

修士論文概要書

Master's Thesis Summary

Date of submission: 01/22/2024 (MM/DD/YYYY)

専攻名 (専門野) Department	Computer Science and Communications Engineering	氏名 Name	Smita Priyadarshani	指導 教員 Advisor	Hiroshi Watanabe 印 Seal
研究指導名 Research guidance	Research on Audiovisual Information Processing	学籍番号 Student ID number	5122FG16		
研究題目 Title	Attention-enhanced Spatial-Temporal Graph Convolutional Network for Assessing Rehabilitation Exercises				

1. Introduction

Physical rehabilitation exercises are crucial in the post-operative recovery and treatment of various musculoskeletal conditions. Usually, rehabilitation exercises involve the movement of more than one body part that is far apart. Therefore, for an accurate assessment, it is crucial to understand the joint-specific roles and correlation between these non-local features. However, most methods treat all body joints equally and fail to capture the correlation information between all joints.

Hence, to address this limitation, we propose an attention-enhanced spatial-temporal graph convolutional network that captures the spatial and temporal dependencies among the body joints, as illustrated in Figure 1. Our model consists of mainly 3 blocks; the spatial layer which performs spatial convolution of graph data to extract spatial features, the attention layer calculates the weighted sum of these features at all positions and, the temporal layer performs convolution in the temporal domain.

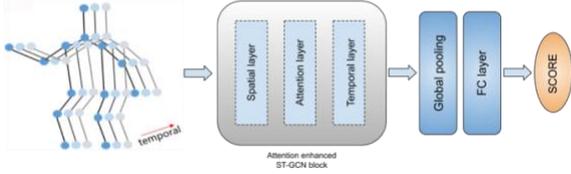


Figure 1. An Overview of the proposed model.

2. Related Work

The Spatial-Temporal Graph Convolutional Network (ST-GCN) that we use for our architecture was originally introduced by Yan et al. [1] in 2018. It operates on graph-structured data, where nodes represent human joints and edges define their spatial-temporal connections. ST-GCN performs convolution operations in both spatial and temporal dimensions, effectively capturing the dynamics of human movements over time. The key equation governing its operation is,

$$f_{out}(v_i) = \sum_{v_j \in B(v_i)} \frac{1}{Z_{ij}} f_{in}(v_j) \times w(l_{ij}), \quad (1)$$

where (f_{out}) and (f_{in}) are the output and input features,

$(B(v_i))$ denotes the neighbour set of joints (v_i) , (Z_{ij}) is a normalization factor, (w) represents the learnable weight, and (l_{ij}) is the predefined spatial configuration label.

In 2019, Wang et al. [2] proposed a non-local neural network that captures the long-range dependencies by computing response at a position as a weighted sum of the features at all positions. The formulation for non-local operation can be written as,

$$y_i = (1/C(x)) \sum (\forall j) f(x_i, x_j) g(x_j), \quad (2)$$

where (y_i) represents the output at position (i) , (x_i) and (x_j) are the positions in the input signal, (f) is a pairwise function measuring the relationship between (i) and (j) , (g) is a unary function representing the input signal at position (j) , and $(C(x))$ is a normalization factor

3. Proposed Method

Our proposed attention-enhanced ST-GCN model is illustrated in Figure 2, which captures the spatial and temporal dependencies among the body joints incorporating an attention mechanism [2] to learn global information for the joint-specific roles to provide better assessment results. Our objective is to provide a point-in-time assessment of the patient's quality of motion linked to one or more body parts.

This input is fed into the ST-GCN [1] which combines the non-local attention module adapted from [2] between the spatial and temporal convolution layers. Afterwards, we employ an attention module as shown in Figure 2 (right), which computes 1×1 convolution and generates three sets of features: θ , φ , and g . Next, an element-wise multiplication of θ and φ is calculated, which allows us to determine the feature auto-correlation. The updated graph convolution function can be expressed using the equation,

$$f_{out} = \sum_j \left(\Lambda^{-\frac{1}{2}} (A + I) \Lambda^{-\frac{1}{2}} \otimes M_j \right) f_{in} W_j, \quad (3)$$

where the output feature map (f_{out}) is generated by applying a Hadamard product of a normalized adjacency matrix and a trainable self-attention map (M) to the input feature map (f_{in}) , then convolving with node-specific weight vectors (W_j) , focusing on the relevant inter-node relationships in a spatiotemporal graph structure.

