J. COMBIN. MATH. COMBIN. COMPUT. 127a (2025) 5069--5083

Journal of Combinatorial Mathematics and Combinatorial Computing
www.combinatorialpress.com/jcmcc

Combinatorial Press

Deep Learning-based Modeling of Musical Instrument Performance Movements and Human Biomechanics

Honghe Li[1], Guopeng You[2,] ✉

[1] *Humanities and Arts Media Department, Changzhi, Shanxi, 047100, China*

[2] *Department of Physical Education, Xiamen University of Technology, Xiamen, Fujian, 361000, China*

ABSTRACT

The aim of this study is to construct a deep learning-based biomechanical model of musical instrument playing action that integrates skeletal pose estimation and action recognition techniques. PHRNet-based human pose estimation can extract the skeletal key points of a player from video data, and these key points provide basic data for instrumental performance action recognition and analysis. The human skeletal action recognition method based on diversity rewarded reinforcement learning framework (DDRL-GCN) classifies the extracted key point sequences into specific playing actions, and the musical instrument playing actions are successfully modeled. The biomechanical model of musical instrument playing action designed in this paper is applied to recognize the playing action of five different musical instruments, and the recognition accuracy can reach more than 90%. This paper is designed to distinguish between different musical instruments, the recognition effect is more satisfactory.

Keywords: deep learning, gesture estimation, action recognition, musical instrument playing

1. Introduction

The use of body movement in instrumental performance and the role it plays can be explored from two dimensions. In the utilization dimension, it refers to the method of expressing music through body movements along with the music rhythm, feeling the sentiment contained in the music, and body movements driven by the music rhythm [1-3]. In the role dimension, it is the counteraction and influence of body movements on the quality and state of instrument playing. Performers have a full understanding of both are more likely to show the beauty [4-5].

"Beauty is an apt harmony and moderation", a famous quote by the French philosopher Descartes, could not be more appropriately applied to the body language of instrumental performance [6-7]. The use of messy, contrived, and redundant body language in inappropriate places will inevitably add

✉ Corresponding author.
E-mail address: lihonghe@czmc.edu.cn (G. You).
Received 05 February 2024; Revised 12 April 2024; Accepted 05 December 2024; Published Online 15 April 2025.
DOI: 10.61091/jcmcc127a-286

to the use of body language, which not only destroys the value of the work of art, destroys the image of the work of art, lowers the level of the work of art, hinders the dissemination of the work of art, but also distorts the meaning, concept, and form of beauty, which is ugly [8-11]. Therefore, the beauty of body movements lies in the appropriateness and inappropriateness, coordination and incoherence, which is the aesthetic law of body language in instrumental performance [12-13].

Cavdir, D et al. used quantitative and qualitative methods to analyze the musical gestural interactions involved in Bodyharp to inform innovations in music performance body movements [14]. Wang, G et al. explored the design and performance of digital musical instruments with body movements (gestures) as the core logic, enabling convergent practices in musical and non-musical domains [15]. Blanco-Piñeiro, P et al. discussed the requirements of music for musicians' playing postures and the physical damage these postures can cause to musicians, with bagpipes, percussion instruments, and other instruments being played in the best postures [16]. Erdem, C et al. empirically investigated the interaction of sound and movement in guitar performance and found that the dynamics of vocal movement characteristics can be used to predict the musician's audio energy signature [17]. Ohlendorf, D et al. conducted a survey related to music performance posture and learned that static strain and lameness in the backs of string players during performance that violates body mechanics can cause musculoskeletal disorders in the players [18]. Kim, J. H elaborated that the interaction between instrument and movement during instrumental performance somewhat contributes to the performer's perception of the depth of musical quality and musical experience, as well as being a way of materializing music concretely [19].

The study firstly designed PHRNet-based human posture estimation. The model selects HRNet high-resolution network as the human posture estimation network model, and the PSA polarized attention mechanism is used to improve the structure of this network. Then a diversity reward based human skeletal action recognition method is designed. The method models the diversity frame selection process of video as a discrete Markov decision process. The inter-frame similarity of the video frames is used as a reward to select key frames, and a variant of the cosine distance is used as the inter-frame similarity measure of the video frames. Finally, the gesture estimation model and action recognition model are combined to classify and recognize musical instrument playing actions.

2. Method

Biomechanical modeling of musical instrument performance refers to capturing and analyzing the movement characteristics of musical instrument performance by technical means, in order to achieve accurate modeling and understanding of the performance movement. In this paper, we will design a biomechanical model of musical instrument performance based on “Skeletal Posture Estimation + Action Recognition”.

2.1.Skeletal posture estimation

Skeletal pose estimation is the basis for instrumental performance action modeling. Deep learning algorithms can extract the key points of the player's skeleton from the video data, which include the position of the joints and the trajectory of the movement, to provide basic data for subsequent action recognition and analysis.

2.1.1. HRNet attitude estimation modeling. The task of human posture estimation involves in-depth understanding of human movements, behaviors, and interactions, and its goal is to detect and identify various parts of the human body, such as the head, shoulders, and knees, in an image or video, so as to obtain a skeletal representation of the human body, and then analyze the human body's posture and movements.The HRNet network model adopts a multi-branching structure, where high-pixel and low-pixel feature branches are arranged in parallel, and the mutual integration of the features of different The mutual fusion of different resolution features improves the feature

extraction ability of the network, which enables the network to acquire different levels of feature information and effectively supplements the information loss caused by the reduction of branch resolution, which makes it show excellent ability in semantic segmentation and human posture estimation. The key point heat map is obtained through the four stages of feature extraction operation of the neural network, and then the position information of the skeletal points in the image is determined by the heat value, and the position information of the skeletal points is connected according to the a priori knowledge to obtain the human skeleton structure map, which accomplishes the task of estimating the human body's posture on the image.

The human pose is constructed using the coordinate information of key points of human skeleton including head, shoulder, elbow, wrist, ankle, etc., and the MPII dataset is used for validation, in which the head is regarded as a key point.The structure of the human skeleton in the MPII dataset is shown in Fig. 1.

