EGCN: An Ensemble-based Learning Framework for Exploring Effective Skeleton-based Rehabilitation Exercise Assessment

Bruce X.B. Yu , **Yan Liu*** , **Xiang Zhang** , **Gong Chen** and **Keith C.C. Chan**
The Hong Kong Polytechnic University
{csxbyu, csyliu, csxgzhang, csgchen, cskcchan}@comp.polyu.edu.hk

Abstract

Recently, some skeleton-based physical therapy systems have been attempted to automatically evaluate the correctness or quality of an exercise performed by rehabilitation subjects. However, in terms of algorithms and evaluation criteria, the task remains not fully explored regarding making full use of different skeleton features. To advance the prior work, we propose a learning framework called Ensemble-based Graph Convolutional Network (EGCN) for skeleton-based rehabilitation exercise assessment. As far as we know, this is the first attempt that utilizes both two skeleton feature groups and investigates different ensemble strategies for the task. We also examine the properness of existing evaluation criteria and focus on evaluating the prediction ability of our proposed method. We then conduct extensive cross-validation experiments on two latest public datasets: UI-PRMD and KIMORE. Results indicate that the model-level ensemble scheme of our EGCN achieves better performance than existing methods. Code is available: https://github.com/bruceyo/EGCN.

1 Introduction

It has been expected to have a growing worldwide burden caused by musculoskeletal disorders [Sebbag *et al.*, 2019]. Physical therapists often conduct therapeutic exercises for different rehabilitation phases of musculoskeletal disorders such as back pain, sprains, and epicondylitis [Wyss and Patel, 2012]. However, regular rehabilitation therapy episodes in hospital settings are often unaffordable for patients due to their inadequate working ability [Machlin *et al.*, 2011]. Accordingly, home- or office-based rehabilitation programs initiated with the supervision of a therapist becomes a cost-effective alternative [Jessep *et al.*, 2009]. Keeping exercise regimens in a home-based setting, however, is hard for patients to adhere to, which can even lead to higher healthcare expenditure [Bassett and Prapavessis, 2007]. In recent years, some home-based physical therapy systems that utilize the 3D skeleton data collected by motion sensors have been attempted to evaluate the quality or correctness of exercises performed by rehabilitation patients [Komatireddy *et al.*, 2014; Saraee *et al.*, 2017]. Such exercise assessment systems could provide patient-identified barriers for increasing adherence to home-based rehabilitation exercises, which works as a professional therapist to motivate patients to do the therapeutic exercises [Karmali *et al.*, 2014].

Motion sensors such as Kinect and motion capture are used to collect skeleton data in existing solutions [Ahad *et al.*, 2019]. The 3D skeleton data is a sequence of skeleton joints featured with 3D position and 3D orientation (i.e., the angle of a skeleton joint) as shown in Figure 1. By analyzing the joint movement patterns thereof, abnormalities in exercises can then be detected [Lei *et al.*, 2019]. This tasks is a more challenging task, especially comparing with the action recognition task.Specifically, different rehabilitation exercises such as "deep squat" and "sit to stand" can be easily recognized, but evaluating the correctness or quality of a single recognized exercise is a more fine-grained task.

In addition to the challenges, exercise assessment in the 3D skeleton data remains not well tackled. First, existing methods [Williams *et al.*, 2019; Liao *et al.*, 2020a; Bruce *et al.*, 2021b] did not make good use of the position and orientation features together as they mainly rely on single modal skeleton feature. Second, existing evaluation standards used in [Williams *et al.*, 2019; Liao *et al.*, 2020a] impede the exploration of effective algorithms as they do not rely on the prediction results and can be hardly further improved based on the results in [Bruce *et al.*, 2021b].

We observe that the position and orientation streams of skeleton data (see Figure 1) are structurally homogeneous but they are heterogeneous in terms of their physical meanings. Specifically, the position feature represents the global structural movement of an exercise. While the orientation feature describes the local characteristic of skeleton joints, which is relatively more independent from one to another. These two skeleton features are not mutually convertible. To explore effective algorithms for skeleton-based exercise assessment with both the position and orientation features, we propose a learning framework called Ensemble-based Graph Convolutional Network (EGCN) that contains various ensemble strategies. Our main contributions are as follows:

- As far as we know, we are the first to explore the ef-

[*] Corresponding Author

fects of position and orientation features of skeleton data for rehabilitation exercise assessment. Our EGCN explores ensemble strategies at various levels including data level, feature level, decision level, and model level.

- We provide in-depth analysis regarding the properness of existing evaluation criteria [Liao *et al.*, 2020a; Williams *et al.*, 2019] for skeleton-based exercise assessment methods. Based on the analysis, we adopt cross-validation to validate the effectiveness of methods regarding their prediction abilities.
- We conduct extensive experiments on two latest public datasets: UI-PRMD [Vakanski *et al.*, 2018] and KIMORE [Capecci *et al.*, 2019], the proposed model-level ensemble in our EGCN significantly outperforms not only other ensemble strategies but also state-of-the-art GCN-based single modal methods.

2 Related Work

2.1 Skeleton-based Exercise Assessment

We briefly review the related work of skeleton-based exercise assessment in two perspectives: datasets and algorithms. Some exercise assessment datasets have been briefly reviewed in [Ahad *et al.*, 2019], where only the UI-PRMD [Vakanski *et al.*, 2018] dataset is relevant to this study since the other datasets are either focusing on action classification or not skeleton-based. Comparing with UI-PRMD, KIMORE is collected with real patient subjects and it also provides clinical evaluation. According to the surveyed datasets in [Liao *et al.*, 2020b], we use UI-PRMD and KIMORE for our experiments as they are the two latest public datasets for skeleton-based exercise assessment.

