Applied Soft Computing Journal 164 (2024) 111963

Contents lists available at ScienceDirect

Applied Soft Computing

journal homepage: www.elsevier.com/locate/asoc

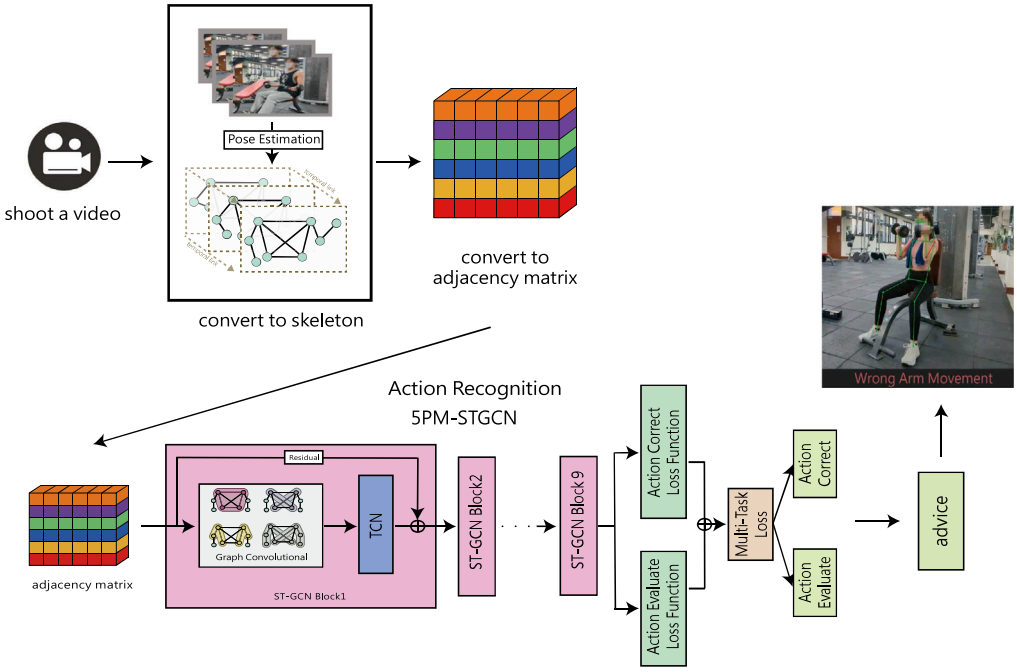
# Human movement science-informed multi-task spatio temporal graph convolutional networks for fitness action recognition and evaluation

Jia-Wei Chang [a], Ming-Hung Chen [b], Hao-Shang Ma [a,*], Hao-Lan Liu [a]

[a] Department of Computer Science and Information Engineering, National Taichung University of Science and Technology, No. 129, Section 3, Sanmin Road, North District, Taichung City, 404, Taiwan

[b] Physical Education Section, National Taichung University of Science and Technology, No. 129, Section 3, Sanmin Road, North District, Taichung City, 404, Taiwan

## GRAPHICAL ABSTRACT




## ARTICLE INFO

*Keywords:*
Action recognition
Five Primary Kinetic Chains(5PKC) partitioning
Multi-task objective loss
Auto-fitness advisor system

## ABSTRACT

In recent years, with the rise of health consciousness, people's demand for fitness has steadily increased. Utilizing automated human action recognition technology to monitor users' movements during exercise continuously would help prevent situations where incorrect movements lead to injuries while working out. Skeleton-based human action recognition methods can overcome the susceptibility of past color-based and depth-based methods to various external backgrounds and noise, becoming a more successful solution in recent years. In this study, the auto-fitness advisor system we propose not only identifies the category of the action but also assesses the quality of the action and provides suggestions. We integrate human movement science,

* Corresponding author.
*E-mail addresses:* jwchang@nutc.edu.tw (J.-W. Chang), allen@nutc.edu.tw (M.-H. Chen), hsma@nutc.edu.tw (H.-S. Ma).

https://doi.org/10.1016/j.asoc.2024.111963
Received 10 October 2023; Received in revised form 25 June 2024; Accepted 29 June 2024
Available online 9 July 2024

such as the Five Primary Kinetic Chains (5PKC), which defines the primary physiological principles in human movement, to enhance the accuracy of fitness action recognition by providing a more precise relationship between the human skeleton and muscles. For assessing the quality of movements and providing suggestions, we have designed a multi-task objective function within our model. Overall, the proposed model is a multi-task model based on Spatio Temporal Graph Convolutional Networks (ST-GCN), which employs the Five Primary Kinetic Chains (5PKC) as a partitioning strategy for skeletal information. In our experiments, we not only collected a certain amount of datasets in gyms to validate the performance of our model but also compared it with other current methods using existing public datasets.

## 1. Introduction

In recent years, there has been a global fitness trend, with fitness centers popping up like mushrooms. People have more interest in exercise and weight training. Improper exercise movements may lead to injuries. Even though there are fitness trainers to assist, there is still a possibility of overlooking certain dangers due to environmental or human factors. For gym owners, enhancing the gym's safety and reducing operational costs are of utmost importance. Human action recognition is the task of recognizing human movements. Applying human recognition technology to the monitor system can let owners continuously monitor and assess users' movements automatically for safety purposes. The system may detect and assess whether the movements meet the standards immediately. If the movements are incorrect, the system can issue an alert, reminding users to make adjustments. This can effectively prevent users from ignoring potential risks due to environmental or human factors, thus reducing the risk of injuries caused by improper exercise movements. Furthermore, automated human action recognition technology may help gym owners reduce labor costs. In general, gyms require hiring many professional fitness trainers to guide users, which should be expensive. The professional sometimes may lose the attention of some users. By utilizing automated human action recognition technology, the demand for specialized trainers can be reduced, thereby lowering labor costs. Additionally, applying the system may provide better guidance to users when they exercise or workout at home, further enhancing fitness effectiveness.

Human action recognition has been widely applied in multimedia computing, including intelligent monitoring, virtual reality, human–computer interaction, and more. Despite significant advancements in recent years, challenges such as the complexity and variability of human action and diverse environmental conditions still leave room for improvement in recognition accuracy. Early existing human action recognition methods are primarily image-based and take factors like backgrounds into account. However, the same action may appear significantly different in various lighting, angles, and backgrounds. This introduces unnecessary noise when performing action recognition. Researchers have proposed using human skeleton information for the human action recognition task to address this issue. Skeleton-based action recognition effectively mitigates background noise and variations in body posture within images. Furthermore, compared to image-based methods, skeletal information is more focused on capturing action information, enhancing the efficiency of human action recognition.

In human action recognition, deep neural networks [1] have become the primary tools. Researchers have recently explored using skeleton-based temporal CNNs or RNNs to perform action recognition, representing human skeletal data as vector sequences for model input. However, due to the non-Euclidean nature of skeletons, directly inputting skeleton coordinates into models cannot effectively analyze the spatial distribution of skeletal joints [2]. To address this issue, researchers have applied graph convolutional networks (GCN) [3] to skeleton-based motion recognition, as GCNs efficiently analyze the structural information within graphs. One prominent example is the spatio-temporal graph convolutional network (ST-GCN), a representative work in skeleton-based motion recognition. ST-GCN performs graph convolutional operations by using joints as vertices and joint connections as edges to construct a skeletal graph and designing spatial configuration partitioning based on the adjacency relationships between joints. This method has proven effective in improving human action recognition. Several variants derived from ST-GCN have achieved remarkable results [4–6], making ST-GCN one of the most commonly used frameworks for this task.

The GCN-based methods usually consider the position and connection information within skeletal structure. The information of human skeletons is represented using adjacency matrices, and convolution operations are utilized to extract skeletal relationships relevant to human body movements. However, in the field of exercise science, the stable and accurate completion of a movement is not solely determined by the action of adjacent muscle groups. Therefore, we have introduced the concept of the Five Primary Kinetic Chains (5PKC) [7] from exercise science. 5PKC is a theory that integrates body structure and kinematic principles, categorizing the body into five kinetic chains: intrinsic core, deep longitudinal, lateral stabilizing, posterior oblique, and anterior oblique chains. 5PKC allows us to understand the interaction relationship between a muscle with the rest of the structure Applying 5PKC theory to partitioning strategies can better reflect the holistic features of movements, enabling a more comprehensive analysis of action characteristics. In addition, most existing skeleton-based action recognition methods primarily focus on full-body movements. For avoiding compensation during exercise is crucial to ensure action accuracy and maximize muscle effectiveness. Compensation occurs when the body uses other muscle groups or incorrect postures to compensate for deficiencies or incorrect positions of the primary muscles, which can lead to poor movement outcomes or even injuries. During fitness training, focusing on action accuracy, posture stability, and using target muscle groups without compensation is essential. Therefore, focusing specifically on a portion of the body structure related to target movements may be better than focusing on the full body. For example, when users perform chest and shoulder training, we only need to detect the upper part of the body.