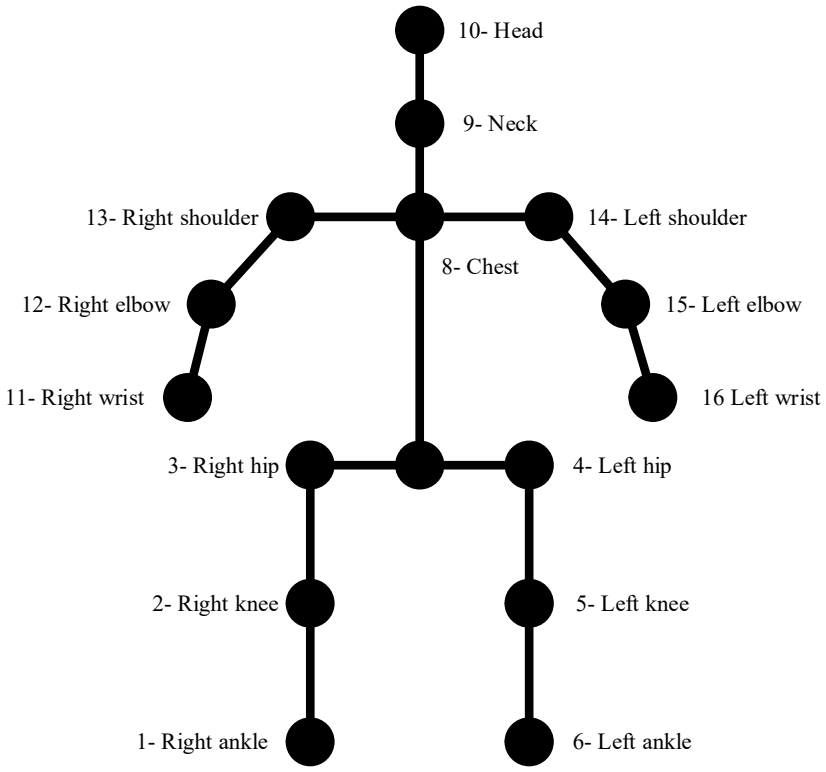


Fig. 1. MPII human skeleton

2.1.2. PHRNet incorporating PSA-polarized self-attention. Attention mechanism can effectively make the network pay more attention to the effective features, there are two types of attention mechanism, channel attention and spatial attention. Polarized self-attention can maintain a relatively high resolution in the channel and spatial dimensions, which further can reduce the loss of information caused by dimensionality reduction, improve the feature extraction ability of the network model, and thus improve the accuracy of human pose estimation [20].

1) PSA polarized self-attention module. The attention mechanism is initially a single-branch structure, such as SE, ECANett and other channel attention mechanisms. This single-branch attention mechanism only applies different weights in the channel, but applies the same weights to different spatial locations, which results in the spatial location features not being better utilized, and thus it is often difficult for separate spatial or channel attention to comprehensively capture the key information in the image. In contrast, dual-attention mechanisms [21] such as DANet and CBAM are able to adjust the weights of features in both spatial and channel dimensions in a targeted manner respectively, so that the model can further explore the deeper high-level features embedded on the

image.

When modeling both spatial and channel, each pixel or feature point needs to be computed in both spatial and channel dimensions to generate the corresponding weight map, and the amount of computation and memory consumption will increase dramatically, which will increase exponentially with the increase in the size of the input image or feature map. In order to make the information transfer more concise and efficient, the PSA polarized self-attention module employs a polarization filtering method to selectively amplify or filter out features. Next, the PSA module uses polarization filtering to completely compress the features in a particular dimension. For example, when the PSA module performs full compression of the channel dimension, the high resolution of the spatial dimension is maintained so that the spatial feature information can be fully extracted and utilized. This polarization filtering approach helps the model to focus on key features in a certain dimension, improving the relevance and effectiveness of feature extraction. In order to recover the compressed features and restore their original information content, the PSA module also adopts a high dynamic range mapping technique by applying the Softmax function to expand the minimum tensor in the attention module, which can effectively expand the range of attention by calculating the exponential value of each element and normalizing it, and then performs a dynamic mapping operation by using the Sigmoid function, which is a dynamic mapping operation of the features that have gone through the Softmax processed feature values are mapped to the range of [0,1] to recover the original information content of the compressed features. The PSA module is specified in two structures: series and parallel.

For channel attention $A^{ch}(X)$ and spatial attention $A^{r}(X)$, the formulas are, respectively:

$$A^{ch}(X) = F_{SG}[W_z|_{SM}(\sigma_1(W_v(X)) \times F_{SM}(\sigma_2(W_q(X)))] \tag{1}$$

$$A^{\varphi}(X) = F_{SG}[\sigma_3(F_{SM}(\sigma_1(F_{GP}(W_q(X))))) \times \sigma_2(W_v(X))] \tag{2}$$

Eqs:

W_q, W_v, W_z - 1×1 convolutional layers.

$F_{SM}(\cdot)$ - SoftMax operator.

$\times$ - Dot product operation for matrices.

F_{GP} - Global pooling operation.

F_{SG} - The sigmoid function.

$\sigma_1, \sigma_2, \sigma_3$ - Tensor reshaping operator.

The channel attention and spatial attention are combined by either parallel or tandem, and the combination method for parallel (PSA_p) and tandem $(PSAs)$ is expressed by Eq:

$$PSA_p(X) = Z^{sp} + Z^{sh} = A^{ch}(X) \square^{ch} X + A^{sp}(X) \square^{sp} X \tag{3}$$

$$PSA_z(X) = Z^{sp}(Z^{ch}) = A^{sp}(A^{ch}(X) \square^{ch} X) \square^{sp} A^{ch}(X) \square^{ch} X \tag{4}$$

Eq:

Z^{sp} - Output of the spatial attention module.

Z^{ch} - Output of the channel attention module.

$\square^{ch}$ - Channel multiplication operator.