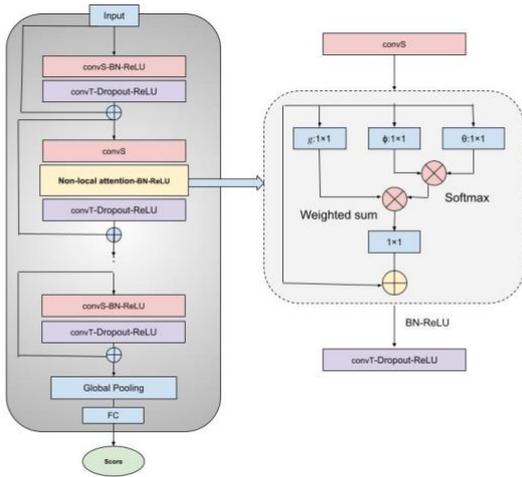


Figure 2. The proposed architecture of attention-enhanced spatial-temporal graph convolutional network (left). Non-local attention block(right).

4. Experiment

For our experiment, we use two publicly available datasets; the UI-PRMD consisting of 10 subjects performing 10 exercises and the KIMORE with 78 subjects performing 5 exercises. We stack such 9 layers of spatial-temporal blocks together and incorporate the attention layer in the second and third blocks after spatial convolution. We implemented the proposed model using the pyTorch framework and two Nvidia RTX 3060 GPUs, each with 12 GB RAM. We train our model using Adam optimizer for 1500 epochs with a 0.0004 learning rate. For the Huber loss function, after experimenting with different values, we set the value of (δ) to 0.1. In Table 1, we report the results for the average of Mean Absolute Deviation (MAD) across all the exercises from the UI-PRMD and KIMORE dataset and compare it with Liao et al [3]. Lower scores represent better performance. Hence, our model has slightly better performance for the UI-PRMD dataset, but a considerable improvement of 0.182 for the KIMORE dataset.

Table 1. Comparison of mean absolute deviation (MAD) scores method on the UI-PRMD and KIMORE datasets.

Dataset	Ours	Liao et al. [3]
UI-PRMD	0.020	0.025
KIMORE	0.7840	0.966

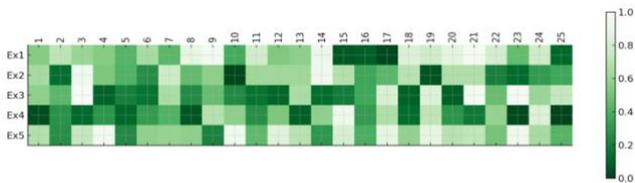


Figure 3. A visualization of how the role of joints varies with different exercises as determined by the attention value computed by our method for the KIMORE dataset.

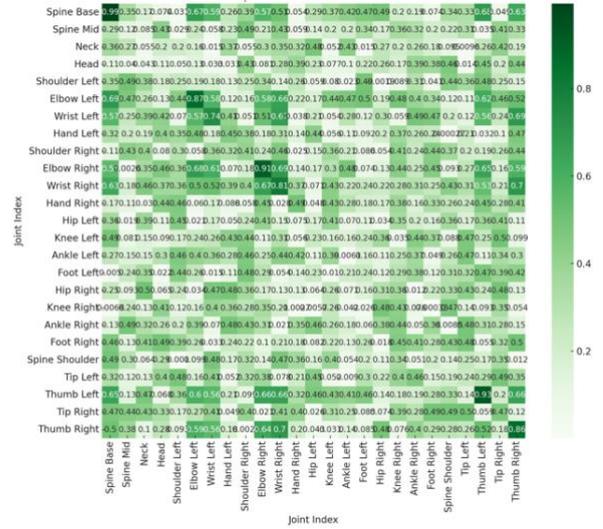


Figure 4. A visualization of how the role of joints varies with exercise 1 (lifting arms) as determined by the attention value computed by our method for the KIMORE dataset.

Figure 3 shows the obtained correlation map for exercise 1 in the KIMORE dataset is visualized. Exercise 1 is a lifting arm that involves four key joints; elbow, spine, thumb, and wrist. In the attention map, these key joints are in darker shades, which represents a stronger correlation between these joints for a given exercise sequence. Whereas, the remaining joints are in lighter shades which represent less important interactions.