In terms of algorithms, they can be roughly grouped to regression- and non-regression-based methods. For regression-based methods, early works using various Hidden Markov Model (HMM) models were compared in [Tao *et al.*, 2016]. [Elkholy *et al.*, 2019] proposed a similar HMM-based method that has less computational overhead than [Tao *et al.*, 2016]. Recently, a deep learning framework [Liao *et al.*, 2020a] was proposed to encode the skeleton data of the UI-PRMD dataset, which is supervised by a quality score function. The training process of [Elkholy *et al.*, 2019] is supervised by the score of abnormality degree (on the scale of 1 to 5) evaluated by a professional specialist. Unlike the regression-based methods that require the supervision of clinical scores or a score function, some non-regression-based methods [Bruce *et al.*, 2020; Bruce *et al.*, 2021b] were proposed to deliver a numerical exercise evaluation score by utilizing the probability results of the SoftMax classifier or transforming outputs before the SoftMax layer via a sigmoid function. We follow the works of non-regression-based methods that treat the problem as abnormality prediction by considering the effects of different skeleton features.

2.2 Ensemble Learning

Ensemble learning is a hot research topic that aims to integrate data fusion, data modeling, and data mining into a unified framework [Dong *et al.*, 2020]. Typical ensemble methods usually achieve better performance through a proper combination mechanism. Otherwise, simply combining ensemble members might jeopardize the overall performance. For the classification task, [Dong *et al.*, 2020] categorized ensemble methods to data-level, feature-level, decision-level, and model-level. These different levels are also known as data fusion strategies [Du and Swamy, 2019]. Motivations behind ensemble learning can be forcing the diversity or independence of submodels, focusing on local information, and aggregation mechanism [Sagi and Rokach, 2018]. In the regime of deep learning, ensemble methods have seldom been surveyed although there are many ensemble-based methods being proposed. With deep learning, typical ensemble methods usually combine or fuse the feature-level representations of different data streams, or aggregate their results at the decision level [Baltrušaitis *et al.*, 2019]. Meanwhile, it is also possible to follow the traditional motivation of ensemble methods by forcing the feature-level diversity or independence of small classifiers [Ross *et al.*, 2020]. Otherwise, proper learning methods that fuse different data streams at the model level need to be proposed based on the comprehensive understanding of the data characteristics, which is also known as modal-based fusion [Bruce *et al.*, 2021a] or co-learning [Baltrušaitis *et al.*, 2019]. In our proposed learning framework, we design a model-level fusion method with an effective training strategy and compare it with various general ensemble strategies proposed in our EGCN.

3 Proposed Method

In this section, we introduce our skeleton-based exercise assessment method. We first introduce the GCN model adopted to extract features from the position and orientation streams of the skeleton data. Then, we introduce various ensemble strategies in our EGCN.

3.1 Data Structure and Notation

With N data samples, the exercise repetitions of a dataset can be represented as $S = \left\{S^{(n)} | n = 1, \ldots, N\right\}$. An exercise repetition $S^{(n)}$ that begins at time $t = 1$ and ends at time T with skeleton frames collected at regular intervals can, therefore, be represented as a set of $T \times J$ skeleton joints $S^{(n)} = \{S^{(n)}_{ti} |\ t = 1, ...T,\ i = 1, \ldots, J\}$, where J is the total number of skeleton joints. Here, a skeleton joint $S^{(n)}_{ti} = (P^{(n)}_{ti}, O^{(n)}_{ti})$ has two groups of features include the position feature $P^{(n)}_{ti}$ and the orientation feature $O^{(n)}_{ti}$. The position feature $P^{(n)}_{ti} = (x, y, z)$ has 3 attributes featured the 3D cartesian coordinates of the position. The orientation feature $O^{(n)}_{ti} = (X, Y, Z)$ also has 3 attributes that could be transformed to pitch, roll, and yaw values of the joint.

3.2 Representing the Skeleton Data

The skeleton frame is streamed as an ordered list of skeleton joints that has the position and orientation attributes. A complete exercise repetition contains varied lengths of such skeleton frames. We adopt a graph to represent the spatially and temporally structured information among these joints. The structure and the traversal rules of the graph follows