$\square^{p}$ - Spatial multiplication operator.

The PSA module uses the Self-Attention mechanism to obtain the attention weights.The Self-Attention mechanism is an effective feature processing method with global modeling capability that enables long range modeling of features. This means that the Self-Attention mechanism is able to capture the connections between pixel points farther away in the image, rather than just information about localized regions. By calculating the correlation between each location in the

feature map and all other locations, the Self-Attention mechanism is able to generate a weight vector for each location that reflects the feature associations between that location and other locations. The tandem PSAs architecture effectively reduces computational complexity and graphics memory consumption, and also maintains the integrity of feature information and improves the model's sensitivity to key information in the image. This design enables the PSA module to capture and recognize the key information in the image more accurately, thus improving the overall performance of the model.

2) Basicblock module with fused PSA. The PSA module can improve the model's sensitivity to key information in the image while maintaining the integrity of feature information. Therefore, the PSA module is introduced into the Basicblock module to construct the PBasicblock (PSA-Basicblock) module, and the flow of the PBasicblock module is shown in Figure 2. All the Basicblock modules of the four stages in the model are replaced with the modified PBasicblock module to improve the feature extraction ability of the model for the human target bone points, to obtain richer feature expressions and to realize more accurate human pose estimation.

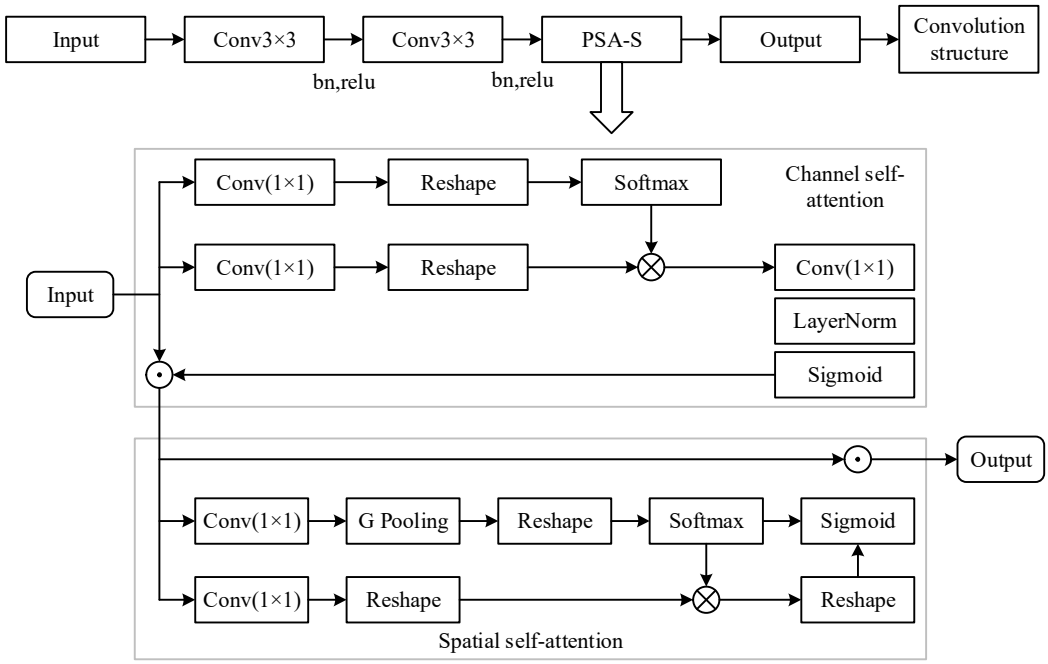


Fig. 2. PBasicblock structure

2.2. Skeletal action recognition

After obtaining the skeletal pose data, the next step is to perform skeletal action recognition. The purpose of action recognition is to classify the extracted sequence of keypoints into specific playing actions. In this section, a framework of diversity reward-based deep reinforcement learning algorithm (DDRL-GCN) for human skeletal action recognition is proposed. The general framework of the algorithm is first described, after which a variant of the cosine distance is defined as the diversity reward for video frames. Finally, the skeletal keyframe selection process is then problem defined and modeled.

2.2.1. General DDRL-GCN framework. The overall framework of DDRL-GCN is shown in Fig. 3.The DDRL-GCN framework consists of two parts: a diversity reward-based frame selection network (DDRL) and a spatio-temporal feature model selector, GCN-Selector, which contains the 2s-AGCN and MS-G3D models.The DDRL-GCN is firstly used by the DDRL network to select diversity frame index sequences based on the diversity reward $Rdiv$. Sequence, after which the dataset is directly feature

extracted based on the diversity frame index sequence, and finally recognized with accuracy via GCN-Selector.

Among them, the GCN-Selector directly adopts the existing powerful spatio-temporal feature extractors 2s-AGCN and MSG3D. The input of the framework is the tandem combination of the original skeletal sequences F and the selected skeletal sequences Fm in the temporal dimension, and finally, the GCN-Selector performs the optimal output based on the accuracy of the action recognition computed by the two spatio-temporal feature extractors mentioned above.