In this work, we aim to improve the fitness experience through automated human action recognition technology to enhance gym safety and reduce operational costs. To achieve this, we propose a multi-task graph convolutional network based on spatio-temporal graph convolutional networks (ST-GCN), introducing a new partitioning strategy based on the 5PKC theory to explore joint relationships. Furthermore, to assess the correctness of users' actions and provide adjustment suggestions, we designed a comprehensive loss function that integrates two tasks. This allows us to accurately judge both the correctness and erroneous parts of movements, providing users with feedback for optimizing and improving their workout quality. In experiments, to assess whether utilizing different skeletons for localized movements improves accuracy, we collect a real dataset in a gym focused on upper-body movements, utilizing extracted upper-body skeletons for analysis. We also prove that the 5PKC partitioning strategy effectively recognizes human action on commonly used datasets such as Kinetics and NTU-RGB+D. 5PKC can improve performance compared to the position and connection skeleton information.

Overall, the contributions of this study are listed as follows.

- Developed a multi-task model using Spatio Temporal Graph Convolutional Networks for gym fitness.

- An automated fitness adviser system is designed to enhance the exercise experience in fitness gyms.
- Integrated Five Primary Kinetic Chains for better partitioning using skeleton structure and kinematic principles.
- Collected custom fitness data focused on upper body training for real-world experiments.

The remainder of the paper is organized as follows. The related works are discussed in Section 2. The proposed scheme and the application are detailed in Section 3. Experiment results are presented in Section 4, and conclusions are drawn in Section 5.

## 2. Related works

### 2.1. Image-based human action recognition

Traditional image-based recognition tends to utilize the information from RGB features. RGB features capture the colors and texture details of objects, which can provide rich visual information. The RGB features are also easy to obtain from the dataset. Tsai et al. [8] propose the OF-MHI method to incorporate an optical flow and a revised motion history image into a Motion History Representation. OF-MHI describes the local actions of each body part for action recognition. Eum et al. [9] propose a Depth-MHI-HOG(DMH) to separate the foreground and background of the image effectively. DMH extracts features by using depth information and recognizes actions through the HMM-based spotter model.

Inspired by the deep learning model such as Convolutional Neural Networks(CNNs) success in the computer vision area. There are many researches realize the advantages of CNNs in capture the RGB features for action recognition. CNNs are deep models capable of directly processing raw input and automating the feature construction process. Ji et al. [10] propose a 3D CNN model to investigate the fully automated recognition of movements in uncontrolled environments. The proposed 3D CNN model employs 3D convolutions to extract features from both spatial and temporal dimensions, capturing motion information from multiple adjacent frames. The proposed 3D CNNs also overcome the limitation of CNNs, which can only handle 2D inputs at that time. Tran et al. [11] propose a 3D ConvNets that is trained on a large-scale supervised video dataset. ConvNets applies a homogeneous architecture with a small $3 \times 3 \times$ three convolution kernels in all layers to capture the spatio-temporal feature. Many research studies have been proposed to enhance the performance of 3D CNN-based models through the integration of two-stream or multi-stream architecture. For example, Carreira et al. [12] introduce the two stream Inflated 3D CNN (I3D) to expand the filter and pooling kernels from the 2D CNNs into the 3D ConvNets. Wang et al. [13] propose a two-stream fusion function to integrate the 3D ConvNet and LSTM model for learning the long-term temporal dependencies.

Although 3D CNN-based models are efficient for modeling discriminative features in the spatiotemporal dimension, this information may not be enough for human action recognition. Some researchers try to adopt the depth of information in images for action recognition. Depth images provide distance information between objects and the environment, which is valuable for action recognition. Depth information helps differentiate relative positions and motion relationships among different objects or body parts, providing more detailed context for actions. Moreover, compared to RGB images, depth images are more robust in terms of background variations. Depth sensors directly measure the distance between objects and the sensor, reducing the impact of changes in light intensity and color. This enables depth images to provide motion information even in the presence of background changes. Wang et al. [13] proposed a novel method for human action recognition from depth images, which combines Weighted Hierarchical Depth Motion Maps (WHDMM) and 3ConvNets. Multiple time-scale WHDMMs are constructed to encode the spatiotemporal motion patterns of actions into a 2D spatial structure. Then, transforming WHDMMs into pseudo-colored images further enhanced the 2D spatial structure's recognition capability. Yang et al. [14] presented an approach for recognizing human actions from depth image sequences, where these sequences provide body shape and motion information for action recognition. In this approach, depth images were projected onto three orthogonal planes, and global activities were accumulated across the entire video sequence to generate Depth Motion Maps (DMM). Directional Gradient Histograms (HOG) were then computed from DMM as a representation of the action video. The method was tested on the Microsoft Research (MSR) Action3D dataset, which contains human actions captured by a depth camera. Yang et al. [15] proposed a novel framework for human action recognition using depth camera-captured video sequences. This method clustered hypersurface normal vectors from the depth sequences to form polynomials, which were used to jointly describe local motion and shape information. To capture spatial and temporal orders, an Adaptive Spatio-Temporal Pyramid was introduced, dividing the depth videos into a set of spatio-temporal grids. A new approach was then introduced to aggregate low-level polynomials into Super Normal Vectors (SNV), enhancing recognition accuracy.

Overall, image-based human action recognition methods continue to face numerous challenges, including variations in viewpoints, occlusions, lighting conditions, differences in human body size, and the speed at which actions are performed. Although depth information facilitates the discernment of the relative positioning and motion dynamics between various objects or body segments, offering a richer contextual understanding of actions. In practical applications, there are still many limitations and issues with poor performance.

### 2.2. Skeleton-based human action recognition

Image-based methods that do not utilize RGB or depth information will be affected by factors such as lighting conditions and poor image quality. For example, intense lighting may cause certain areas to be overexposed or distorted, while insufficient lighting can result in images being too dark or lacking clear details. These variations in lighting can potentially decrease the accuracy of action recognition and assessment, as critical motion details might not be accurately extracted from the images. Therefore, many studies have proposed focusing on the information on the skeleton structure for action recognition. Skeleton-based methods try to capture the key information from skeleton data instead of the original images so that it can avoid the effect of the quality of images. In real applications, we can obtain the skeleton sequence by applying pose estimation algorithms on RGB or depth images. Various methods adopt RNNs and LSTMs to model the sequential information within the skeleton sequences for human action recognition. Liu et al. [16] propose a spatio-temporal LSTM with trust gates to control the effect of context information from a sequence. This architecture allows the model to learn the tree-based skeleton structure rather than the traditional skeleton sequence. Song et al. [17] adopt the attention mechanism into human action recognition and propose the Spatio-Temporal Attention Model. A spatial attention module and a temporal attention module work jointly under the LSTM model. Lee et al. [18] introduce a novel approach to address skeleton feature representation and temporal dynamics modeling for recognizing human actions based on poses. The proposed generalized Temporal Sliding Long Short-term Memory (TS-LSTM) networks capture diverse temporal action dynamics for human action recognition. Liu et al. [19] introduce an improved method for skeleton visualization aimed at recognizing human actions regardless of viewing angle. Initially, a sequence-based, view-invariant transformation is applied to counter the effects of viewing angle variations on the spatial–temporal arrangement of skeleton joints. The modified skeletons are then represented as a sequence of color images, subtly embedding the joints' spatial–temporal information, enhanced further by visual and motion techniques to highlight local patterns. Finally, a convolutional neural network model

is utilized to extract robust and distinctive features from these enhanced color images. The improved skeleton visualization technique translates spatio-temporal skeleton data into color images, enriched with visual and motion enhancements, offering a concise yet uniquely identifiable representation.

Graph Convolutional Networks (GCN) [20] may be a more powerful network to effectively analyze skeleton information since skeletons are non-Euclidean graph structures with complex interactions and spatial distribution. GCN can capture the relative positions and interactions between joints, providing a better understanding of the structural features of skeletons. It can also consider the local information of each joint in the skeleton and propagate information on the graph through neighboring relationships between nodes. With the development of GCN, it has been widely applied in skeleton-based action recognition. Li et al. [21] introduced a CNN-based framework that simultaneously performs action classification and detection. Ke et al. [22] presented a new approach for 3D action recognition based on human skeletal data. This method first employs Deep Convolutional Neural Networks (DCNNs) Li et al. [23] introduced a co-occurrence feature learning approach for action recognition and detection based on skeleton data. This method leverages hierarchical aggregation to learn context information at different levels. Yan et al. [24] apply GCN to skeleton-based action recognition and propose a dynamic skeleton model called Spatial–Temporal Graph Convolutional Networks (ST-GCN). ST-GCN learns spatial and temporal structures from dynamic skeleton data, capturing not only the spatial relationships between joints but also extending the joint connections to the concept of time to capture relationships between joints in space and time across consecutive frames. This allows for leveraging the dynamic information of skeletons and providing more accurate action recognition capabilities. In addition, ST-GCN also proposes a spatial configuration partitioning strategy to redefine the neighbor relation of joint nodes within the human skeleton structure. Chen et al. [25] introduce a new approach called Channel-wise Topology Refinement Graph Convolution (CTR-GC), designed to dynamically learn various topologies and efficiently consolidate joint features across different channels for recognizing actions based on skeletons. Li et al. [26] introduce the Actional-Structural Graph Convolution Network (AS-GCN), combining actional-structural graph convolution with temporal convolution into a foundational building block. This structure is crafted to capture both spatial and temporal characteristics vital for action recognition. Additionally, a future pose prediction module is integrated alongside the recognition module, employing self-supervision to discern more intricate action patterns. Shi et al. [6] introduce a Two-Stream Adaptive Graph Convolutional Network (2s-AGCN) for human action recognition. 2s-AGCN is based on Graph Convolutional Networks (GCN), and it builds upon the foundation of the existing ST-GCN model. The motivation behind incorporating the two-stream network is that ST-GCN utilizes joint information from the skeleton while failing to fully exploit skeleton information, such as bone length and orientation, which are crucial for action recognition. To leverage this information for action recognition, the two-stream network was introduced into the model, consisting of the joint stream (J-Stream) and the bone stream (B-Stream). The J-Stream focuses on learning the features of joints and the relationships between them, while the B-Stream is responsible for learning the features of different bones and the relationships between bones. By incorporating the two-stream network to fully utilize both joint and bone features, the accuracy of action recognition is enhanced.

