firstly after OpenPose extracts the key point information of each frame of the human body, then according to the data specification format defined by Algorithm 1, the json file obtained by OpenPose is redefined into the canonical json file shown in Algorithm 1 according to the prescribed format, which can be converted into the input format of the spatiotemporal graphic convolutional neural network. By comparing these probability values, the final action category prediction results can be obtained.

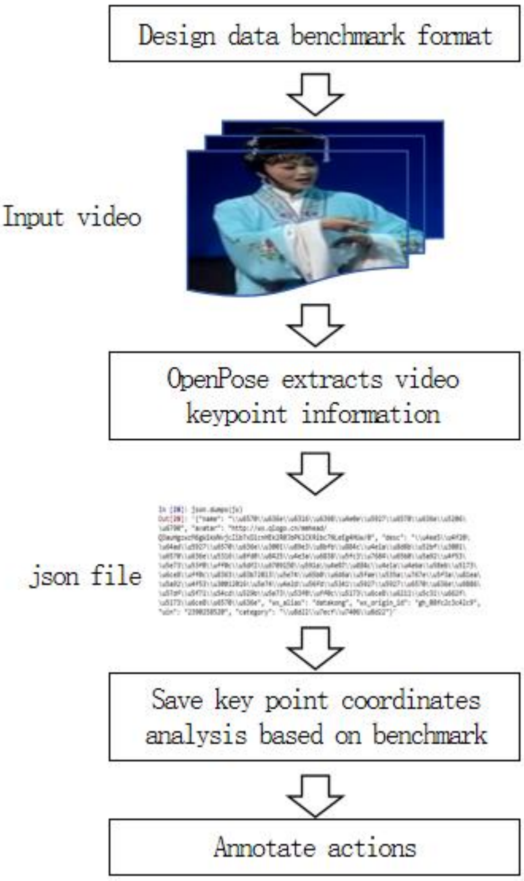
Design data benchmark format
Input video
OpenPose extracts video keypoint information
json file
Save key point coordinates analysis based on benchmark
Annotate actions

Fig. 4: Flow of classification of Dan opera actions

3 Analysis of Experimental Results

As shown in Table 1, in order to determine the influence of the number of heads on the effect of the model in multi-head attention, this experiment used ST-GCN as the baseline to compare and verify the best number of heads. This group of experiments was carried out on an NTU-RGB+D data set, using two different classification standards of X-Sub and X-View. X-Sub indicates that the training set and test set are from different people, and X-View indicates that the training set and test set are from different camera perspectives. K is used to represent the number of heads in multi-head attention. The minimum value of K is 1 and the maximum value is 8. This group of experiments only verified the effectiveness of ST-GCN.

Table 1. Header validation of st-gcn

Method	X-Sub(%)	X-View(%)
ST-GCN(K=1)	81.6	88.5
ST-GCN(K=2)	80.3	87.7
ST-GCN(K=3)	81.2	88.2
ST-GCN(K=4)	82.4	89.1
ST-GCN(K=5)	83.8	91.5
ST-GCN(K=6)	83.2	91.1
ST-GCN(K=7)	82.6	90.5
ST-GCN(K=8)	82.1	90.2

Table 2. Openpose validity verification

Method	X-Sub (%)	X-View (%)
ST-GCN(K=5)	83.8	91.5
ST-GCN(K=5)+OpenPose	84.5	92.0

According to the experimental results in Table 2, the accuracy of X-Sub and X-View is improved by 0.7% and 0.5% respectively after adding the OpenPose module on the basis of the ST-GCN model, which shows that the OpenPose module is effective.

4 Conclusion

As an important traditional art component of the Chinese nation, the Dan opera is an important cultural and intangible asset of Chinese civilization, [14]. However, in the modern social environment, Dan opera has gradually lost people's love and attention, and the inheritance is facing challenges. This paper explores how to inject new vitality into the development of Dan opera from a technological perspective. It is found that the integration of modern information technology, such as virtual reality and other advanced means, with the performing art of traditional Dan opera can achieve diversified artistic presentations and revitalize the Dan opera, a long-standing opera culture. The action modeling of Dan opera performance is standardized, which brings considerable challenges to computer recognition technology. Aiming at Jiangsu Dan opera, a typical traditional Chinese opera, this paper proposes a character action recognition algorithm ST-GCN+OpenPose based on deep learning. The algorithm makes full use of the advantages of deep learning in visual feature learning and temporal modeling, and can accurately capture the complex visual features of opera actions and achieve high recognition accuracy on large-scale datasets. This provides strong technical support for automated analysis and understanding of opera performances.