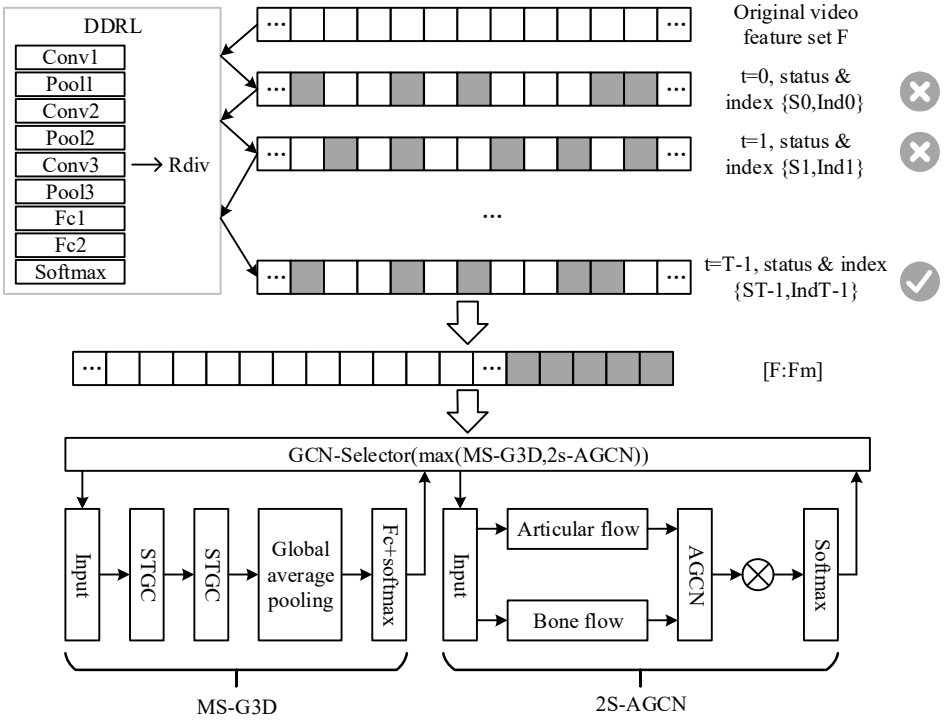


Fig. 3. DDRL-GCN network framework

2.2.2. Cosine similarity. Cosine similarity is essentially a measure that takes the cosine value of the angle of the corresponding vectors between any two samples in the feature space as a measure of their similarity [22], and in this paper, it is introduced into the human skeleton action recognition problem as a measure of the difference between any two video frames. The value range of the cosine function is [-1, 1], if the cosine value is closer to 1, i.e., the angle between the two eigenvectors is closer to 0 degrees, at this time, these two vectors in the direction of the degree of similarity is higher.

For two vectors A and B in space, the cosine similarity formula is as follows:

$$cos(A,B) = \frac{A \cdot B}{\| A \|_2 \| B \|_2} \tag{5}$$

One of the characteristics of cosine similarity is that it focuses on the difference in direction between any two vectors in the feature space and is not sensitive to the length of the distance between the two vectors. If the angles between two vectors are equal but the numerical differences are very large, the cosine similarity can bias the experimental results due to the lack of a measure of the numerical differences in each dimension.

2.2.3. Diversity incentives. Assuming that each video frame is a distance unit in the time dimension, the index number of a video frame is its absolute distance in the time dimension. As for the diversity of video frames, the most different video frames should be used as key frames, so this section needs

to find the cosine distance $div(x_t, x_t^{'})$ between video frames based on the cosine similarity, which is a function used to measure the degree of dissimilarity between the features x_t and $x_t^{'}$ of any two selected frames in the feature space, which is calculated as follows:

$$div(x_t, x_{t'}) = 1 - \frac{x_t^T x_t^{'}}{\| x_t \|_2 \| x_t^{'} \|_2} \tag{6}$$

However, it is still important to note that the cosine similarity is only sensitive to the direction of the inter-frame eigenvectors and ignores the effect of the absolute distance between the vectors. After finding out the dissimilarity between video frames, in order to better measure the similarity in distance between different frame vectors in the same cosine similarity, this paper will use a variant of the cosine similarity, namely, the ratio of the cosine distance to index $Ind_i \in \square^{1\times m}$, as a measure of the diversity between video frames. In this way, in the case of equal cosine angles, if the product of the index numbers of the two selected frames is larger, i.e., the further apart the two frames are in the time dimension, the smaller the inter-frame diversity will be, and this approach is extremely suitable for skeletal video data such as data with multiple repetitions of frames. With equal cosine angles, it is easier to select frames that are relatively close in the time dimension, preserving the continuity of the video frames as much as possible while reducing the degree of redundancy in the selected frames. Assuming that the selected frames index sequence $Ind_m = \{Ind_i \mid A_{Ind_i} = 1, i = 1, 2, ..., m\}$, the cosine distance variant D_{div} is calculated as follows:

$$D_{div} = \frac{1}{| Ind_i | (| Ind_i | - 1)} \left(\sum_{t \in Ind_i} \sum_{t' \in Ind_i} div(x_t, x_t^{'}) \right) \tag{7}$$

The diversity bonus R_{div} is measured by the cosine distance variant between video frames in the feature space. The smaller the cosine angle between the two selected frames, the greater the cosine similarity, the smaller the cosine distance, and therefore the smaller the cosine distance variant, i.e., the similarity between the two cries is inversely proportional to the cosine distance variant. This is exactly in line with the diversity reward needs to capture the distribution of scattered video frames and low similarity "mutation" frame needs, so the cosine distance variant is selected as the diversity reward, the intelligent body in the iterative process to select the high diversity reward value of the frame. The diversity reward R_{div} is calculated as follows:

$$R_{div} = D_{div} = \frac{1}{| Ind_i | (| Ind_i | - 1)} \left(\sum_{t \in Ind_i} \sum_{t' \in Ind_i} div(x_t, x_t^{'}) \right) \tag{8}$$

2.2.4. Problem modeling. The process of extracting video key frames by RL technique can be regarded as an MDP process. Each iteration selects a fixed number (m) of key frames to form a frame index sequence Ind_m, and each frame of the selected frame index sequence Ind_m is equivalent to a sliding window, and the intelligent body will generate an action sequence A_t according to the policy to guide the actions of the sliding window ("move left", "keep still", "move right") to guide it so as to generate a new frame index sequence $Ind_m^{'}$. At the t moment, traversing the skeletal samples V_τ in the sample space V, the important components of the MDP are as follows.

State: the state (S_t) consists of the features $F_m \in \square^{m\times j\times c}$ of the selected skeletal sequence and the selected index sequence Ind_m, i.e., $S_t = [F_m, Ind_m]$, in the initial state S_0, F_m, Ind_m are uniformly sampled from the original feature set V_n of the current skeletal sample n, and uniformly sampled from the index sequence from 0 to 299, respectively. where m is the number of frames in the selected frame, j is the number of joints in the human skeleton map, c is the number of axes $(c = 3)$. F_m provides the accuracy of the frame, and Ind_m provides part of the visualization

content.