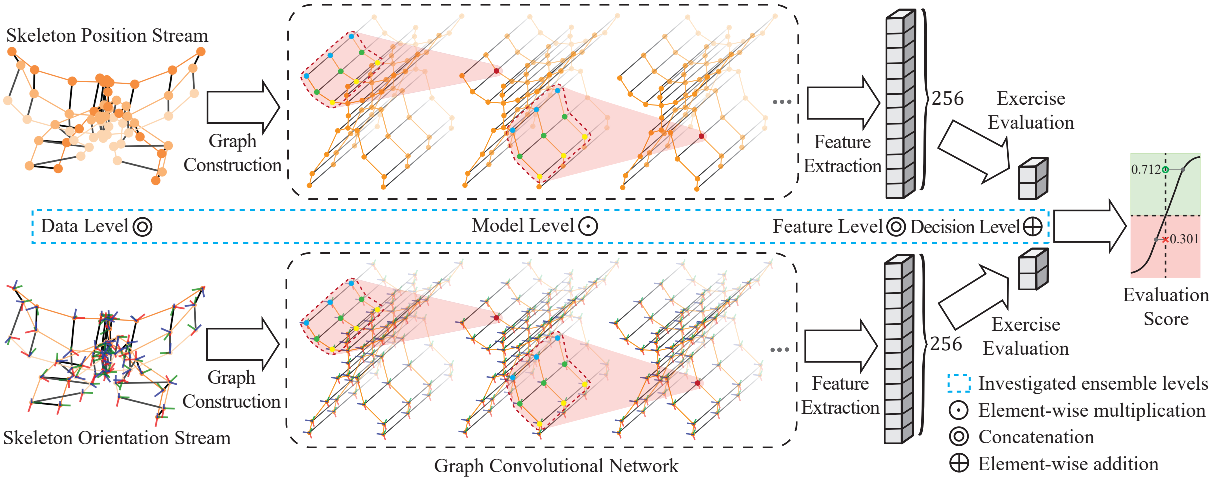

Figure 1: Illustration of our EGCN learning framework. The framework has two inputs (i.e., skeleton position and skeleton orientation streams) that are fed into graph convolutional networks for feature extraction. Four ensemble strategies at different levels (i.e., data level, model level, feature level, and decision level) are illustrated in the blue dashed line rectangle area.

the works of 2T-GCN [Bruce *et al.*, 2021b]. The skeleton graph at time t could be represented as $\vartheta_t = \{\upsilon_t, \varepsilon_t\}$, where $\upsilon_t = \{\upsilon_{ti} | \upsilon_{ti} = S_{ti}^{(n)}, i = 1, \ldots, J\}$ denotes the graph vertexes containing all the skeleton joints. While ε_t denotes the spatial edges representing the skeleton bones. The attributes of each graph vertex are featured with the position and orientation streams of the corresponding skeleton joint.

To perform convolutional operations, the convolutional sampling area of a graph vertex υ_{ti} is defined as a neighbor set $N(\upsilon_{ti})$. Specifically, the strategy empirically uses 3 spatial subsets: the vertex itself, the centripetal subset that contains the neighboring vertexes being closer to the center of gravity, and the centrifugal subset that contains the neighboring vertexes being farther from the gravity center. Assuming there are a fixed number of K subsets in $N(\upsilon_{ti})$, these subsets will be numerically indexed with a mapping $l_{ti} : N(\upsilon_{ti}) \rightarrow \{0, \ldots, K-1\}$. For a graph vertex υ_{ti}, the convolutional operation on the spatial dimension could be calculated as

$$f_{out} = \sum\nolimits_{\upsilon_{tj} \in N(\upsilon_{ti})} \frac{1}{Z_{ti}(\upsilon_{tj})} f_{in}(\upsilon_{tj}) W(l(\upsilon_{tj})) \qquad (1)$$

where υ_{tj} is one graph vertex of the defined neighbor set, $f_{in}(\upsilon_{tj})$ is a mapping getting the attribute vector of υ_{tj}, $W(l(\upsilon_{tj}))$ is a weight function $W(\upsilon_{ti}, \upsilon_{tj}) : N(\upsilon_{ti}) \rightarrow \mathbb{R}^c$ that could be implemented with a tensor of (c, K) dimensions. Here, c is the number of feature attributes. $Z_{ti}(\upsilon_{tj}) = |\{\upsilon_{tk} | l_{ti}(\upsilon_{tk}) = l_{ti}(\upsilon_{tj})\}|$ equals to the cardinality of the corresponding subset, which performs as a normalization term.

For a spatial skeleton frame, the spatial convolutional layer could be implemented by an adjacency matrix $\mathbf{A}$ with its elements indicating if a vertex υ_{tj} belongs to a subset of $N(\upsilon_{ti})$. The graph convolution is implemented by performing a 1×1 classical 2D convolution and multiplies the output tensor with a normalized adjacency matrix $\mathbf{\Lambda}^{-\frac{1}{2}} \mathbf{A} \mathbf{\Lambda}^{-\frac{1}{2}}$ on the second dimension, where $\mathbf{\Lambda}^{ii} = \sum_j (\mathbf{A}^{ij}) + \alpha$ is a diagonal matrix with α set to 0.001 to avoid empty rows. With K sampling strategies $\sum_{k=1}^{K} \mathbf{A}_k$, Equation 1 could be transformed as

$$\mathbf{f}_{out} = \sum\nolimits_{k=1}^{K} \mathbf{\Lambda}_k^{-\frac{1}{2}} \mathbf{A}_k \mathbf{\Lambda}_k^{-\frac{1}{2}} \mathbf{f}_{in} \mathbf{W}_k \odot \mathbf{M}_k \qquad (2)$$

where $\mathbf{W}_k$ is a weight tensor of the 1×1 convolutional operation with $(C_{in},\ C_{out},\ 1,\ 1)$ dimensions, which represents the weighting function of Equation 1. $\mathbf{M}_k$ is an attention map with the same size of $\mathbf{A}_k$, which indicates the importance of each vertex. $\odot$ denotes the element-wise product between two matrixes.

For the temporal dimension, the convolutional operation is the same as 2T-GCN [Bruce *et al.*, 2021b], i.e., performing a $1 \times \Gamma$ convolution on the feature map $\mathbf{f}_{out}$, where Γ is the temporal kernel size. Both the spatial and temporal graph convolutional layers are followed by a batch normalization layer and a ReLU layer. A basic GCN block is the combination of a spatial convolution layer, a temporal convolution layer, and an additional dropout layer to avoid overfitting.