In human action recognition, most studies, such as DCNNs, ST-GCN, 2S-AGCN, etc., adopt accuracy as the evaluation standard for action identification. A minority of research also utilizes statistical tests to discuss the reliability of the models [27]. The skeleton-based approach yields better accuracy compared to methods using depth and sensor data alone. In addition, GCN-based approaches typically focus on the positional and connectivity data within the skeletal framework. Human skeleton information is depicted through adjacency matrices, with convolution processes applied to identify skeletal connections pertinent to the movements of the human body. However, within exercise science, the consistent and precise execution of a movement isn't just influenced by the functioning of neighboring muscle groups.

## 3. Methodology

In this section, we first introduce the proposed system for automated fitness advice in real applications. We would like to involve the human movement science concept to provide a more accurate relationship between the human skeleton and muscles. Five Primary Kinetic Chains(5PKC) is a human movement science that defines the primary physiological principle in human movement and describes the connection between bones and muscles. We propose a 5 Primary Kinetic Chains-based Multi-task Spatio Temporal Graph Convolutional Networks (5PM-STGCN). The proposed model adopts Five Primary Kinetic Chains(5PKC) as a partitioning strategy to explore the skeleton partition. For addressing the action recognition and evaluation issues, we design a multi-task objective function based on the Spatio Temporal Graph Convolutional Networks(ST-GCN) framework. In addition, we propose a skeleton partitioning strategy and the multi-task loss for model training in 5PM-STGCN. We will discuss each component of the proposed system and model in the following subsections.

### 3.1. System architecture

The architecture of the auto-fitness advised system is shown in Fig. 1. First, the user's fitness routine is captured using recording equipment or monitors. A segment of action footage is inputted and processed using the YOLO V7 model [28], transforming the video into a sequence of body skeleton information presented in 2D coordinates. This skeleton information includes joint coordinates and angles, which serve as features for nodes on the graph. Subsequently, we construct a spatial–temporal graph with the joints as graph nodes and the natural connectivities in human body structures and over time as graph edges. We then describe the connectivity between each pair of nodes on this graph with an adjacency matrix. The adjacency matrix is typically represented as a binary matrix, where the rows and columns correspond to all the nodes. The elements within the adjacency matrix indicate the connection relationships between nodes. This adjacency matrix is inputted into a human action recognition model for action classification and quality assessment. The human action recognition model used throughout this framework is our proposed 5PKC-based Multi-task Spatio Temporal Graph Convolutional Networks. Lastly, based on the analytical results from the model, we provide suggestions for refining the actions, assisting the user in improving and adjusting their execution technique for better results.

### 3.2. Spatio temporal graph convolutional networks

ST-GCNs [24] consists of multiple layers of spatial–temporal graph convolution operations. In each layer, it consisted of a spatial convolution network and a temporal convolution network. Spatial convolution networks deal with the skeleton information in a single frame of video. The temporal convolution network(TCN) models the spatial–temporal dynamics within skeleton sequences. For the spatial operations, the convolution formula is defined as follows,

$$f_{out}(v_{ti}) = \sigma_{v_{tj} \in B(v_{ti})} \frac{1}{Z_{ti}(v_{tj})} f_{in}(v_{vtj} \cdot w(l_{ti}(v_{tj}))) \quad (1)$$

where $f_{in}$ and $f_{out}$ represent the input and output feature maps, respectively. $v_{ti}$ serves a role similar to the center location in a traditional convolution, except it represents a vertex on the graph. $v_{tj}$ corresponds to all the nodes within one stride of $v_{ti}$, analogous to the neighboring nodes surrounding the center of a convolution filter. The weight function $w$ can be thought of as the parameters within a convolution filter, which are used to weigh each node. The term $l_{ti}(v_{tj})$ within $w$ refers to the subset composed of each neighboring node. The part $Z_{ti}$ acts as a normalizing term to ensure that each node's influence on the output is equalized.

In TCN, the temporal dimension of the graph is established by linking identical joints from one frame to the next. This method adopts

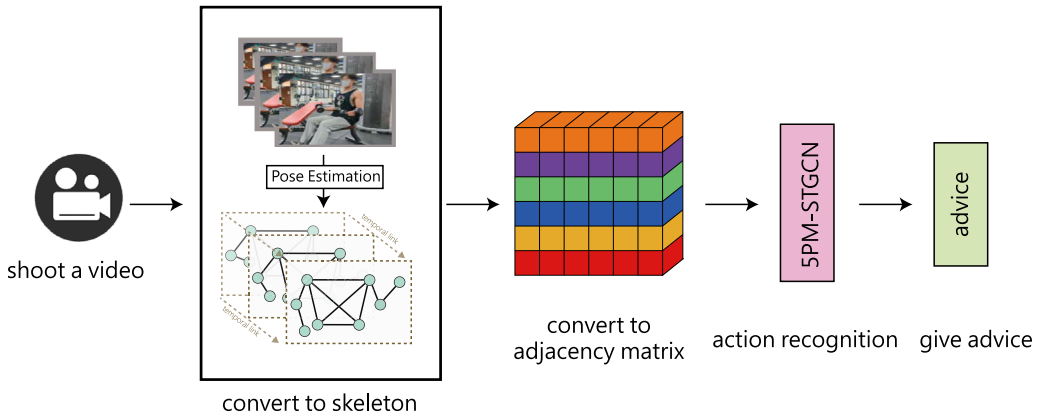



**Fig. 1.** Auto-fitness advisor system.

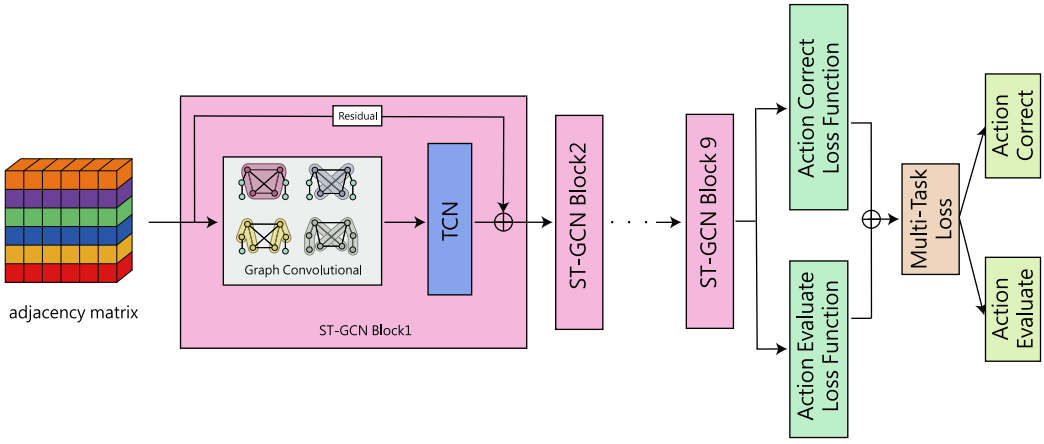



**Fig. 2.** 5PM-STGCN model architecture.

a straightforward approach to expand the spatial CNN into the spatial–temporal realm. Essentially, we broaden the notion of adjacency to encompass joints that are connected over time.

### 3.3. 5PKC-based multi-task spatio temporal graph convolutional networks

The 5PM-STGCN model architecture is shown in Fig. 2, which is based on ST-GCN. In the 5PM-STGCN, the ST-GCN blocks are the spatial–temporal graph convolution layers of the ST-GCN. Each block consists of a spatial convolution network and a temporal convolution network. When an adjacency matrix derived from an image is input, we introduce the concept of Five Primary Kinetic Chains (5PKC) partitioning to establish a relationship matrix that dictates how skeletal parts move in tandem during human motion. The obtained 5PKC-based skeleton partitioning matrix is used to replace the adjacency matrix that represents human skeletal information in the spatial convolution operation of the original ST-GCN. Then, to capture the temporal correlations of joints across frames, we employ a temporal convolution network. The complete model incorporates nine layers of ST-GCN spatial–temporal graph convolution layers. For the model's learning function, we designed a multi-task loss function that combines action correctness loss and action evaluation loss. This proposed multi-task loss function allows the model to simultaneously output classification predictions for both action correctness and action evaluation.