Action: Action (A_t) $A_t \in \square^{m\times 1}$ is the sequence of action selection for each sampled frame, and three types of actions are defined in this paper: "Left Shift" $(j=0)$, "Keep" $(j=1)$, and "Right Shift" $(j=2)$. In the iteration In the iteration process, the action set of the current frame i selecting action j is denoted as $A_i, j \in \square^{m\times 3}$, and its corresponding action selection probability $Pob_{i,j} \in [0,1]$, A_t is the sequence of actions selected by $A_{i,j}$ according to the highest probability in $Pob_{i,j}$, and the initial action sequence A_0 is randomly selected from the set of actions with equal probability.

State Transfer Function The state model for each action at moment $P_a^t : t$, i.e. $P(S_t+1=S^{'} \mid S_t=s, A_t=a)$. Reward: the reward (R_t) is the feedback from the intelligence about how good or bad the behavior A_{t-1} taken in the previous round of state S_{t-1} was.

Discount factor: discount factor $\gamma \in [0,1]$.

3. Results and discussion

3.1. Experimental data set

This experiment uses two datasets, UCF101 and Kinetics, to evaluate the performance of the method proposed in this paper.

The UCF101 dataset is a public dataset in the field of human behavior recognition. This dataset is mainly obtained from video websites, and is divided into 101 action categories, including the category of "playing a musical instrument". The UCF101 dataset specifies that each category of behavior consists of 25 instances, and each instance has 4~7 groups of video data, so there are a total of 13,320 videos as human behavior instances. The UCF101 dataset has a great diversity in terms of human appearance, video background, action amplitude, and expression because the videos are randomly shot in real environments without any limitation or pre-rehearsal. There are also large variations in camera type, video lighting, and video resolution.

The Kinectis dataset is a large-scale, high-quality dataset focused on video behavior recognition. This dataset contains 400 different human action categories, each with at least 400 video clips. Each clip lasts approximately 10 seconds and are all from different YouTube videos. The data covers a wide range of scenarios from human-object interactions to human-human interactions. The action category "playing a musical instrument" is also included.

3.2. Experimental Analysis of Skeletal Posture Estimation

3.2.1. Evaluation indicators. In this experiment, the percentage of correct joint point localization (PCK), the percentage of correct site (PCP), and the number of parameters were used as evaluation indexes.

PCK: It mainly measures whether the Euclidean distance between the predicted joint point and the real joint point is within a specific threshold. The calculation formula is shown in equation (9):

$$PCK_i^k = \frac{\sum_P \delta\left(\frac{d_{pi}}{d_p^{def}} \le T_k\right)}{\sum_p 1} \tag{9}$$

where $\delta(*)$ denotes the determination of condition *, $\delta(*)=1$ when condition * holds, and $\delta(*)=0$ otherwise. i denotes the ith joint. T_k denotes the kth threshold. p denotes the pth body. d_{pi} denotes the Euclidean distance between the predicted and manually labeled values of the

ith joint point in the pth person. d_p^{def} denotes the scale factor of the pth person. PCK_i^k denotes the PCK metric of the ith joint point at the T_k th threshold.

3.2.2. Analysis of results. The PHRNet method was compared with other attitude estimation methods on two datasets, UCF101 and Kinetics. Parametric quantities and PCK are used to evaluate the models respectively.

The experimental results of the comparison between the PHRNet method and other methods on the UCF101 dataset are shown in Table 1. In terms of model parameter count metrics, the proposed method reduces 2.7M and 3.3M compared to the TokenPose model and SH-Net model, respectively.In terms of recognition accuracy metrics, the proposed method improves 9.84% and 4.83% compared to the SE-Net model and ViTPose model, respectively. From the comparison results of prediction accuracy of joint points in several key parts of the human body, it can be seen that the prediction accuracy of the knee part is slightly lower compared to the SH-Net model, but the method in this paper achieves the highest value in the other six parts.

Table 1. Comparison of PHRNet and other methods on Datasets of UCF101

Method	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Parameter quantity	PCK
HRNet	74.29	48.84	40.71	34.18	36.56	34.35	35.24	8.1	44.12
SE-Net	95.88	91.74	81.71	72.76	82.8	73.34	66.5	7.3	81.32
DEKE	95.07	91.76	83.06	76.76	81.86	74.67	69.39	7.9	82.53
TokenPose	96.23	93.16	86.85	82.51	85.24	81.37	74.12	6.6	86.17
ViTPose	97.32	94.17	86.37	81.31	86.79	80.65	73.44	6.8	86.33
OpenPose	97.79	95.19	88.72	84.1	88.52	82.76	79.44	7	88.58
SH-Net	98.26	96.2	91.16	87.05	90.04	87.48	83.61	7.2	91.01
PHRNet	98.19	96.34	91.36	87.45	91.07	87.19	83.7	3.9	91.16

The experimental results of the comparison between the PHRNet method and other methods on the Kinetics dataset are shown in Table 2. In terms of the number of model parameters, the proposed method reduces 4.9M and 3.9M compared with DEKE model and SH-Net model, respectively.In terms of the recognition accuracy, the proposed method improves 5.27% and 8.13% compared with TokenPose model and SE-Net model, respectively.