5. Conclusion

Our model captures both spatial and temporal dependencies among body joints and applies an attention mechanism to understand the roles and interrelations of these joints more effectively. The network architecture, based on the spatial-temporal graph convolutional network (ST-GCN), is adapted to include a non-local attention module that enables the model to consider global information and relationships between non-adjacent joints, essential for a nuanced assessment of rehabilitation exercises.

References

- [1] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," AAAI Conference on Artificial Intelligence, No.912, pp.7444-7452, Feb. 2018.
- [2] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp.7794-7803, Jun. 2018.
- [3] Y. Liao, A. Vakanski, and M. Xian, "A deep learning framework for assessing physical rehabilitation exercises," IEEE Transactions on Neural Systems and Rehabilitation Engineering, Vol. 28, No.2, pp.468-477, Feb.2020.

Attention-enhanced Spatial-Temporal Graph Convolutional Network for Assessing Rehabilitation Exercises

A Thesis Submitted to the Department of Computer Science and Communications Engineering, the Graduate School of Fundamental Science and Engineering of Waseda University in Partial Fulfillment of the Requirements for the Degree of Master of Engineering

January 22nd, 2024

Smita Priyadarshani
(5122FG16-5)

Advisor: Prof. Hiroshi Watanabe
Research guidance: Research on Audiovisual Information Processing

Acknowledgement

I would like to express my deepest gratitude to my advisor, Prof. Hiroshi Watanabe, whose expertise, understanding, and patience added considerably to my graduate experience. I am immensely thankful for the resources he provided to conduct my experiments and for the insightful feedback that encouraged me to expand my research from multiple perspectives.

I must express my very profound gratitude to my parents for providing me with unwavering support and continuous encouragement throughout my years of study. Thank you for instilling in me a work ethic that has served me well throughout this process. My sincere thanks also go to my colleagues in the laboratory. Their assistance when I encountered problems and their stimulating discussions have been instrumental in broadening my knowledge and enriching my growth as an academic.

Finally, I would like to acknowledge my friends for making my study abroad a memorable and fun journey. Living and studying in Tokyo has been a unique and delightful experience. Their presence and support helped me to stay balanced and focused during my academic pursuits.

Abstract

Physical rehabilitation exercises are crucial in the post-operative recovery and treatment of various musculoskeletal conditions. An automated vision-based rehabilitation exercise assessment is a portable and cost-effective way of evaluating the patients' performance in a home-based setting and predicting a performance score by analyzing the correct and incorrect exercise sequences performed by the patient. Recent works have shown that exploring spatial and temporal features of the skeleton data is vital for this task. Usually, rehabilitation exercises involve the movement of more than one body parts that are far apart. Therefore, for an accurate assessment, it is crucial for the model to understand the joint-specific roles and correlation between these non-local features. However, most methods treat all body joints equally and fail to capture the correlation information between all joints.

Hence, to address this limitation, we have proposed an attention-enhanced spatial-temporal graph convolutional network that captures the spatial and temporal dependencies among the body joints. This is achieved by incorporating an attention mechanism to learn global information for the joint-specific roles, thereby providing better assessment. It does so by computing the response at a position as the weighted sum of features at all positions, regardless of their positional distance.

Keywords: Graph Convolutional Network, Rehabilitation Exercises, Attention Mechanism, Movement Assessment, Temporal Dynamics

Contents

Abstract	3
List of Figures	6
List of Tables	7
1 INTRODUCTION	1
1.1 Background	2
1.1.1 Why Vision-based Techniques Are Needed	2
1.1.2 Existing Methods for Vision-based Assessment	3
1.2 Research Motivation and Objective	4
1.3 Thesis Outline	5
2 RELATED WORK	7
2.1 Discrete Movement Score Approaches	7
2.2 Rule-based Approaches	8
2.3 Template-based Approaches	8
2.4 Deep Learning-based Approaches	9
3 PROPOSED METHOD	10
3.1 An Overview of the Approach	10
3.2 Problem Formulation	12
3.3 Attention-enhanced Spatial-Temporal Graph Convolutional Network	12
3.3.1 Graph Construction	13
3.3.2 Partitioning Strategy	15
3.3.3 Spatial and Temporal Modeling	15
3.3.4 Attention Mechanism	16
3.4 Network Architecture	17
3.5 Training Loss	19