It is conducive to the inheritance and protection of Jiangsu Danju, an intangible cultural heritage. By accurately identifying and recording the movements of the characters in the Danju opera, the classic performance movements can be digitally saved, providing rich information for future generations to study and study the Danju opera. At the same time, it is also conducive to the dissemination and promotion of Dan opera, so that more people can understand and understand this local opera, and stimulate people's interest and love for traditional culture.

Considering the problems of costume occlusion, complex movements, and unique style in Danju performance, the depth learning model will be improved in the future. For example, using the method of expanding convolution partition, the joint coordinate vector of the joint sequence in the graph structure is used as the network input, different subsets of the convolution integration area are divided, and the convolution operation with different parameters is performed for each subset, so that the feature information of the joints involved or possibly involved in the role action in the drama video can be extracted, and the accuracy of the action recognition of the characters in the Danju opera is improved.

References:

[1] Fan T, Wang H, Hodel T, Multimodal knowledge graph construction of Chinese traditional operas and sentiment and genre recognition, *Journal of Cultural Heritage*, Vol.62, 2023, pp. 32-44. DOI: 10.1016/j.culher.2023.05.003.

[2] Tang X, Creation of Drama Art Based on Deep Learning and Evolution Strategy, *Scientific programming*, 2022, Vol. 2022, pp. 6217325.1-6217325.9. DOI: 10.1155/2022/6217325.

[3] Fang M, Peng S, Zhao Y, Yuan H, Hung CC, Liu S, 3s-STNet: three-stream spatial–temporal network with appearance and skeleton information learning for action recognition, *Neural Computing and Applications*, Vol.35, No.2, 2023, pp. 1835-1848. DOI: 10.1007/s00521-022-07763-8.

[4] Wu Q, Huang Q, Li X. Multimodal human action recognition based on spatio-temporal action representation recognition model. *Multimedia Tools & Applications,* Vol.82, No.11, pp. 16409-16430, 2023. DOI: 10.1007/s11042-022-14193-0.

[5] Cheng B, Xiao B, Wang J, Shi H, Huang TS, Zhang L, Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA*, 2020, pp. 5386-5395. DOI: 10.48550/arXiv.1908.10357.

[6] Wang L, The Application of Dance Techniques in Dance Performance, International Journal of Social Science and Education, Vol. 5, No.12,2022, pp. 153-158. DOI: 10.6918/IJOSSER.202212_5(12).0025

[7] Chen J, Chen L, Movement Evaluation Algorithm-Based Form Tracking Technology and Optimal Control of Limbs for Dancers, *Mathematical Problems in Engineering*, Vol. 2022, 2022, pp. 7749324. DOI: 10.1155/2022/7749324.

[8] Richardson E, Alaluf Y, Patashnik O, Nitzan Y, Azar Y, Shapiro S, Cohen-Or D, Encoding in style: a stylegan encoder for image to-image translation.Computer Vision and Pattern Recognition, *IEEE, Nashbille, TN, USA*, 2021, pp. 2287-2296. DOI: 10.1109/CVPR46437.2021.00232.

[9] Alsawadi MS, El-Kenawy ESM, Rio M, Using BlazePose on Spatial Temporal Graph Convolutional Networks for Action Recognition, *Computers, materials & continua*, Vol. 74, No. 1, 2022, pp. 19-36. DOI: 10.32604/cmc.2023.032499.

[10] Jiang H, Tsai SB, An Empirical Study on Sports Combination Training Action Recognition Based on SMO Algorithm Optimization Model and Artificial Intelligence, *Mathematical Problems in Engineering*, Vol. 2021, 2021, pp. 1-11. DOI: 10.1155/2021/7217383.

[11] Liu Y, Ma R, Li H, Wang C, Tao Y, RGB-D Human Action Recognition of Deep Feature Enhancement and Fusion Using Two-Stream ConvNet, *Journal of Sensors*, 2021, Vol. 2021, pp. 8864870. DOI: 10.1155/2021/8864870.

[12] Zahra SB, Khan MA, Abbas S, Khan KM, Ghamdi MAA, Almotiri SH, Marker-Based and Marker-Less Motion Capturing Video Data: Person and Activity Identification Comparison Based on Machine Learning Approaches, *Computers, Materials & Continua,* Vol.66, No.2, 2023, pp. 1269-1282. DOI: 10.32604/CMC.2020.012778.