Table 2. Comparison of PHRNet and other methods on Datasets of Kinetics

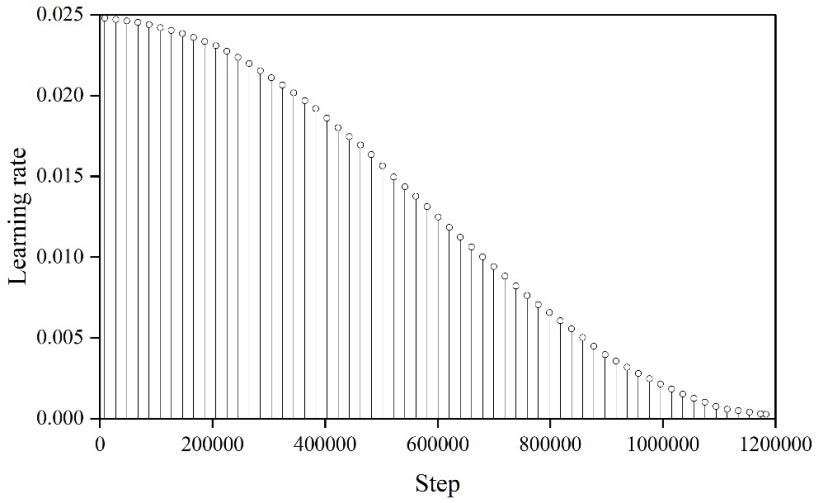
Method	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Parameter quantity	PCK
HRNet	75.32	49.97	44.25	41.23	37.55	39.51	39.83	7.9	46.76
SE-Net	95.75	93.65	85.71	79.18	84.55	75.26	71.28	7.3	84.39
DEKE	96.3	93.07	85.2	78.2	83.18	76.31	70.44	8	86.58
TokenPose	97.37	94.91	88.63	84.75	87.72	83.38	76.25	6.8	87.25
ViTPose	97.59	94	88.11	84.23	87.99	82.2	75.38	6.6	88.11
OpenPose	97.88	96.29	90.19	86.48	89.66	84.11	82.7	7.2	89.11
SH-Net	98.29	96.47	92.33	88.98	93.72	88.37	86.31	7	91.49
PHRNet	98.76	97.4	93.58	89.94	91.55	89.98	86.16	3.1	92.52

It can be seen that the PHRNet model proposed in this paper shows superior performance in the human pose estimation task, maintaining high recognition accuracy and site prediction accuracy while reducing model complexity. In addition, although the PHRNet method is not optimal for some specific site predictions, it performs very well overall and achieves the highest prediction accuracy

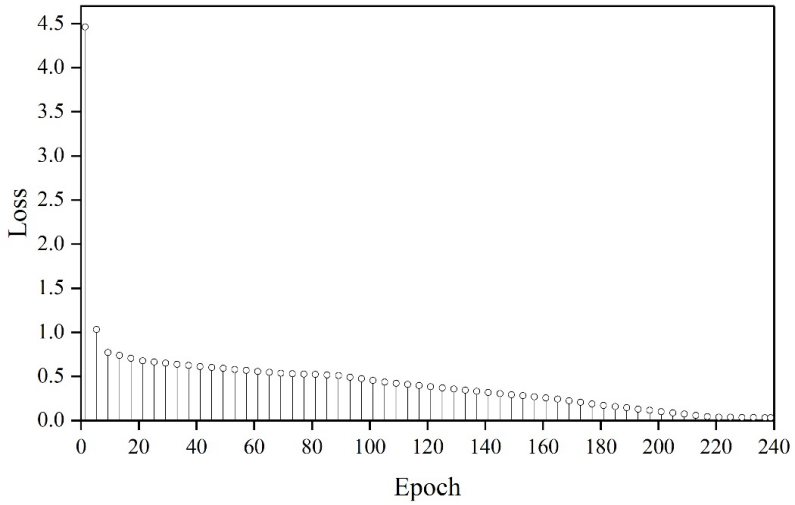
for most of the critical sites. This experiment fully demonstrates that the method in this paper is an efficient and robust human posture estimation method with good generalization ability and potential for practical applications.

3.3. Experimental Analysis of Skeletal Action Recognition

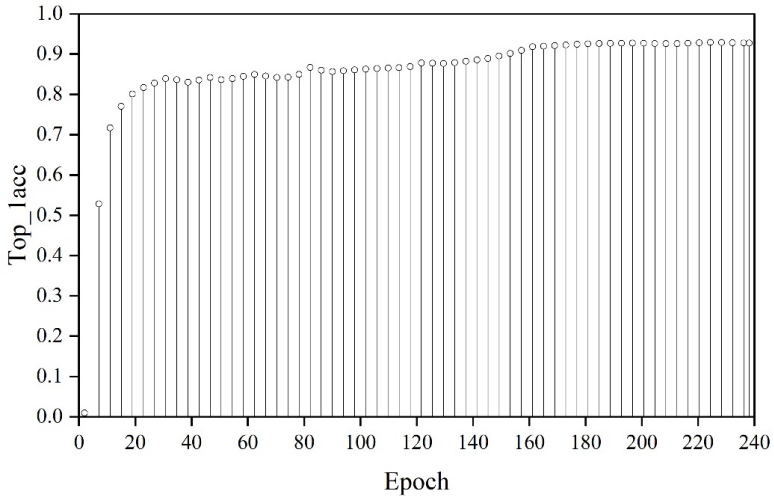
Based on the key point data of musical instrument playing skeleton extracted by PHRNet model for action recognition, the model training results on UCF101 dataset are shown in Fig. 4. (a)~(c) represent the learning rate curve, the loss value curve, and the recognition accuracy during the recognition process, respectively. The loss value is higher at the beginning of the learning process, but it decreases rapidly, and finally tends to converge, and the learning rate also tends to a steady state, and continues to search for the optimal, and basically converges to the lowest at about 90% of the training process, and tends to a steady state. Validation is performed every 5 rounds during the training process. At the beginning of training, the classification accuracy value grows rapidly, there is a little twist in the middle stage, but the overall trend is upward, and at about 180 rounds, the model accuracy is close to steady, and peaks at about 190 rounds.