Our GCN model is a stack of 9 basic GCN blocks. The first three blocks have 64 output channels. The middle three blocks have 128 output channels. And the last three blocks have 256 output channels. The temporal kernel size Γ is set to 9. To stabilize the training, the residual mechanism is applied to each GCN block. The strides of the 4th and the 7th blocks are set to 2, while all the other blocks use a stride size of 1. A global average pooling layer is added at the last GCN block to pool the GCN feature map to a 256-dimensional feature vector. The last layer of the GCN model is a 1×1 2D convolutional layer, transforming the feature vector to our desired outputs (i.e., correct or incorrect).

3.3 Ensemble-based Learning Framework

As shown in Figure 1, we use the defined GCN model to extract features from the different skeleton input streams. In the middle of the two separate learning pipelines, different fusion strategies could be performed. We use $g(P^{(n)}, \theta_g)$

and $h(O^{(n)}, \theta_h)$, where θ_g and θ_h are learnable parameters, to denote the GCN submodels for learning features from the skeleton position and orientation streams, respectively. The goal is to improve the abnormality prediction performance of our EGCN with an effective ensemble strategy. As surveyed in [Baltrušaitis *et al.*, 2019], fusion methods for ensemble learning at the data level, feature level, and decision level are commonly adopted for ensemble-based methods. Otherwise, special data fusion design needs to be performed. In the following, we introduce four groups of ensemble-based methods proposed in our EGCN.

Data-Level Ensemble. The data-level ensemble method is also known as Sample-level Ensemble (SLE). Our SLE follows [Bruce *et al.*, 2021b] that concatenated the position and orientation streams at the data level and feed them to a single GCN model, which could be represented as

$$y \ = \ \sigma(Conv(GAP(g(S^{(n)}, \theta_g)))) \tag{3}$$

where GAP is the global average pooling layer, $Conv$ is the fully connected convolutional layer, and σ is the SoftMax classifier.

Feature-Level Ensemble. We investigate two Feature-Level Ensemble (FLE) strategies. The first one, FLE-1, separately extracts features from different skeleton input streams and concatenate the extracted features at the feature level. The whole model could be optimized with an end-to-end learning process. FLE-1 can be formulated as

$$\begin{aligned} y \ = \ \sigma(Conv(GAP(Cat(g(P^{(n)}, \theta_g), \\ h(O^{(n)}, \theta_h))))) \end{aligned} \tag{4}$$

where Cat is the concatenation operation.

The second FLE, FLE-2, is based on FLE-1 by forcing the feature-level diversity of FLE-1. Since forcing the diversity of small classifiers is one of the main motivations for ensemble-based methods, we follow the Local Independence Training method [Ross *et al.*, 2020] that was implemented by penalizing the cosine similarity between the features to approximate the feature-level diversity. The loss objective of Cosine Independence Error (E_{CI}) could be formulated as

$$E_{CI}(f,\, g) \ = \ E[cos^2(g(P^{(n)}, \theta_g), h(O^{(n)}, \theta_h))] \tag{5}$$

To incorporate E_{CI} to our EGCN, we use the ensemble strategy defined in Equation 4 and optimize the E_{CI} together with the cross-entropy loss of FLE-1. The whole FLE-2 model is trained end-to-end.

Decision-Level Ensemble. Decision-Level Ensemble (DLE) could have different training strategies for optimizing the whole learning framework. We investigate two DLE strategies: DLE-1 and DLE-2. For DLE-1, we aggregate the prediction results at the decision level and train the submodels together with an end-to-end learning process, which could be written as

$$\begin{aligned} y \ = \ &\sigma(Conv(GAP(g(P^{(n)}, \theta_g))) \\ &+ Conv(GAP(h(O^{(n)}, \theta_h)))) \end{aligned} \tag{6}$$

For DLE-2, the submodels are separately trained and then their prediction results are aggregated, which could be represented as

$$\begin{aligned} y \ = \ &\sigma(Conv(GAP(g(P^{(n)}, \theta_g)))) \\ &+ \sigma(Conv(GAP(h(O^{(n)}, \theta_h)))) \end{aligned} \tag{7}$$

Model-Level Ensemble. Given the intuition that the position and orientation features respectively represent the global and local characteristics of an exercise, we utilize the neuron activation values of g that represents the position stream as the spatial and temporal importance of skeleton joints to regulate the training of the orientation stream. This is also inspired by existing model-based fusion methods [Baradel *et al.*, 2018; Si *et al.*, 2019; Bruce *et al.*, 2021a] that utilize an attention mechanism by averaging the neuron activation values along specific dimensions for the action recognition task. We do not follow this operation as it tends to smooth out the joint importance if we average the neuron activation values along the spatial or temporal dimensions of the GCN feature map. In our Model-Level Ensemble (MLE), we fuse the joint importance derived from g, which is a $C_{out} \times T_{out} \times J_{out}$ tensor, with the model representation of skeleton orientation stream $h(O^{(n)}, \theta_h)$ by element-wise multiplication along their three dimensions. The MLE model can be written as

$$y \ = \ \sigma(Conv(GAP(\ g(P^{(n)}, \theta_g) \odot h(O^{(n)}, \theta_h)\ ))) \tag{8}$$

where $g(P^{(n)}, \theta_g)$ is pretrained with the action recognition task that classifies different exercises of a whole dataset. During the training process of $h(O^{(n)}, \theta_h)$, the pre-trained parameters θ_g of $g(P^{(n)}, \theta_g)$ is fixed to maintain the mutual independence of submodels g and h.