4	EXPERIMENTAL RESULTS	20
4.1	Experimental Setup	20
4.1.1	Dataset	20
4.1.2	Implementation Details	21
4.1.3	Evaluation Metrics	21
4.2	Ablation Study	22
4.2.1	Effect of Attention Block Incorporated with the ST-GCN	22
4.2.2	Effect of Position of Attention Block	22
4.2.3	Effect of Adding Multiple Attention Blocks	23
4.3	Results	23
4.3.1	Quantitative Results of Rehabilitation Exercise Assessment on the UI-PRMD and KIMORE dataset	23
4.3.2	Feature Visualization	25
5	CONCLUSION	27
5.1	Limitations	27
5.2	Application for Real-time Assessment and Feedback	28
5.3	Future Work	28

List of Figures

1.1	General overview of vision-based physically impaired patient assessment system and techniques.	3
1.2	A flowchart of machine learning-based approach for assessment and feedback process.	5
3.1	An overview of the proposed methodology for the rehabilitation exercise assessment.	10
3.2	The architecture of the proposed attention-enhanced spatial-temporal graph convolutional network (left). Non-local attention block (right).	13
3.3	The spatial and temporal graph of a skeleton sequence proposed by [1]. Yellow dots denote the body joints and edges are connected in inter-frame and intra-body fashion.	14
3.4	An illustration of the partitioning strategy adopted in ST-GCN architecture.	15
3.5	An illustration of the non-local neural network proposed by [2] used as the attention block.	18
4.1	A visualization of the attention map for exercise 1 (lifting arms) in the KIMORE dataset. Darker shades represent a stronger correlation and vice-versa.	25
4.2	A visualization of how the role of joints varies with different exercises as determined by the attention value computed by our method for the KIMORE dataset. Darker shades represent stronger correlation and vice-versa.	26

List of Tables

4.1	Summary table of KIMORE and UI-PRMD dataset.	21
4.2	Comparison of our proposed model with and without Attention block on KIMORE dataset.	22
4.3	Comparison table of the position of one attention block in different spatial convolution (convS) layers of the ST-GCN model.	23
4.4	Comparison table of adding multiple attention blocks in the second and third spatial layers of the ST-GCN.	23
4.5	Comparison of mean absolute deviation (MAD) scores of existing methods and our proposed method on the UI-PRMD dataset. A lower value indicates better result.	24
4.6	Comparison of mean absolute deviation (MAD), Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE) scores of the model by Liao et al. [3] and our proposed method on the KIMORE dataset.	24

Chapter 1

INTRODUCTION

Physical therapy also known as physiotherapy is a treatment for people suffering from certain types of diseases such as Parkinson’s Disease, post-stroke, or injuries that cause loss of natural limb mobility. The treatment requires expert guidance for effective outcomes. However, frequent visits to the therapist for each session make it cost-intensive and infeasible for patients suffering from severe musculoskeletal impairment. Hence, an automated computer-aided assessment of rehabilitation exercises is a solution to provide a virtual analysis in the home-based setting. It involves using sensor devices such as an RGB-D camera to record the body movements and provide an assessment score based on the evaluation of the data sequence.

In computer vision, several skeleton-based methods proposed for recognition tasks have proven to be one of the most effective ways for pose classification. However, not enough work has been done in the action evaluation domain on the skeleton data making it a potential research interest. Most of the work addressing the assessment task fails to capture the patterns embedded in the spatial configuration of the joints as well as their temporal dynamics. Hence, to address this issue we have utilized the spatial-temporal graph convolutional network (ST-GCN) [1] originally proposed for skeleton-based action recognition that effectively learns the spatial and temporal patterns from the data. However, ST-GCN can capture only local features since the receptive field of the convolution operation is the neighbour of the root node. Thus, we introduce a non-local attention module proposed by Wang et al. [2] into the network to capture long-range dependencies directly by computing interactions between any two positions, regardless of their positional distances.

1.1 Background

1.1.1 Why Vision-based Techniques Are Needed

Rehabilitation is a critical process in healthcare, aiming to restore or improve functional ability and quality of life in individuals who have experienced impairments due to injury, illness, or a chronic health condition. It encompasses a wide range of therapies, including physical, occupational, and speech therapy, tailored to individual patient needs. The goal of rehabilitation is multi-fold: to enhance physical functionality, promote patient independence, and improve overall well-being.