[13] Emanuel AWR, Mudjihartono P, Nugraha JAM, Snapshot-Based Human Action

Recognition using OpenPose and Deep Learning, *IAENG Internaitonal journal of computer science*, Vol.48, No.4 Pt.1, 2021, pp. 862-867, [Online]. https://www.zhangqiaokeyan.com/journal-foreign-detail/0704060238570.html (Accessed Date: October 22, 2024).

[14] Kraehea AM. The Future of Art Curriculum: Imagining and Longing Beyond "the Now." *Art Education*, Vol. 73, No. 3, 2020, pp. 4–5. Doi: 10.1080/00043125.2020.1738843

Contribution of Individual Authors to the Creation of a Scientific Article (Ghostwriting Policy)

The authors equally contributed in the present research, at all stages from the formulation of the problem to the final findings and solution.

Sources of Funding for Research Presented in a Scientific Article or Scientific Article Itself

No funding was received for conducting this study.

Conflict of Interest

The authors have no conflicts of interest to declare.

Creative Commons Attribution License 4.0 (Attribution 4.0 International, CC BY 4.0)

APPENDIX

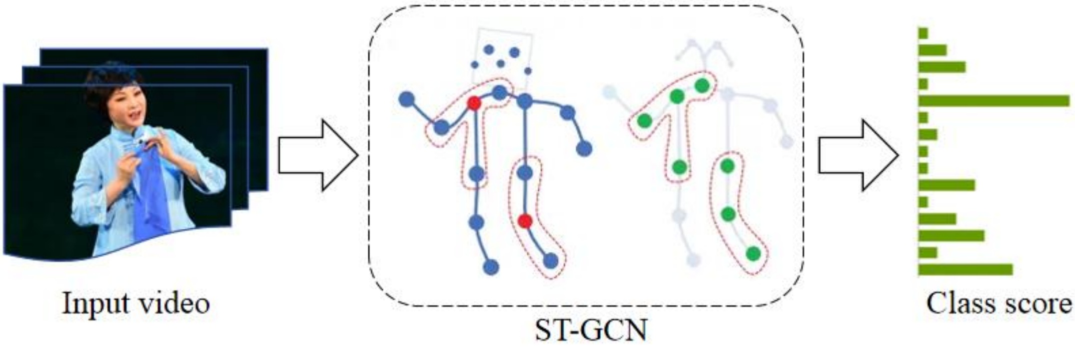

Fig. 1: ST-GCN Process

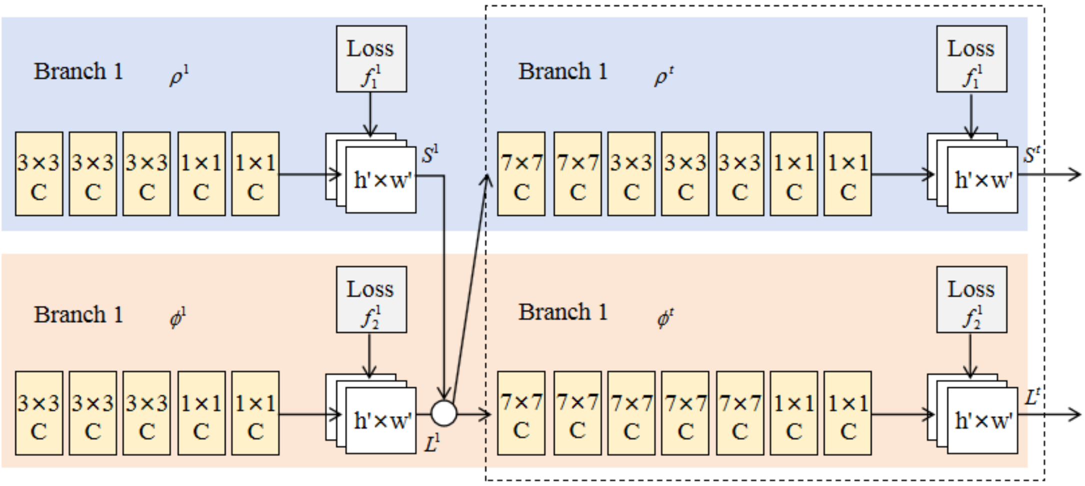

Fig. 3: OpenPose network structure