In this application, we would like to achieve fitness action recognition and provide action suggestions. Therefore, 5PM-STGCN model is designed with three primary functionalities: action recognition, action correctness assessment and action refinement suggestions. Through the action recognition capability of the 5PM-STGCN model, we can accurately identify the type of action the user is performing. Additionally, 5PM-STGCN model assesses the correctness of actions, determining whether the user is executing the movements and poses correctly. Specifically, We can separate the main idea of the 5PM-STGCN model into two parts, such as the 5PKC-based Skeleton Partitioning Strategy and the Multi-task objective loss.

#### 3.3.1. 5PKC-based skeleton partitioning strategy

According to the suggestions of the fitness expert, the actions usually focus on a part of the muscles for training. Therefore, we proposed to utilize the partial body information and focus on the related skeleton structures with target muscles. If we perform the workouts of the upper body, we only use the upper skeleton structure. In the later case, we will use the upper body as the example. Suppose we are given a skeletal sequence in which each frame consists of $N$ joints. Following the human skeletal diagram in Fig. 3, we construct a graph defined as $G = (V, E)$. The node features for the entire graph are defined as $V = v_i \in R^C | i = 1, \ldots, N$, where each node in the graph is defined as $v_i$. Here, $v_i$ is a C-dimensional feature vector, where C is the dimension of joint features. Each joint feature contains information about the $(x, y, z)$ coordinates and joint angles. The edges of the graph represent the connecting relationships between adjacent joints. All edges in the graph are defined as $E \in e_{ij}$, where each edge $e_{ij} = (v_i, v_j)$ signifies a connection between joints $v_i$ and $v_j$. This representation enables us to capture the connectivity between joints and their neighboring joints, facilitating analysis and modeling of the human skeletal structure.

Moreover, in the skeletal Graph Convolutional Network (GCN), we incorporated Schwartz's Five Primary Kinetic Chains (5PKC) [7] to devise a novel skeleton partitioning strategy. This strategy aims to extract valuable information from the human skeletal structure.

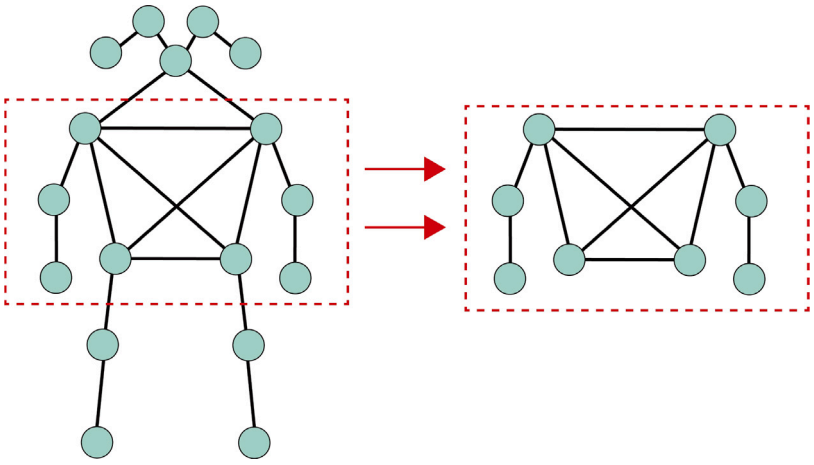


**Fig. 3.** Skeleton structure example.

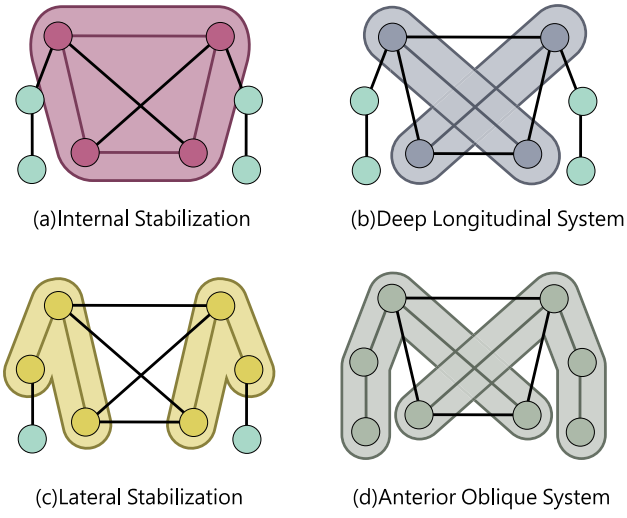



**Fig. 4.** The adjacency relation within four primary kinetic chains.

In the theory of the Five Primary Kinetic Chains, the human body consists of numerous interconnected joints. During the execution of movements, these joints form interconnected chain structures by transmitting forces among each other, referred to as Kinetic Chains. These kinetic chains serve distinct functional roles within the human body, and their distribution follows specific logic. The Five Primary Kinetic Chains are defined based on these functional roles and encompass the Internal Stabilization (IS) chain, the Deep Longitudinal System (DLS), the Lateral Stabilization (LS), the Posterior Oblique System (POS), and the Anterior Oblique System (AOS).

By integrating 5PKC into the partitioning strategy of GCN, the results of action recognition can closely resemble real human movement patterns. 5PKC offers a systematic approach to human body analysis rooted in anatomical structure and principles of action. However, since the human skeleton used in fitness action does not distinguish between front and back, we chose to exclude the posterior oblique system (POS) and only utilize the other four kinetic chains, which are shown in Fig. 4. Specifically, in each ST-GCN block, the skeleton structure adjacency matrix is divided into four submatrices, which correspond to the four primary kinetic chains. For each kinetic chain, we first divide the original adjacency matrix into the submatrix of the upper body. We redefine the adjacent relation as the node connection within each kinetic chain to fit the kinetic chain relation. Then, we can define the graph convolution function as follows:

$$V^{l+1} = \sigma\left(\sum_{i=1}^{4} \bar{A}_i V^l W_i^l\right) \tag{2}$$

where $V^l \in \mathbb{R}^{N\times C_l}$ represents all node features in layer $l$, and $V^{l+1} \in \mathbb{R}^{N\times C_{l+1}}$ represents all node features in layer $l+1$. $C_l$ and $C_{l+1}$ are the numbers of feature channels in layers $l$ and $l+1$, respectively. Here, the adjacency matrix $A$ is divided into four submatrices, so we need to combine them. $A_i$ refers to the $i$th submatrix extracted from the adjacency matrix $A$, and i is the index of the submatrix. $W_i^l$ represents the trainable weights for the $i$th submatrix. The network updates the features of each node based on the propagation of adjacent node features. This partition definition and strategy enables a more precise utilization of information within each submatrix for action recognition and analysis. It also better captures the features and structural connections of upper body movements.

### 3.3.2. Multi-task objective loss for training

We aim for our model to achieve two objectives: accurately recognizing the training actions the user is performing and evaluating the quality of the user's movements. Therefore, we employ two loss functions: the action recognizing correctness loss function, denoted as $L_correct$, and the action quality evaluation loss function, denoted as $L_evaluate$. The action recognizing correctness loss is calculated for the differences between the recognition action class and the ground truth class. The computation of the action recognizing correctness loss function $L_correct$ is as follows.

$$L_{correct} = -\frac{1}{N}\sum_{m=1}^{M_{correct}}\sum_{i=1}^{N} y_{im}\log\Big(P(Y_m|X_i)\Big), \tag{3}$$

where $M_{correct}$ signifies the number of action classes, $N$ represents the number of samples, $y_{im} \in 0, 1$ is the indicator function used to indicate whether the recognition class of the input sample $i$ is equal to the true class $m$. $P(Y_m|X_i)$ denotes the probability that sample $X_i$ belongs to class $Y_m$.

For evaluating the quality of the users' training movements, we consider the action during time and correctness of posture which are suggested from the fitness expert. We use four labels to class the action quality: the first label indicates the action during time is perfect and the pose is correct; the second label indicates the action during time is perfect, but the posture does not fit the standard; the third label indicates the action during time does not meet the criteria, but the pose is perfect; fourth label indicates both the action during time and the posture are not correct. Then, the action quality evaluation loss function can be formed as follows:

$$L_{evaluate} = -\frac{1}{N}\sum_{m=1}^{M_{evaluate}}\sum_{i=1}^{N} y_{im}\log\Big(P(Y_m|X_i)\Big), \tag{4}$$

where $M_{evaluate}$ represents the number of labels for evaluating motion quality, which in this case equals 4. $N$ is the number of samples, and $y_{im} \in 0, 1$ is the indicator function used to indicate whether the quality class of the input sample $i$ is equal to the true class $m$. $P(Y_m|X_i)$ denotes the probability that sample $X_i$ belongs to class $Y_m$.

Finally, we use the weighted sum to combine the action recognizing correctness loss and action quality evaluation loss as a multi-task loss. The multi-task loss function is defined as follows.

$$L_{multi} = \alpha L_{correct} + (1 - \alpha) L_{evaluate} \tag{5}$$

This loss function aids in categorizing fitness actions, evaluating their quality, and providing recommendations for adjusting them. By minimizing this loss value, our model can learn more precise and effective strategies for action recognition and evaluation, thereby enhancing the overall performance and accuracy of the model. In addition, we will discuss this $\alpha$ parameter through an ablation study in the experiment section.