(a) learning rate



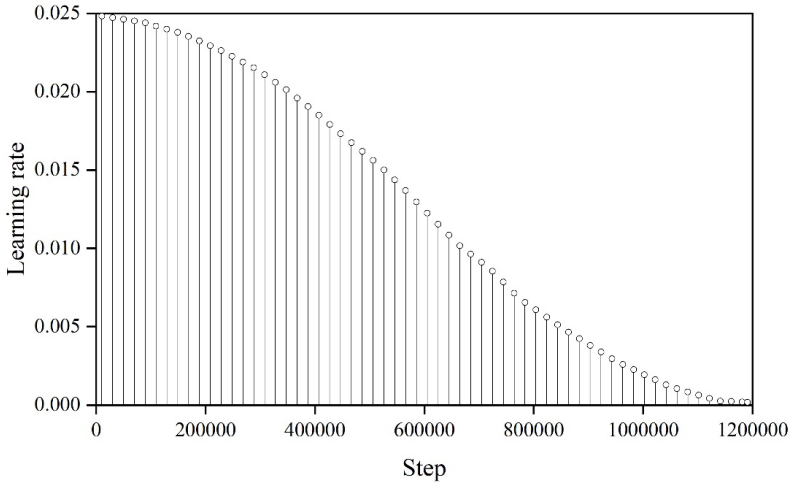
(b) Loss value



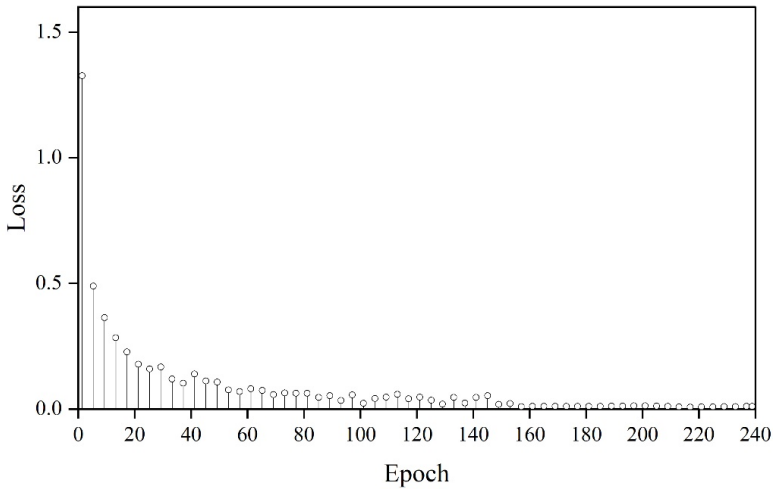
(c) Top 1 precision

Fig. 4. Model training results in UCF101

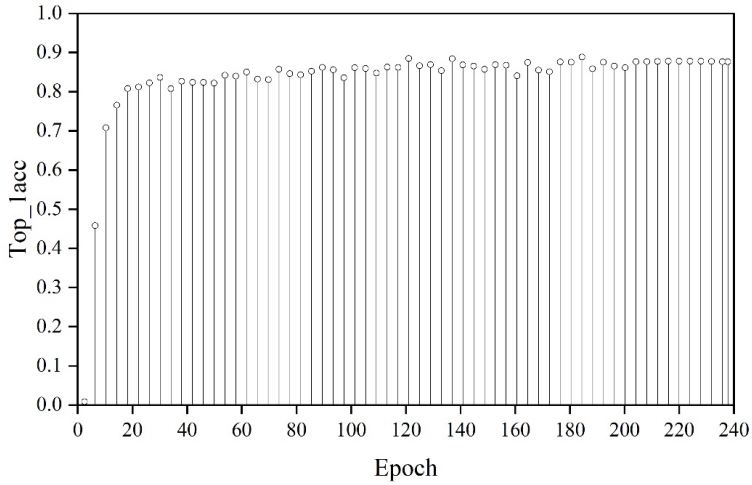
The results of model training on the Kinetics dataset are shown in Fig. 5. (a)~(c) represent the learning rate curve, the loss value curve, and the recognition accuracy during the recognition process, respectively. The pre-training model is used in the training and learning process, so the loss value during training decreases sharply in the first 20 epochs of training, and slowly levels off, and the leveling off part also directly corresponds to the smooth optimization-seeking stage at the end of the learning rate curve, which can be obtained from Fig. (c), where the Top 1 (meaning that the predicted category with the highest confidence level in the categorization and recognition of the validation samples is its true category, usually expressed as a percentage) In the learning to about 160 rounds, along with the loss value infinitely close to 0, Top 1 accuracy also basically smooth to reach the peak value, no longer rise, and finally Top 1 reaches 90.97% accuracy value.



(a) learning rate



(b) Loss value



(c) Top 1 precision

Fig. 5. Model training results in Kinetics

3.4. Performance Action Recognition Accuracy Analysis

The accuracy of the PHRNet-based skeletal pose estimation method and the DDRL-GCN-based action recognition technique method was verified above using two datasets, UCF101 and Kinetics. In this section, the two models are combined to recognize the playing actions of five different musical instruments: piano, violin, accordion, cello and guzheng, and the accuracy of action recognition is shown in Figure 6. The biomechanical models constructed in this paper all have more than 90% accuracy in recognizing the instrumental playing actions, and are able to accurately distinguish the playing actions of different instruments.