Physical rehabilitation is widely recommended for individuals with physical disabilities or those needing to regain functional abilities following injuries or surgeries. It has been established as crucial for enhancing patient outcomes, with a noted connection between the intensity of exercises in rehabilitation programs and their effectiveness. Despite its benefits, rehabilitation can be costly, both for patients and healthcare systems. In 2007, the United States witnessed an expenditure of about 13.5 billion dollars in this sector, covering a vast number of rehabilitation episodes [4].

Rehabilitation typically involves clinicians guiding and monitoring patients' exercise routines in clinical settings. Due to limitations in clinician availability and the demands on patients' time, home-based rehabilitation programs have become more common. These programs are designed by clinicians and are followed by patients at home, with periodic clinic visits for progress checks. However, adherence to these home-based exercises is often low, attributed to factors like the absence of real-time supervision and feedback from healthcare professionals. This lack of motivation and proper guidance can lead to improperly performed exercises, heightening the risk of re-injury.

Assessments in physical rehabilitation, which rely heavily on visual reporting and clinical judgment, can be inconsistent and subjective. This often leads to inaccuracies in evaluating patient progress. Conversely, assessments integrating kinematic parameters tend to be more robust and accurate. Traditional methods using body-worn sensors or marker-based systems are not only costly but also intrusive to daily activities. On the other hand, marker-less vision-based human motion modelling offers a promising alternative. This approach could provide cost-effective, non-intrusive monitoring suitable for home-based rehabilitation, harnessing the potential to revolutionize patient assessment and monitoring.

1.1.2 Existing Methods for Vision-based Assessment

The research process in vision-based rehabilitation and monitoring is depicted in Figure 1.1 which breaks down the process into key steps. It starts with using a vision-based sensor, like a monocular RGB or depth camera, to collect data. This is followed by extracting low-level features such as human joint positions. The next step involves encoding and representing these features, possibly through grouping joint positions or using human kinematic parameters. These features are then analyzed using either basic graphical/statistical methods or more advanced algorithms. The final stage involves various assessments, broadly divided into three types: kinematic parameter comparisons, impairment classification, and impairment clinical scoring. The first, Comparison, involves extracting patient data like kinematic parameters for analysis but lacks an automated scoring system. This might include statistical methods like ANOVA or visual comparisons of trajectories between ideal and actual patient movements. The second type, Categorization, is more definitive, focusing on classifying movements as correct or incorrect or identifying specific types of movement abnormalities. This encompasses both gesture/posture and activity recognition. The third type, Scoring, assigns a definitive score to patient movements, evaluating their quality or quantifying deviations from ideal motion. This can involve clinical scores like the Fugl-Meyer Assessment (FMA) [5], the Unified PD Rating Scale (UPDRS) [6], or scores proposed by researchers. The score may be for assessing the quality of movement or quantifying the differences from an ideal motion.

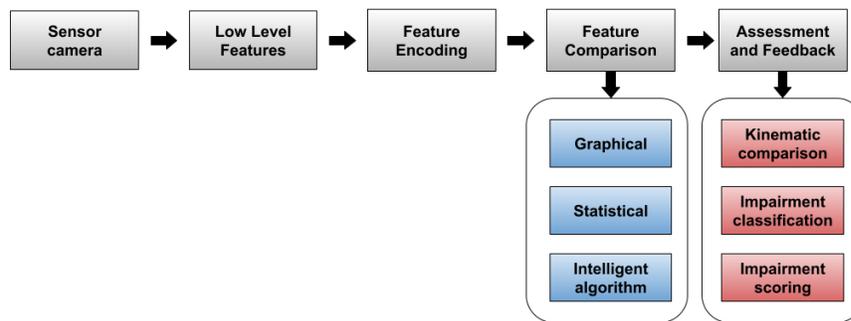


Figure 1.1: General overview of vision-based physically impaired patient assessment system and techniques.

The primary aim of monitoring musculoskeletal patients is to develop an automated system for evaluating their progress. Clinicians typically assess physical rehabilitation by looking at various physical attributes, like the range of elbow flexion, shoulder movement, and speed of motion. Research in this field has predominantly utilized the estimation of joint po-

sitions in the human body as the key low-level features for such assessments. An important aspect of automated evaluation involves comparing a patient’s performance in daily activities or rehabilitation exercises against those of a healthy individual, focusing specifically on the sequence of joint positions.