### 3.3.3. Network architecture and training of 5PM-STGCN

We discuss the architecture setting of 5PM-STGCN in this section. 5PM-STGCN is based on ST-GCN model architecture, which is shown in Fig. 2 so that we process skeletal information through a stack of multiple ST-GCN (Spatial–Temporal Graph Convolutional Network) modules. The number of layers is 9, with the channel numbers per layer being 64, 64, 64, 128, 128, 128, 256, 256, 256. Each ST-GCN layer consists of a Graph Convolutional Network (GCN) and a Temporal Convolutional Network (TCN).

The initial learning rate is set to 0.1, and the model is trained for the first 100 epochs. After 100 epochs, the learning rate is reduced by a factor of 10 every 50 epochs. The Batch Size is set to 64.

The pseudo-code of the fitness action recognition framework using 5PM-STGCN is shown in Alg. 1. The first step is capturing the fitness routine video and segmenting it into $T$ data frame. For each frame, we need to extract the skeleton information and adopt a 5PKC-based skeleton partitioning strategy to represent the skeleton relation. The connections between joints within a single frame are encoded using an adjacency matrix. Then, we merge the adjacency matrix in all frames to a spatial–temporal tensor as the input of 5PM-STGCN. For a single frame, the implementation of ST-GCN using the first partitioning strategy is governed by the equation:

$$f_{out} = \Lambda^{\frac{1}{2}}(A + I)\Lambda^{\frac{1}{2}} f_{in} W \tag{6}$$

In this equation, the weight vectors for multiple output channels are compiled into the weight matrix $W$. For spatial–temporal cases, the input feature map is represented as a tensor with dimensions $(C, V, T)$, where $C$ is the number of channels, $V$ is the number of joints, and $T$ is the temporal dimension. To perform graph convolution, we conduct a $1\times\Gamma$ standard 2D convolution on the input tensor. We then multiply the output tensor with the normalized adjacency matrix $\Lambda^{\frac{1}{2}}(A+I)\Lambda^{\frac{1}{2}}$ along the second dimension, which corresponds to the joint connections.

### 3.3.4. Computation complexity discussion

The computation complexity is presented in detail below. First, in Step 1, the computation complexity is $\mathcal{O}(T)$, corresponding to the number of frames used for segmentation.

For each frame of the video, we need to convert the entire image into a pure skeletal feature map to be input into the subsequent model. This process involves four steps, Step 2 to Step 5: processing the video using YOLO V7 to extract skeleton information, constructing a spatial–temporal graph from the skeleton data, implementing the 5PKC-based skeleton partitioning strategy, and generating the feature map using this partitioning strategy. In Step 2, the computation complexity is dependent on the YOLO V7. From the authors of YOLO V7 presented in their work, YOLO V7 is a real-time object detector. Its recognition processing speed can reach up to approximately 160 FPS. By using YOLO V7 to process each frame of the video, the computational complexity can be considered as $\mathcal{O}(1)$. In Step 3, constructing a spatial graph from the skeleton data just change the skeleton information detecting from YOLO v7 into adjacency matrix format. The computation complexity is $\mathcal{O}(V \times V)$, where $V$ is the number of selected skeleton nodes. It is also equal to the number of node features in Eq. (2). In implementation, the $V$ is 8 since we only use the upper body skeleton structure. In Step 4, the 5PKC-based skeleton partitioning strategy is implemented, and the 5PKC-based adjacency matrix is calculated. The computation complexity of this step is $\mathcal{O}(4 \times V \times V)$ since we involve four primary kinetic chains. In Step 5, we generate the spatial–temporal feature map by concatenating the spatial graphs of each frame into a spatial–temporal feature map. The computation complexity is only $\mathcal{O}(1)$. For all frames of a video, the total computation complexity is the summation of these four steps times $T$. Therefore, the total computation complexity of data preparing for the training model is $\mathcal{O}(T \times (1 + V^2 + 4V^2 + 1)) = \mathcal{O}(T \times V^2)$

Then, we input the spatial–temporal feature map for training 5PM-STGCN model. The training complexity mainly depends on the training epochs and batch size. In each layer of convolution, the computation complexity is $\mathcal{O}(E)$, where $E$ is the number of edges within the 5PKC-based adjacency matrix $normalized_A$.

**Algorithm 1** Fitness Action Recognition using 5PM-STGCN

**Input:** Video footage of fitness routine
Step 1: Capture the fitness routine video
video ← CAPTURE_VIDEO()
segment video into $T$ frame
**for each** frame $t \in T$ **do**
  Step 2: Process video using YOLO V7 to extract skeleton information
  skeleton_data ← PROCESS_VIDEO_WITH_YOLOv7(video)
  Step 3: Construct a spatial graph from the skeleton data
  graph ← CONSTRUCT_SPATIAL_TEMPORAL_GRAPH(skeleton_data)
  Step 4: Implement the 5PKC-based Skeleton Partitioning Strategy
  partitioned_g← APPLY_5PKC_PARTITIONING(graph)
  Step 5: Generate the spatial-temporal feature map
  feature map (C, V, t) ← CREATE_FEATUREMAP(partition_g, t)
  $(C, V, T)$ ← merged_frames.APPEND((C, V, t))
**end for**
Step 6: Training the 5PM-STGCN model with the input feature maps
$f_{in}$ ← INITIALIZE_INPUT_FEATURE_MAP($C, V, T$)
$A$ ← INITIALIZE_ADJACENCY_MATRIX($V$)
$I$ ← IDENTITY_MATRIX($V$)
$D$ ← COMPUTE_DEGREE_MATRIX($A + I$)
$normalized_A$ ← NORMALIZE_ADJACENCY_MATRIX($A, D, I$)
$W$ ← INITIALIZE_WEIGHT_MATRIX($C$, $number\_of\_output\_channels$)
**for** Layer $l = 1 to L$ **do**
  **for each** node $v \in V$ **do**
    $f^l_{out}(v)$ ← GRAPH_CONVOLUTION($f^{l-1}_{in}$, $normalized_A$, $W$, $v$)
  **end for**
**end for**
action_category ← st_gcn_network.CLASSIFY_ACTIONS()
multi_task_loss ← CALCULATE_MULTI_TASK_LOSS(action_category)
Back propagation to adjust the $W$ WEIGHT_MATRIX
**return** $W$

## 4. Experiment

This section describes the setup of the experiments and datasets and introduces some comparing methods. First, all experiments are conducted on the PyTorch framework with an NVIDIA GeForce RTX 3090 GPU.

**Table 1**
Description of datasets statistic.

| Datasets #of | Samples | Subjects | Action Classes | Quality Labels |
|---|---|---|---|---|
| Custom Fitness | 1477 | 16 | 8 | 4 |
| Kinetics | 266,440 | N/A | 400 | N/A |
| NTU-RGB+D | 56,880 | 40 | 60 | N/A |

**Table 2**
Action classes.

| Actions | Correctness | Class Label |
|---|---|---|
| dumbbell curls | Correct | 0 |
| dumbbell curls | Incorrect | 1 |
| dumbbell flyes | Correct | 2 |
| dumbbell flyes | Incorrect | 3 |
| dumbbell front raises | Correct | 4 |
| dumbbell front raises | Incorrect | 5 |
| dumbbell shoulder presses | Correct | 6 |
| dumbbell shoulder presses | Incorrect | 7 |

**Table 3**
Action quality classes.

| During Time | Arm Movements | Class Label |
|---|---|---|
| Standard | Standard | 0 |
| Standard | Non Standard | 1 |
| Non Standard | Standard | 2 |
| Non Standard | Non Standard | 3 |

**Table 4**
Performance comparison on two tasks for different methods.

| Model | Action Recognition | Action Quality Evaluation |
|---|---|---|
| 2s-AGCN | 87.85% | 79.09% |
| ST-GCN | 94.07% | 86.16% |
| 5PM-2sAGCN | 98.31% | 81.37% |
| 5PM-STGCN | **98.59%** | **88.98%** |

### 4.1. Datasets

Currently, available fitness datasets primarily focus on whole-body actions and lack specifically captured data for individual body parts. To obtain more realistic fitness data, we conducted recordings in a gym and created a skeleton-based upper body exercise dataset. Otherwise, we also use the common datasets for human action recognition to evaluate the performance of our proposed method. We discuss these datasets in detail in this section.

#### 4.1.1. Custom fitness dataset

This dataset is designed to concentrate on upper body actions, aiming to reflect localized training scenarios. During the data collection process, fitness trainers with a background in exercise management provided guidance to ensure that participants performed fitness actions in accordance with research requirements. Their professional knowledge and experience were used to assess the correctness of the movements. This dataset encompasses four common fitness actions, as depicted in Fig. 5, including dumbbell curls, dumbbell flies, dumbbell front raises, and dumbbell shoulder presses. The detailed data description is shown in Table 1. We collected the data from 16 users, and each person performed these four actions repeatedly for 10 to 20 users. After we removed the uncompleted and unclear data, the total number of samples was 1477.