4 Experiments

4.1 Datasets

UI-PRMD. The UI-PRMD dataset [Vakanski *et al.*, 2018] is a popular dataset for exercise assessment, which consists of skeletal data collected from 10 healthy subjects with every subject performing 10 repetitions of 10 rehabilitation exercises (E1-10) like "deep squat", "hurdle step", and "sit to stand". The subjects perform every exercise in both correct and incorrect manners. The incorrect manner is simulating the performance of patients with musculoskeletal constraints. Two sensors namely Kinect v2 and Vicon motion capture are utilized to collect the dataset. Both sensors provide positions (i.e., 3D Cartesian coordinates) and orientation features of skeleton joints. According to the results in [Bruce *et al.*, 2021b], Kienct v2 turns out to work better than Vicon motion capture. Hence, we use the data of Kinect v2 for experiments. As the dataset contains inconsistent samples caused by measurement errors and performing with incorrect limbs, we use the consistent version provided by [Liao *et al.*, 2020a], which has $1,326$ exercise repetitions in total.

KIMORE. Another rehabilitation dataset called KIMORE [Capecci *et al.*, 2019] is collected with Kinect v2 from 78 subjects that are categorized to three groups including Control Group Expert (CG-E), Control Group Non-Expert (CG-NE) and Group with Pain and Postural disorders (GPP). The GPP group has 34 subjects that have different motor dysfunctions like stroke, Parkinson's disease and back pain. All subjects perform 5 exercises (Es1-5) like "lifting of the arms", "lateral tilt of the trunk with the arms in extension", "trunk rotation", "pelvis rotations on the transverse plane" and "squatting". It is verified in [Capecci *et al.*, 2019] that the clinical total score distribution of groups CG-E and GPP are without overlapping, which means we could treat their exercise repetitions as correct and incorrect. Hence, to perform abnormality prediction based on each exercise repetition, we manually segment the datasetand group the repetitions of 17 experts and 34 patients as normal and abnormal, respectively.

4.2 Evaluation Metrics

In existing methods, Distance Metric (D_M) and Separation Degree (S_D) respectively defined in [Williams *et al.*, 2019] and [Liao *et al.*, 2020a] have been used to evaluate the representation ability of a model. On the one hand, D_M and S_D quantify the difference between the correct and incorrect evaluation results but cannot reflect the model's prediction ability. [Liao *et al.*, 2020a] attempted to show the prediction ability of their method by dividing the data into training set and validation set, but only reported the results for the E1 of UI-PRMD. On the other hand, the results of S_D in [Bruce *et al.*, 2020] reached 0.808 (derived from the training accuracy of 99.59%) for UI-PRMD by using the orientation feature. [Bruce *et al.*, 2021b] achieved an even higher S_D of 0.933 for UI-PRMD by using the SoftMax to calculate the exercise evaluation score. The main difference of [Bruce *et al.*, 2020] and [Bruce *et al.*, 2021b] is the way for calculating the exercise evaluation score, where the former used the sigmoid function while the latter used SoftMax. This observation indicates S_D can be irrelevant to a model's representation ability as it can be improved by just changing the calculation method of exercise evaluation score.

For D_M, given two positive sequences $\mathbf{x} = (x_1, \ldots, x_N)$ and $\mathbf{y} = (y_1, \ldots, y_N)$, the D_M could be calculated as

$$D_M(x_n, y_n) = \frac{|(x_n - y_n)|}{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_n - y_n)^2}} \tag{9}$$

To analyze the properness of D_M, we implement the methods in [Bruce *et al.*, 2020] and [Bruce *et al.*, 2021b]. Table 1 gives the D_M results of existing methods, which shows a similar observation with S_D. Hence, we do not continue to use the evaluation criteria D_M and S_D.

Method	E1	E7
MV [Williams *et al.*, 2019]	0.7367 (0.5383)	0.7659 (0.6012)
PCA [Williams *et al.*, 2019]	0.3777 (0.2063)	0.8161 (0.5281)
ANN [Williams *et al.*, 2019]	0.8717 (0.4330)	0.8696 (0.4246)
GCN [Bruce *et al.*, 2020]	0.9294 (0.1455)	0.8824 (0.1576)
2T-GCN [Bruce *et al.*, 2021b]	**0.9870 (0.0640)**	**0.9772 (0.1137)**

Table 1: The results of D_M (Std. deviation) for exercises "deep squat" and "standing shoulder abduction" (i.e., E1 and E7, respectively) in UI-PRMD with the orientation feature of skeleton captured by Vicon motion capture.

Given that both KIMORE and UI-PRMD are relatively small datasets, we extend the attempt of [Liao *et al.*, 2020a] by following the 5-fold cross-validation criterion applied in [Bruce *et al.*, 2021b] to test the prediction ability of different ensemble strategies in our EGCN.

4.3 Implementation Details

For the cross-validation, we split both UI-PRMD and KIMORE based on the subject ID to five folds. The proposed MLE strategy requires to pretrain the GCN model $g(P^{(n)}, \theta_g)$ with the position feature. The pretrained model could then be utilized to retrieve the joint importance from the position feature for model-level data fusion with the orientation feature. To do so, we use the position feature to pretrain a GCN model with the action classification task. In our implementation, we involve all the exercise classes of a dataset. The overall action classification accuracy for UI-PRMD and KIMORE are 96.91% and 98.04%, respectively (please refer to the supplementary for details). Like single modal methods, all the ensemble strategies proposed in EGCN are optimized with the cross-entropy loss using stochastic gradient descent with a base learning rate of 0.01. By training 50 epochs in total, we decay the learning rate by 0.1 at epochs 10 and 30. All experiments are conducted on a workstation with 2 GTX 1080 GPUs.