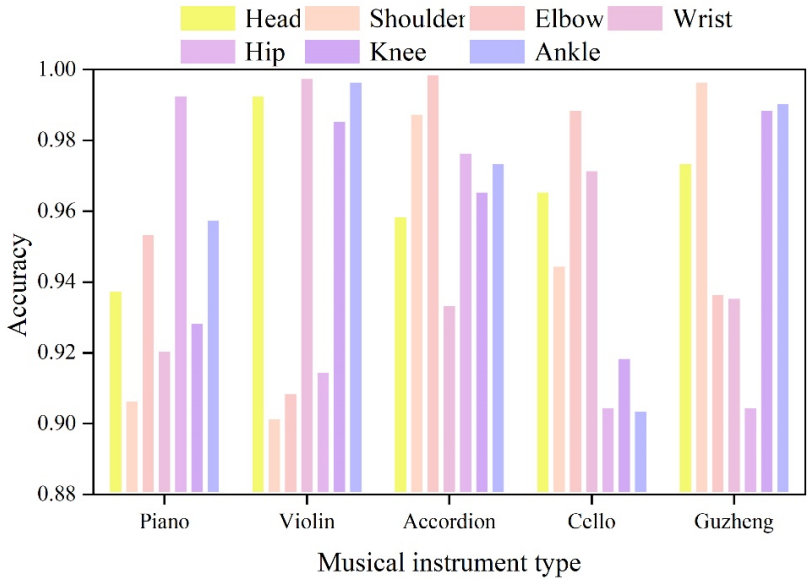


Fig. 6. Action recognition accuracy

4. Conclusion

The biomechanical model of musical instrument playing action constructed in this paper combines PHRNet-based human pose estimation and DDRL-GCN-based action recognition. The PHRNet-based human pose estimation method is compared with other methods on the UCF101 dataset and the Kinetics dataset, respectively, and the recognition accuracies of PHRNet are 91.16% and 92.52%, respectively. The recognition accuracy can reach 90.97% when the motion recognition is based on the key point data of the instrument playing skeleton extracted from the PHRNet model. Applying the biomechanical model designed in this paper to recognize the playing movements of piano, violin, accordion, cello and guzheng, the recognition accuracies of head, shoulder and ankle movements can reach more than 90%. It shows that the method in this paper can accurately classify and recognize the playing movements of different musical instruments.

References

[1] Frizziero, A., Gasparre, G., Corvo, S., Gamberini, J., Finotti, P., Masiero, S., & Maffulli, N. (2018). Posture and scapular dyskinesis in young bowed string instrumental musicians. Muscles, Ligaments & Tendons Journal (MLTJ), 8(4).

[2] Fintoni, L. (2020). Bedroom Beats & B-sides: Instrumental Hip Hop & Electronic Music at the Turn of the Century. Velocity Press.

[3] Visi, F. G., Östersjö, S., Ek, R., & Röijezon, U. (2020). Method development for multimodal data corpus analysis of expressive instrumental music performance. Frontiers in Psychology, 11, 576751.

[4] Tanaka, A., & Donnarumma, M. (2019). The body as musical instrument. The Oxford handbook of music and the body, 79-96.

[5] Wyroślak, P. (2023). Accusative-instrumental alternation in Polish. Zeitschrift für Slawistik, 68(1), 41-82.

[6] Doğantan-Dack, M. (2016). The role of the musical instrument in performance as research: The piano as a research tool. In Artistic practice as research in music: Theory, criticism, practice (pp. 169-202). Routledge.

[7] Nikolić, K. (2023). The Vocal-Instrumental Dance Form Gara: Vital Manifestation of Dance Genre Kolo u

Tri within the Wedding Ritual in Serbia. Conservatorium, 10(1), 14-27.

[8] De Souza, J. (2017). Music at hand: Instruments, bodies, and cognition. Oxford University Press.

[9] Williams, K. E. (2018). Moving to the beat: Using music, rhythm, and movement to enhance self-regulation in early childhood classrooms. International Journal of Early Childhood, 50(1), 85-100.

[10] Rì, F., Masu, R., Graziani, M., & Roncador, M. (2020). From the Body with the Body: Performing with a Genome-Based Musical Instrument. EAI Endorsed Transactions on Creative Technologies, 7(23).

[11] Alla, C. (2019). PERFORMING SEMANTICS OF SILENCE IN INSTRUMENTAL MUSIC OF THE XX-XXI CENTURIES. National Academy of Managerial Staff of Culture & Arts Herald/Vìsnik Deržavnoï Akademìï Kerìvnih Kadrìv Kul'turi ì Mistectv, (1).

[12] Bremmer, M., & Nijs, L. (2020, June). The role of the body in instrumental and vocal music pedagogy: a dynamical systems theory perspective on the music Teacher's bodily engagement in teaching and learning. In Frontiers in Education (Vol. 5, p. 79). Frontiers Media SA.

[13] López-Íñiguez, G., & McPherson, G. E. (2024). Using a music microanalysis protocol to enhance instrumental practice. Frontiers in Psychology, 15, 1368074.

[14] Cavdir, D., & Dahl, S. (2022, June). Performers' Use of Space and Body in Movement Interaction with A Movement-based Digital Musical Instrument. In Proceedings of the 8th International Conference on Movement and Computing (pp. 1-12).

[15] Cavdir, D., & Wang, G. (2021). Borrowed gestures: The body as an extension of the musical instrument. Computer Music Journal, 45(3), 58-80.

[16] Blanco-Piñeiro, P., Díaz-Pereira, M. P., & Martínez Vidal, A. (2018). Variation in posture quality across musical instruments and its impact during performances. International Journal of Occupational Safety and Ergonomics, 24(2), 316-323.

[17] Erdem, C., Lan, Q., & Jensenius, A. R. (2020). Exploring relationships between effort, motion, and sound in new musical instruments. Human Technology, 16(3), 310-347.

[18] Ohlendorf, D., Wanke, E. M., Filmann, N., Groneberg, D. A., & Gerber, A. (2017). Fit to play: posture and seating position analysis with professional musicians-a study protocol. Journal of occupational medicine and toxicology, 12, 1-14.

[19] Kim, J. H. (2020). From the body image to the body schema, from the proximal to the distal: Embodied musical activity toward learning instrumental musical skills. Frontiers in Psychology, 11, 101.

[20] Liu Shengjie,He Ning,Wang Cheng,Yu Haigang & Han Wenjing. (2022). Lightweight human pose estimation algorithm based on polarized self-attention. Multimedia Systems(1),197-210.

[21] Yuxiao Duan,Jiuyang Tang,Hao Xu,Changsen Liu & Weixin Zeng. (2024). Commonsense-Guided Inductive Relation Prediction with Dual Attention Mechanism. Applied Sciences(5).

[22] Hei Hongzhong,Jian Xianzhong & Xiao Erliang. (2021). Sample weights determination based on cosine similarity method as an extension to infrared action recognition. Journal of Intelligent & Fuzzy Systems(3),3919-3930.