To accommodate the requirements of two tasks such as action recognition and quality evaluation, we employed two types of labels for each video in this dataset. The first type of label is used for action recognition and correctness assessment, as outlined in Table 2. This label set encompasses correctness classification labels for each action, distinguishing whether the action was performed correctly or not. Consequently, there are a total of 8 classes in this label.

The second type of label is utilized to determine both the duration of an action and its quality, as described in Table 3. These labels aid in determining whether the actions meet the predefined standards. The duration ranges from 1.7 to 1.9 s. The quality evaluation of actions primarily focuses on the arms, given that the dataset predominantly concentrates on the upper body region. Errors in arm movements effectively reflect quality in actions and are assessed based on erroneous segments. We determine whether the arm movements satisfy the quality standard from the expert's review and separate into standard and non standard classes. Consequently, the second type of label comprises four distinct labels.

#### 4.1.2. Kinetics dataset

Kinetic [12] is a large video dataset for human action recognition, which is commonly used for training and evaluating human action recognition models. This dataset includes up to 650,000 video clips corresponding to 400/600/700 different classes of action, for example, running, dancing, and swimming. Moreover, the videos include human–object interactions such as playing instruments, as well as human–human interactions such as shaking hands and hugging. The length of video clips is usually a few seconds to dozens of seconds. The Kinetics dataset aims to provide a rich and diverse data source for research and development. Researchers can use this dataset to advance the development of human action recognition. In this study, we use the kinetic dataset's fewer-data version, which contains around 266,440 video clips covering 400 classes.

#### 4.1.3. NTU-RGB+D dataset

NTU-RGB+D [29] is also a common dataset for human action recognition, which is proposed by Nanyang Technological University, Singapore. The dataset provider claims that this dataset is for academic research only and free to researchers for educational or non-commercial purposes. This dataset includes 56,880 video samples and 60 classes of human action, which refers to three major categories: daily actions, mutual actions, and medical conditions. The length of video samples is usually only a few seconds. In addition, the dataset contains two types of sub-dataset, namely Xsub and Xview. Xsub splits the dataset into 20 participants in the training set and 20 participants in the test set. Xview divides the data set according to the shooting perspective; the front perspective is used as the training set, and the side perspective is used as the test set.

### 4.2. Performance validation with comparing method

First, we aim to determine which model performs better on our custom fitness dataset. The input data utilized is our custom fitness dataset, capturing skeletal structure, including node features such as x, y, z coordinates and joint angles. The partitioning strategy follows the spatial structural partitioning strategy used in the two mentioned models, ST-GCN and 2s-AGCN. The accuracy results are shown in Table 4. The accuracy of action correctness recognition and quality evaluation for 2s-AGCN and ST-GCN are the results of the model performing the task respectively since these two models cannot deal with two tasks simultaneously. 5PM-STGCN is our proposed model for fitness action recognition. Similar with 5PM-STGCN, we use the upper body skeleton and adopt 5PKC skeleton partitioning strategy into the 2s-AGCN. 5PM-2sACGN is a designed method based on the 2s-AGCN model structure. Note that both 5PM-STGCN and 5PM-2sACGN are trained by multi-task objective loss and perform these two tasks simultaneously.

Although the 2s-AGCN paper suggests that their model outperforms ST-GCN, our experiments observe that the ST-GCN model yields superior results. This discrepancy might be due to the relatively small size of our dataset. A smaller dataset could limit the model's performance during training, as it may struggle to learn more complex patterns with fewer samples. In fitness action recognition, the ST-GCN model might

**Fig. 5.** Actions types in custom fitness dataset.

be better suited for our custom fitness dataset, as it tends to handle learning and training on smaller datasets more effectively. Furthermore, this outcome might also be attributed to differences between the datasets. The collected fitness dataset only focuses on upper body actions. These workout actions involve fewer bones compared to full-body actions, and the relative movements between bones may not be as intricate. This could make it challenging for the 2s-AGCN model to effectively model spatial relationships between bones, as fewer bones and simpler relationships could limit its learning capacity. In contrast, the ST-GCN model is simpler and more flexible, potentially more adept at capturing upper body movement features when the dataset primarily consists of upper body actions. Based on these experimental results, 5PM-STGCN is designed based on ST-GCN rather than 2s-AGCN. Moreover, 5PM-STGCN outperforms the comparing methods on our collected fitness dataset. The 5PM-STGCN model utilizing the upper body skeleton along with the 5PKC partitioning strategy achieved an action classification accuracy of 98.59% and a quality evaluation accuracy of 88.98%.

### 4.3. Ablation studies

In 5PM-STGCN, we design the model, which includes the component of partial skeleton structure, 5PKC partitioning strategy, and multi-task objective loss. In this section, we will discuss the impact of performance for each component in the proposed model. The input data employed in the ablation studies is the custom fitness dataset collected for this study. The first experiment is designed to compare different skeleton structure information and partitioning strategies under the same multi-task objective loss. We also designed another experiment to compare single-task loss and multi-task objective loss.

#### 4.3.1. Impact of skeleton structure and partitioning strategy

In this section, we conduct a comparative analysis between using different skeleton structure information and partitioning strategies.

**Table 5**
Performance impact of different components in 5PM-STGCN.

| Model | Action Recognition | Action Quality Evaluation |
|---|---|---|
| 5PM-STGCN | **98.59%** | **88.98%** |
| w/ whole&5PKC | 94.35% (−4.24%) | 87.85% (−1.13%) |
| w/ upper&SSP | 96.33% (−2.26%) | 88.42% (−0.56%) |
| w/ whole&SSP | 93.5% (−5.09%) | 86.44% (−2.54%) |

This aims to comprehend whether utilizing 5PKC as a partitioning strategy and focusing on the upper body skeleton can enhance the model's accuracy. We make use of both the full body skeleton and spatial structure partitioning for comparison. In this experiment, we adopt the multi-task objective loss function proposed in this study. This loss function integrates two task-specific loss functions: one focused on action recognition correctness and the other on action quality evaluation. Table 5 presents the experimental results under different configurations. In the conducted experimental setup with the action dataset, we observed that the configuration utilizing the upper body skeleton outperformed the configuration using the full body skeleton, both in terms of action recognition accuracy and action quality evaluation accuracy. This suggests that in the context of an action dataset primarily focused on the upper body, the upper body skeleton is more effective at capturing features relevant to action recognition. Furthermore, the ST-GCN model only employing the Five Primary Kinetic Chains (5PKC) as the skeletal partitioning strategy still improves the performance compared with the original ST-GCN on both tasks. The results demonstrate that 5PKC actually provides more useful information for the model.

#### 4.3.2. Impact of multi-task objective loss

In this experiment, we would like to demonstrate the effectiveness of the proposed multi-task objective loss. We design to use the single task loss to train and test the model for each task. For the action

**Table 6**
Performance impact of multi-task objective loss.

| Loss in 5PM-STGCN | Action Recognition | Action Quality Evaluation |
|---|---|---|
| $L_{correct}$ | **99.71%** | N/A |
| $L_{evaluate}$ | N/A | 86.54% |
| $L_{multi}$ | 98.59% | **88.98%** |

recognition task, we adopt Eq. (3) to train the model. The Eq. (4) is used to train the model for the action quality evaluation task. The input data used is from our custom fitness dataset collected for this research. We utilize the 5PM-STGCN model as the basis model along with the upper body skeleton and 5PKC partitioning strategy.

In order to adapt our dataset for use in the single loss approach, we reorganize the labels from Tables 2 and 3 into 32 labels, which are then employed in the single loss experiment.

The experimental results, as shown in Table 6, reveal that 5PM-STGCN based on single $L_{correct}$ loss achieves a higher action correctness and recognition accuracy of 99.71% compared to the multi-task loss model's accuracy of 98.59%. This could be attributed to the fact that a single loss function is more focused on optimizing the task of action correctness, allowing the model to better discriminate between different action categories. On the other hand, 5PM-STGCN based on single $L_{evaluate}$ loss achieves a lower action evaluation accuracy of 86.54% compared to the multi-task loss model's accuracy of 88.98%. This could be due to the fact that a single loss function might not adequately consider the features relevant to action quality evaluation task. The use of the multi-task loss function allows for more comprehensive learning and optimization of the model's ability in action quality evaluation prediction. In summary, although multi-task loss performs slightly less well on the single action correctness and recognition task than single loss, it can achieve these two tasks well simultaneously.

### 4.4. Hyper parameters discussion

In this section, we aim to discuss the impact of hyperparameters in 5PM-STGCN, such as the training epoch and the weighted sum $\alpha$ in equation Eq. (5). The input data used in this experiment is derived from our custom fitness dataset. We utilize the upper skeleton structure and 5PKC as the partitioning strategy, with node features encompassing xyz coordinates and joint angles. The loss function employed is the multi-task objective loss function proposed in this study.

### 4.5. Training epoch

To assess the stability of the model during training, we examined the convergence of action correctness and recognition accuracy, and loss values in the training process of the 5PM-STGCN model. Fig. 6 illustrates the variations in accuracy and loss values for each epoch during training.