4.4 Comparison of Different Ensemble Strategies

Exercise ID	Single Modal		Ensemble Strategies of EGCN					
	POS	ORI	SLE	FLE-1	FLE-2	DLE-1	DLE-2	MLE
E1	71.1	73.3	67.2	64.4	72.8	73.3	71.7	**83.3**
E2	84.6	83.6	82.7	78.2	83.6	82.7	84.6	**91.8**
E3	63.7	59.8	53.9	53.9	64.7	62.8	58.8	**80.4**
E4	70.0	75.7	74.6	70.7	73.6	67.9	74.3	**79.3**
E5	86.3	79.8	87.8	73.2	83.3	86.3	82.1	**89.9**
E6	83.6	85.6	82.2	87.7	76.7	81.5	**89.0**	**89.0**
E7	80.2	87.3	89.7	72.2	81.0	68.3	88.1	**92.1**
E8	65.1	61.9	71.4	79.4	81.0	73.8	57.9	**81.8**
E9	86.7	84.2	78.3	72.5	76.7	86.7	85.8	**95.8**
E10	74.1	76.9	81.5	79.6	76.9	64.8	73.2	**85.2**
Average	76.5	76.8	76.9	73.2	77.0	74.8	76.6	**86.9**
Es1	78.0	77.7	78.0	68.2	66.3	69.8	**81.5**	79.2
Es2(L)	75.5	72.5	78.6	68.9	78.1	69.4	71.4	**81.1**
Es2(R)	71.1	80.1	80.1	70.7	74.6	75.6	77.1	**80.6**
Es3(L)	73.8	**84.8**	78.1	69.5	72.4	68.6	82.9	77.6
Es3(R)	73.7	74.6	74.6	67.5	67.0	65.1	66.5	**76.1**
Es4(L)	80.3	74.0	76.2	76.2	83.2	76.2	81.6	**84.8**
Es4(R)	83.6	79.1	79.6	78.6	78.6	79.6	83.6	**84.1**
Es5	74.1	77.7	**79.2**	72.6	74.1	64.3	78.8	77.7
Average	76.3	77.5	78.1	71.5	74.3	71.1	77.9	**80.1**

Table 2: Comparison of different ensemble strategies and single modal methods on UI-PRMD (upper table) and KIMORE (lower table). Accuracy in %. POS and ORI represent position and orientation, respectively.

Table 2 showes the correctness prediction results of all exercises in UI-PRMD and KIMORE. General ensemble methods such as SLE, FLEs and DLEs could not effectively take advantage of multiple skeleton input streams. Although ensemble strategies of SLE, FLE-2, and DLE-2 can gain some improvements when compare with single modal methods, the improvements cannot generalize well to different exercises. While FLE-1 and DLE-1 could not perform well as they just simply combine the features. In contrast, our MLE method achieves greatly better performance than both single modal and other ensemble methods for almost all exercises of UI-PRMD and KIMORE datasets, which validates the effectiveness of our model design.

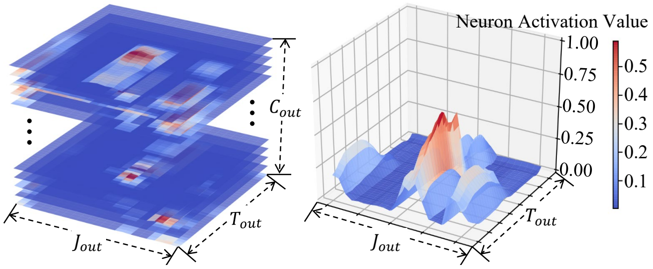

Figure 2: **Left**: visualization of joint importance layers derived from neuron activation values of the pretrained g. **Right**: visualization of mean values along the C_{out} dimension. Larger neuron activation values magnify the joint importance and vice versa.

4.5 Ablations & Comparison with State-of-the-art

Ablations. Our MLE can be implemented with different training strategies. To validate the results of MLE in Table 2, we conduct ablation study with other five different implementations of our MLE. Their results are shown in Table 3. "Self Importance" means using the joint importance derived from h to replace that from g by fixing one θ_h and updating another θ_h. "Swap h and g" means calculating the joint importance from h and updating the θ_g. "No Pretraining" means optimize h and g together without pretraining g. While "Tune θ_g" and "Fix θ_g" respectively indicate optimizing θ_h with and without updating θ_g of the pretrained g. The results consistently show that "Fix θ_g" is the effective setting for our MLE. This verifies that the learned joint importance is effective to facilitate the exercise evaluation of the orientation stream. While tuning the θ_g will affect its physical meaning (i.e., joint importance).

Besides, we also compare our model design with averaging the joint importance along the C_{out} dimension (see Figure 2[right]), which is similar with [Baradel *et al.*, 2018; Si *et al.*, 2019]. The results of "Fix θ_g, Mean Along C_{out}" in Table 3 indicate that it is not as effective as "Fix θ_g", which might be caused by the smoothing out effect of channel-level importance. We can observe this from Figure 2(left), where the joint importance layers are actually changing along the C_{out} dimension. While the averaged joint importance in Figure 2(right) is less informative than the original one.