Starting with accuracy, we notice an initial improvement during the early stages of training. This suggests that the model learns crucial features and information in the initial phases. Subsequently, accuracy continues to improve but at a slower rate. This indicates that the model keeps learning and refining its performance throughout training, though relatively slower than the initial stages. Towards the end, we observe a gradual flattening of the accuracy improvement trend. Despite some fluctuations, the overall accuracy remains at a high level. Moving on to the loss values, we observe a more significant decrease in the model's loss values during the first few epochs. Afterward, the loss reduction rate slows, fluctuations become smaller, and the loss values stabilize at a lower level in the final epochs. This suggests that the model gradually learns and improves its performance over training, resulting in a gradual reduction in loss values. By comparing the accuracy and loss values, we can deduce that accuracy increases gradually

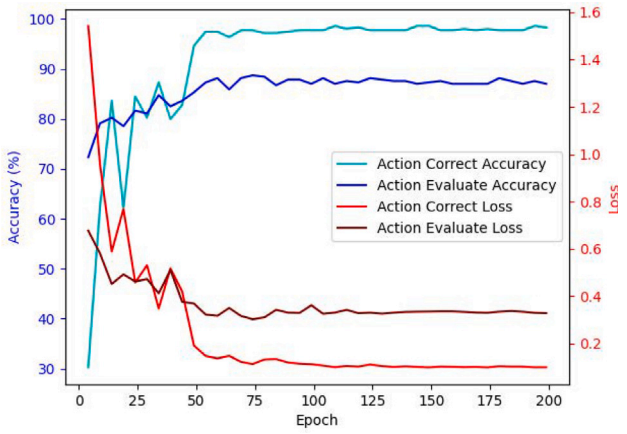



**Fig. 6.** The impact of training epoch on accuracy and loss.

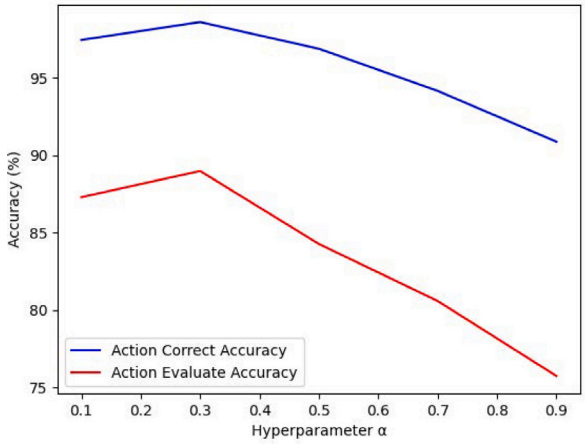



**Fig. 7.** The impact of $\alpha$.

throughout the training process, and simultaneously, loss values diminish gradually. This indicates that the model's performance progressively improves, resulting in more accurate predictions on the data. The sustained stability observed in the final stages signifies that the model has achieved satisfactory performance during training, maintaining a good predictive ability and low error levels. From this experimental result, the 150 training epochs may be the suitable setting in our custom fitness dataset.

### 4.6. Weighted parameter $\alpha$

This weighted parameter $\alpha$ indicates that the model increases focus on the action correctness recognition loss while reducing its focus on the action quality evaluation loss function. The result is shown in Fig. 7.

From the experimental result, we can obtain that the model achieves its highest accuracy when the $\alpha$ setting is at 0.3. This suggests that the model is able to strike a suitable balance between the emphasis on the action correctness recognition loss function and the quality evaluation loss function at this particular hyperparameter configuration. However, with an increase in $\alpha$ hyperparameters, there is a noticeable decrease in action assessment accuracy, and to a certain extent, a decrease in action correctness recognition accuracy as well. This indicates a trade-off between the action correctness loss function and the action assessment loss function as $\alpha$ hyperparameters increase.

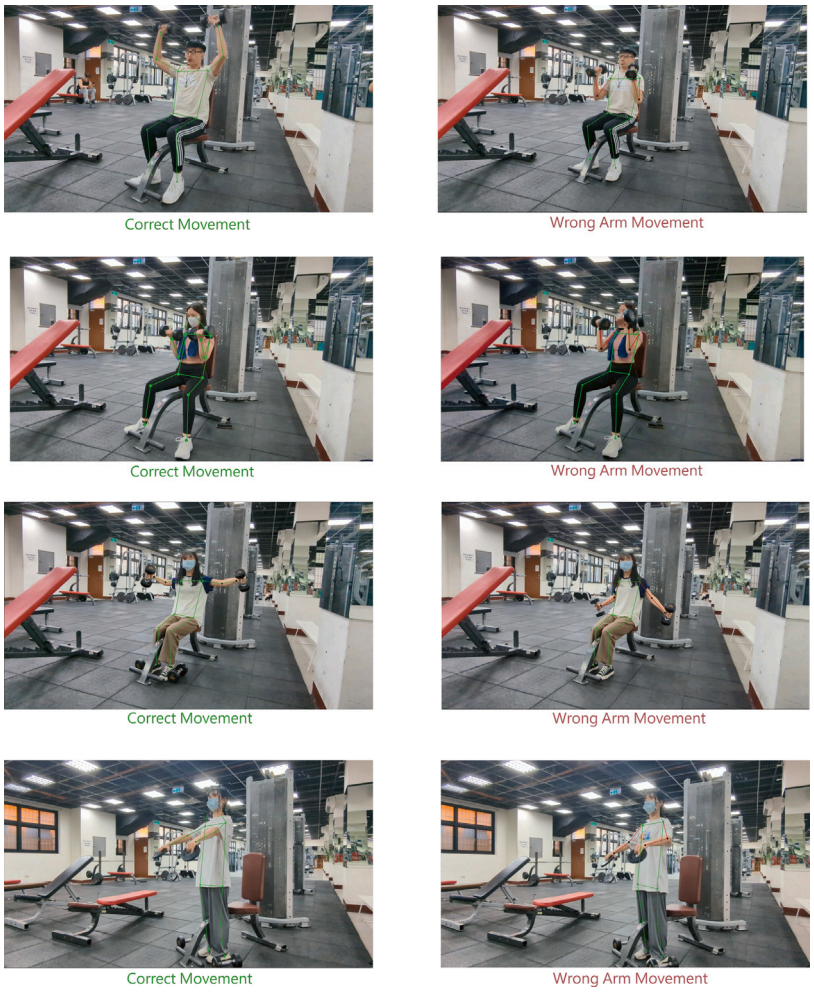



**Fig. 8.** The real case of application scenario.

### 4.7. Case studies

In this section, we show some real cases for the proposed auto-fitness advisor system in the gym. By applying this system, we can recognize the user's action and detect the correctness of action in a visualization presentation. When a user is performing the fitness actions, the system will capture the skeleton structure from the user and draw in the green line. During the action movements, the system continuously detects the correctness of that action. If the duration time of the user's movement does not fit the standard or the action is performed not well, the system will show the advice and draw a better trajectory of movement for users. The visualization is shown in Fig. 8; we select one example for a user performing well and another one for a user who does not qualify for each action in our custom fitness dataset. Take a dumbbell curl as example, when user is performing well, the system give us a correct feedback. Otherwise, the system gives us advice such as the wrong arm movement or duration time does not fit the standard. In addition, the system shows the red line for suggesting a better movement trajectory. Users can clearly understand their own movement errors, thereby helping users improve their movement skills. It also reduces the risk of injury and establishes correct movement memory and body perception to achieve the effect of improving movement accuracy.

### 4.8. Performance evaluation on action recognition datasets

In this section, we further design the experiments on Kinetics and NTU-RGB-D datasets to demonstrate the effectiveness of concepts in 5PM-STGCN. However, applying 5PM-STGCN in these datasets is limited since they do not provide the action quality evaluation label. We only apply the single $L_{c}orrect$ loss instead of the multi-task objective loss in this experiment. We also adopt the whole body skeleton structure since the actions in these dataset are not focus on partial body.

**Table 7**
Action recognition accuracy comparison on different datasets.

| Models | Kinetics | NTU-RGB-D:xsub | NTU-RGB-D:xview |
|---|---|---|---|
| ST-GCN | 30.7% | 81.5% | 86.68% |
| 5PM-STGCN | 31.32% | 81.54% | 88.3% |
| 2s-AGCN | 36.1% | 88.5% | 95.1% |
| 5PM-2sAGCN | **36.40%** | **88.68%** | **95.21%** |

In this experiment, we perform the experiments on Kinetics and NTU-RGB-D datasets to compare the performance. The results are shown in Table 7. In Table 7, the experimental results show that in common action recognition datasets, the objective task for recognition models is only to classify the action categories. Hence, the loss function used is only the action recognition loss function. Our proposed 5PKC-based model uses a multi-task loss function that combines action recognition and action quality evaluation, which may not fully leverage the model's potential on these datasets. As a result, the improvement in accuracy is limited. Nevertheless, there is still an overall improvement, indicating that incorporating the Human Movement Science concept into the upper body convolution operations does enhance action recognition performance to a certain extent.

In conclusion, our experimental results demonstrate that in the 5PKC-based model, utilizing the upper body skeleton in conjunction with the Five Primary Kinetic Chains (5PKC) partitioning strategy yields better action recognition outcomes. Additionally, the multi-task objective loss function that integrates both action correctness and action quality provides complementary information, thus enhancing the model's performance. This outcome holds significance for the development of auto-fitness advisor system. Furthermore, from the experimental results in Section 4.2, the performance ST-GCN is better than 2s-AGCN in our custom fitness dataset, which may be caused by the size and diversity of the dataset. In the future, we plan to apply our model to more larger size and diverse fitness action datasets to further validate its effectiveness.

## 5. Conclusion

In this study, the aim was to enhance fitness facility safety and reduce operational costs. The development of 5PM-STGCN, a multi-task model based on the ST-GCN framework, and an auto-fitness advisor system sought to improve the exercise experience in fitness gyms. Data collected from fitness gyms facilitated the creation of a skeleton-based upper-body exercise dataset. This dataset served as input for the model, allowing investigation into whether using different skeletons for localized movements instead of utilizing the full body skeleton could improve accuracy.

The partitioning strategy incorporated the Five Primary Kinetic Chains concept, which combines body structure and kinematic principles, to enhance the model's skeletal analysis capabilities. This approach better captures the overall characteristics of movements, thereby enhancing action recognition accuracy. Additionally, a multi-task objective loss function was designed to assess and provide suggestions for users' movements, handling multiple tasks simultaneously. This function allows for the simultaneous assessment of movement correctness and identification of movement errors, providing corresponding feedback.

Subsequent experiments on the custom fitness dataset validated the effectiveness of 5PM-STGCN. When using ST-GCN on the dataset, the model achieved higher accuracy compared to 2s-AGCN. Utilizing upper body skeletons for localized fitness movements improved accuracy. Adopting a partitioning strategy focused on the Five Primary Kinetic Chains system enhanced the model's skeletal analysis capabilities. Finally, the multi-task objective loss function yielded better results in the application compared to using single losses.

## CRediT authorship contribution statement

**Jia-Wei Chang:** Supervision, Project administration, Methodology, Formal analysis, Conceptualization. **Ming-Hung Chen:** Validation, Supervision, Methodology, Data curation. **Hao-Shang Ma:** Writing – review & editing, Writing – original draft, Supervision, Methodology, Formal analysis. **Hao-Lan Liu:** Validation, Software, Methodology, Data curation.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Data availability

The authors do not have permission to share data.

## Acknowledgement

This study was supported by a grant from the National Science and Technology Council, R.O.C (No. NSTC 112-2221-E-025-014). We thank the National Science and Technology Council for funding this study.

## References

[1] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (2015) 436–444.
[2] D. Duvenaud, D. Maclaurin, J. Aguilera-Iparraguirre, R. Gómez-Bombarelli, T. Hirzel, A. Aspuru-Guzik, R.P. Adams, Convolutional networks on graphs for learning molecular fingerprints, in: Proceedings of the 28th International Conference on Neural Information Processing Systems, NIPS'15, MIT Press, 2015, pp. 2224–2232.
[3] K. Liu, L. Gao, N.M. Khan, L. Qi, L. Guan, A vertex-edge graph convolutional network for skeleton-based action recognition, in: Proceedings of 2020 IEEE International Symposium on Circuits and Systems, ISCAS, 2020, pp. 1–5, http://dx.doi.org/10.1109/ISCAS45731.2020.9181235.
[4] B. Li, X. Li, Z. Zhang, F. Wu, Spatio-temporal graph routing for skeleton-based action recognition, in: Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, AAAI Press, 2019, pp. 8561–8568, http://dx.doi.org/10.1609/aaai.v33i01.33018561.
[5] X. Gao, W. Hu, J. Tang, J. Liu, Z. Guo, Optimized skeleton-based action recognition via sparsified graph regression, in: Proceedings of the 27th ACM International Conference on Multimedia, 2019, pp. 601–610, http://dx.doi.org/10.1145/3343031.3351170.
[6] L. Shi, Y. Zhang, J. Cheng, H. Lu, Two-stream adaptive graph convolutional networks for skeleton-based action recognition, in: Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2019, pp. 12018–12027.
[7] J. Schwartz, The 5 Primary Kinetic Chains, Lenka Tulenka, 2019.
[8] D.-M. Tsai, W.-Y. Chiu, M.-H. Lee, Optical flow-motion history image (OF-MHI) for action recognition, Signal Image Video Process. 9 (2015) 1897–1906.
[9] H. Eum, C. Yoon, H. Lee, M. Park, Continuous human action recognition using depth-MHI-HOG and a spotter model, Sensors 15 (3) (2015) 5197–5227.
[10] S. Ji, W. Xu, M. Yang, K. Yu, 3D convolutional neural networks for human action recognition, IEEE Trans. Pattern Anal. Mach. Intell. 35 (1) (2013) 221–231, http://dx.doi.org/10.1109/TPAMI.2012.59.
[11] D. Tran, L. Bourdev, R. Fergus, L. Torresani, M. Paluri, Learning spatiotemporal features with 3D convolutional networks, in: Proceedings of the 2015 IEEE International Conference on Computer Vision, ICCV, IEEE Computer Society, 2015, pp. 4489–4497.
[12] J. Carreira, A. Zisserman, Quo vadis, action recognition? A new model and the kinetics dataset, in: Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2017, pp. 4724–4733.
[13] P. Wang, W. Li, Z. Gao, J. Zhang, C. Tang, P.O. Ogunbona, Action recognition from depth maps using deep convolutional neural networks, IEEE Trans. Hum.-Mach. Syst. 46 (4) (2016) 498–509.
[14] X. Yang, C. Zhang, Y. Tian, Recognizing actions using depth motion maps-based histograms of oriented gradients, in: Proceedings of the 20th ACM International Conference on Multimedia, 2012, pp. 1057–1060.
[15] X. Yang, Y. Tian, Super normal vector for activity recognition using depth sequences, in: Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 804–811, http://dx.doi.org/10.1109/CVPR.2014.108.
[16] J. Liu, A. Shahroudy, D. Xu, G. Wang, Spatio-temporal LSTM with trust gates for 3D human action recognition, in: Proceedings of European Conference on Computer Vision, ECCV 2016, 2016, pp. 816–833.
[17] S. Song, C. Lan, J. Xing, W. Zeng, J. Liu, An end-to-end spatio-temporal attention model for human action recognition from skeleton data, in: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, AAAI Press, 2017, pp. 4263–4270.
[18] I. Lee, D. Kim, S. Lee, 3-D human behavior understanding using generalized TS-LSTM networks, IEEE Trans. Multimed. 23 (2021) 415–428, http://dx.doi.org/10.1109/TMM.2020.2978637.
[19] M. Liu, H. Liu, C. Chen, Enhanced skeleton visualization for view invariant human action recognition, Pattern Recognit. 68 (2017) 346–362.
[20] T.N. Kipf, M. Welling, Semi-supervised classification with graph convolutional networks, in: Proceedings of the 5th International Conference on Learning Representations, 2017.
[21] C. Li, Q. Zhong, D. Xie, S. Pu, Skeleton-based action recognition with convolutional neural networks, in: Proceedings of IEEE International Conference on Multimedia and Expo Workshops, ICMEW, 2017, pp. 597–600.
[22] Q. Ke, M. Bennamoun, S. An, F. Sohel, F. Boussaid, A new representation of skeleton sequences for 3D action recognition, in: Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2017, pp. 4570–4579, http://dx.doi.org/10.1109/CVPR.2017.486.
[23] C. Li, Q. Zhong, D. Xie, S. Pu, Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation, in: Proceedings of the 27th International Joint Conference on Artificial Intelligence, AAAI Press, 2018, pp. 786–792.

[24] S. Yan, Y. Xiong, D. Lin, Spatial temporal graph convolutional networks for skeleton-based action recognition, in: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence, AAAI'18/IAAI'18/EAAI'18, AAAI Press, 2018, pp. 7444–7452.

[25] Y. Chen, Z. Zhang, C. Yuan, B. Li, Y. Deng, W. Hu, Channel-wise topology refinement graph convolution for skeleton-based action recognition, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 13359–13368.

[26] M. Li, S. Chen, X. Chen, Y. Zhang, Y. Wang, Q. Tian, Actional-structural graph convolutional networks for skeleton-based action recognition., in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 3595–3603.

[27] D. Li, M.-R. Jiang, M.-W. Li, W.-C. Hong, R.-Z. Xu, A floating offshore platform motion forecasting approach based on EEMD hybrid ConvLSTM and chaotic quantum ALO, Appl. Soft Comput. 144 (2023) 110487.

[28] C.-Y. Wang, A. Bochkovskiy, H.-Y.M. Liao, YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors, 2022, arXiv preprint arXiv:2207.02696.

[29] A. Shahroudy, J. Liu, T.-T. Ng, G. Wang, NTU RGB+D: A large scale dataset for 3D human activity analysis, in: Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2016, pp. 1010–1019, http://dx.doi.org/10.1109/CVPR.2016.115.