Comparison with State-of-the-art. We first compare with state-of-the-art GCN models such as AGCN [Shi *et al.*, 2019] or MS-G3D [Liu *et al.*, 2020] via single modal setting. Table 4 shows the prediction results of these baselines implemented with single modal settings (i.e., use either position or orientation). We also use these GCN baselines as the backbone of our EGCN to further explore different ensemble strategies proposed in our EGCN on the two datasets (see Table 4). We can observe that changing the backbone with other advanced GCN baselines does not lead to stable improvements. For example, MS-G3D can improve the single modal setting of orientation feature for UI-PRMD, but it does improve the performance of other single modal settings on two datasets. This might be due to the fact that these GCN baselines are originally designed for the action recognition task rather than for exercise evaluation. With our model-level fusion design, the MLE method using the basic GCN achieves better performance than the SLE using MS-G3D on both the UI-PRMD and KIMORE datasets. This further verifies our intuition that the joint importance learned from the position stream can regularize the training of the orientation stream.

Model Implementations	Dataset	
	UI-PRMD	KIMORE
MLE (Self Importance)	67.9	76.6
MLE (Swap h and g)	67.7	77.0
MLE (No Pretraining)	76.5	67.8
MLE (Tune θ_g)	80.0	72.9
MLE (Fix θ_g, Mean Along C_{out})	77.1	77.0
MLE (Fix θ_g)	**86.9**	**80.1**

Table 3: Results of ablation studies for MLE on UI-PRMD and KIMORE (Accuracy in %).

Method	UI-PRMD (Kinect v2)			KIMORE		
	GCN	AGCN	MS-G3D	GCN	AGCN	MS-G3D
Position	76.5	65.7	76.0	76.3	72.5	74.8
Orientation	76.8	77.7	83.7	77.5	72.3	75.5
SLE	76.9	82.3	85.1	78.1	76.7	77.2
FLE-1	73.2	64.4	73.8	71.5	70.6	73.4
FLE-2	77.0	70.6	79.0	74.3	73.3	73.9
DLE-1	74.8	71.4	78.9	71.1	76.8	72.2
DLE-2	76.6	74.1	81.1	77.9	69.7	74.4
MLE	**86.9**	71.3	84.3	**80.1**	73.3	75.2

Table 4: Average prediction results implemented with state-of-the-art GCN models (see rows of Position and Orientation) and different ensemble methods of our EGCN implemented with different backbones, i.e., GCN, AGCN, and MS-G3D (Accuracy in %).

5 Conclusion

In this paper, we propose the EGCN framework with various ensemble strategies for the exploration of effective skeleton-based exercise assessment. This is the first work that uses both position and orientation skeleton features for the task. With extensive experiments on two latest datasets: UI-PRMD and KIMORE, the proposed MLE outperforms other ensemble strategies in terms of the prediction accuracy. Meanwhile, we further verify the proper design of our MLE with ablations of different training strategies and other baselines. In the future, we will investigate if our EGCN can generate evaluation scores that are consistent with human evaluation and design online exercise evaluation methods.

Acknowledgements

This work is supported by The Hong Kong Polytechnic University for the project: Artificial Intelligence and Robotics (AIR) Group: Artifical Intelligence Art under grant No.: P0033738.

References

[Ahad *et al.*, 2019] Md Atiqur Rahman Ahad, Anindya Das Antar, and Omar Shahid. Vision-based action understanding for assistive healthcare: A short review. In *Proc. CVPR Workshops*, pages 1–11, 2019.

[Baltrušaitis *et al.*, 2019] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Patt. Analy. and Mach. Intell.*, 41(2):423–443, 2019.

[Baradel *et al.*, 2018] Fabien Baradel, Christian Wolf, Julien Mille, and Graham W Taylor. Glimpse clouds: Human activity recognition from unstructured feature points. In *Proc. CVPR*, pages 469–478, 2018.

[Bassett and Prapavessis, 2007] Sandra F Bassett and Harry Prapavessis. Home-based physical therapy intervention with adherence-enhancing strategies versus clinic-based management for patients with ankle sprains. *Physical Therapy*, 87(9):1132–1143, 2007.

[Bruce *et al.*, 2020] XB Bruce, Yan Liu, and Keith CC Chan. Skeleton-based detection of abnormalities in human actions using graph convolutional networks. In *Proc. TransAI*, pages 131–137. IEEE, 2020.

[Bruce *et al.*, 2021a] XB Bruce, Yan Liu, and Keith CC Chan. Multimodal fusion via teacher-student network for indoor action recognition. In *Proc. AAAI*, pages 3199–3207, 2021.

[Bruce *et al.*, 2021b] XB Bruce, Yan Liu, Keith CC Chan, Qintai Yang, and Xiaoying Wang. Skeleton-based human action evaluation using graph convolutional network for monitoring alzheimer's progression. *Patt. Recog.*, page 108095, 2021.

[Capecci *et al.*, 2019] Marianna Capecci, Maria Gabriella Ceravolo, Francesco Ferracuti, Sabrina Iarlori, Andrea Monteriù, Luca Romeo, and Federica Verdini. The kimore dataset: Kinematic assessment of movement and clinical scores for remote monitoring of physical rehabilitation. *IEEE Transactions on Neur. Syst. and Rehabil. Engi.*, 27(7):1436–1448, 2019.

[Dong *et al.*, 2020] Xibin Dong, Zhiwen Yu, Wenming Cao, Yifan Shi, and Qianli Ma. A survey on ensemble learning. *Frontiers of Computer Science*, pages 1–18, 2020.

[Du and Swamy, 2019] Ke-Lin Du and MNS Swamy. Combining multiple learners: Data fusion and ensemble learning. In *Neural Netw. and Stati. Lear.*, pages 737–767. Springer, 2019.

[Elkholy *et al.*, 2019] Amr Elkholy, Mohamed Hussein, Walid Gomaa, Dima Damen, and Emmanuel Saba. Efficient and robust skeleton-based quality assessment and abnormality detection in human action performance. *IEEE J. of biom. and heal. inform.*, 2019.

[Jessep *et al.*, 2009] Sally A Jessep, Nicola E Walsh, Julie Ratcliffe, and Michael V Hurley. Long-term clinical benefits and costs of an integrated rehabilitation programme compared with outpatient physiotherapy for chronic knee pain. *Physiotherapy*, 95(2):94–102, 2009.

[Karmali *et al.*, 2014] Kunal N Karmali, Philippa Davies, Fiona Taylor, Andrew Beswick, Nicole Martin, and Shah Ebrahim. Promoting patient uptake and adherence in cardiac rehabilitation. *Cochrane database of systematic reviews*, (6), 2014.

[Komatireddy *et al.*, 2014] Ravi Komatireddy, Anang Chokshi, Jeanna Basnett, Michael Casale, Daniel Goble, and Tiffany Shubert. Quality and quantity of rehabilitation exercises delivered by a 3-d motion controlled camera: A pilot study. *International J. of physi. medic. & rehabi.*, 2(4), 2014.

[Lei *et al.*, 2019] Qing Lei, Ji-Xiang Du, Hong-Bo Zhang, Shuang Ye, and Duan-Sheng Chen. A survey of vision-based human action evaluation methods. *Sensors*, 19(19):4129, 2019.

[Liao *et al.*, 2020a] Yalin Liao, Aleksandar Vakanski, and Min Xian. A deep learning framework for assessing physical rehabilitation exercises. *IEEE Trans. on Neur. Syst. and Rehabi. Engi.*, 28(2):468–477, 2020.

[Liao *et al.*, 2020b] Yalin Liao, Aleksandar Vakanski, Min Xian, David Paul, and Russell Baker. A review of computational approaches for evaluation of rehabilitation exercises. *Computers in bio. and medi.*, 119:103687, 2020.

[Liu *et al.*, 2020] Ziyu Liu, Hongwen Zhang, Zhenghao Chen, Zhiyong Wang, and Wanli Ouyang. Disentangling and unifying graph convolutions for skeleton-based action recognition. In *Proc. CVPR*, pages 143–152, 2020.

[Machlin *et al.*, 2011] Steven R Machlin, Julia Chevan, William W Yu, and Marc W Zodet. Determinants of utilization and expenditures for episodes of ambulatory physical therapy among adults. *Physical therapy*, 91(7):1018–1029, 2011.

[Ross *et al.*, 2020] Andrew Slavin Ross, Weiwei Pan, Leo A Celi, and Finale Doshi-Velez. Ensembles of locally independent prediction models. In *Proc. AAAI*, pages 5527–5536, 2020.

[Sagi and Rokach, 2018] Omer Sagi and Lior Rokach. Ensemble learning: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4):e1249, 2018.

[Saraee *et al.*, 2017] Elham Saraee, Saurabh Singh, Kathryn Hendron, Mingxin Zheng, Ajjen Joshi, Terry Ellis, and Margrit Betke. Exercisecheck: remote monitoring and evaluation platform for home based physical therapy. In *Proc. PETRAE*, pages 87–90, 2017.

[Sebbag *et al.*, 2019] Eden Sebbag, Renaud Felten, Flora Sagez, Jean Sibilia, Hervé Devilliers, and Laurent Arnaud. The worldwide burden of musculoskeletal diseases: a systematic analysis of the world health organization burden of diseases database. *Annals of the rheumatic diseases*, 78(6):844–848, 2019.

[Shi *et al.*, 2019] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In *Proc. CVPR*, pages 12026–12035, 2019.

[Si *et al.*, 2019] Chenyang Si, Wentao Chen, Wei Wang, Liang Wang, and Tieniu Tan. An attention enhanced graph convolutional lstm network for skeleton-based action recognition. In *Proc. CVPR*, pages 1227–1236, 2019.

[Tao *et al.*, 2016] Lili Tao, Adeline Paiement, Dima Damen, Majid Mirmehdi, Sion Hannuna, Massimo Camplani, Tilo Burghardt, and Ian Craddock. A comparative study of pose representation and dynamics modelling for online motion quality assessment. *Computer vision and image understanding*, 148:136–152, 2016.

[Vakanski *et al.*, 2018] Aleksandar Vakanski, Hyung-pil Jun, David Paul, and Russell Baker. A data set of human body movements for physical rehabilitation exercises. *Data*, 3(1):2, 2018.

[Williams *et al.*, 2019] Christian Williams, Aleksandar Vakanski, Stephen Lee, and David Paul. Assessment of physical rehabilitation movements through dimensionality reduction and statistical modeling. *Medical engineering & physics*, 74:13–22, 2019.

[Wyss and Patel, 2012] James Wyss and Amrish Patel. *Therapeutic programs for musculoskeletal disorders*. Demos Medical Publishing, 2012.