1.2 Research Motivation and Objective

In the field of detection and activity recognition, objectives like classification are clear-cut. However, when it comes to patient assessment, the goal can vary significantly based on clinical needs. These range from statistical analyses to the automatic generation of clinical scores. Most studies in this field have relied on manually selecting key movement features or creating new features from motion capture data. Traditional feature engineering techniques like PCA [7] and manifold learning [8] are used to represent exercise-related features, but they often come with specific prerequisites. These methods have their limitations, as they need specialized knowledge in motion kinematics for feature extraction, and the features derived are not adaptable for new exercises. Consequently, there’s a need to develop methods for feature analysis based solely on joint positions.

Deep learning methods, on the other hand, are increasingly popular across various domains for their ability to encode feature representations without requiring specialized domain knowledge. Deep learning models are capable of capturing complex relationships between observed and hidden variables and can represent data features at multiple levels of abstraction, making them suitable for motion modelling and analysis. For instance, Vakanski et al. [3] is the first to use a CNN architecture for feature extraction. However, because of the use of CNN, it ignores topological structure information from interaction among neighbourhood joints. Other research has explored various neural network architectures for learning spatial and temporal features from movement data in the action recognition domain. Song et al. [9] introduced a deep network incorporating spatial and temporal attention sub-networks, with spatial attention focusing on key joints and temporal attention emphasizing important time frames.

Despite the extensive research in deep learning for motion modelling, such as in recognition and classification, there is still limited research focused on evaluating movement in rehabilitation exercises. In the field of action recognition, Graph Convolutional Networks (GCNs) have proven effective, particularly with models influenced by Spatio Temporal-GCN. Motivated by the achievements of GCN-based methods and their ability to preserve the topological structure of skeletal data, we have proposed an attention-enhanced spatial-temporal graph convolutional network that captures the spatial and temporal dependencies among the

body joints incorporating an attention mechanism to learn global information for the joint-specific roles to provide better assessment results. By integrating graph-based methods with attention, this strategy is poised to deliver precise and visually comprehensible feedback for the evaluation of rehabilitation exercises. A high-level overview of a computational approach to rehab assessment is depicted in Figure 1.2.

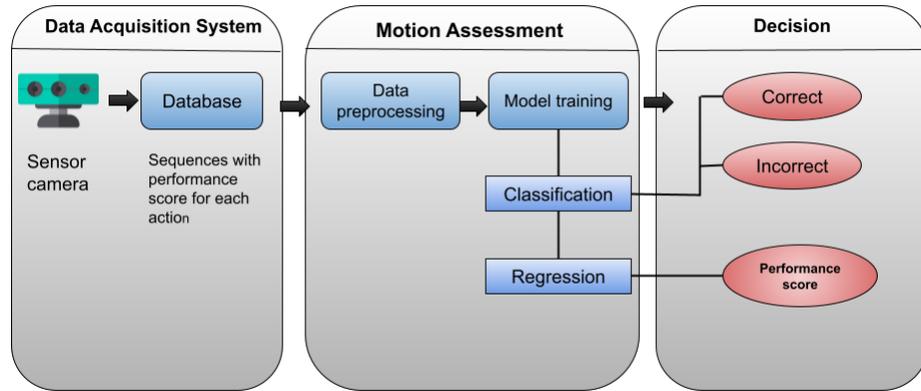


Figure 1.2: A flowchart of machine learning-based approach for assessment and feedback process.

Our objective is to provide a point-in-time assessment of the patient’s quality of motion linked to one or more body parts. For this, we first extract spatial as well as temporal features as a unified network and further, assign varying weights to the nodes using the attention mechanism to learn the relationship between the non-adjacent joints. This is important because, for different movements, interaction between different non-neighbouring nodes could be crucial. Paying more attention to these nodes may improve assessment accuracy.

1.3 Thesis Outline

The outline of the thesis is as follows: Chapter 1: We provide an introduction to traditional rehabilitation exercise assessment approaches and explain why vision-based techniques are needed. Further, we provide the background of existing computational methods and their drawbacks. Then we briefly describe the motivation for this research and its objectives.

Chapter 2: In this chapter, we provide an overview of the existing computational approaches for rehabilitation assessment. We start with older techniques such as discrete movement score, rule-based and template-based approach. We discuss their drawbacks and then move on to new and more robust deep learning-based techniques.

Chapter 3: In this chapter, we discuss our proposed method, problem formulation, attention block and network architecture.

Chapter 4: In this chapter, we provide the details on the experimental setup, the dataset used, implementation details, various ablation studies and obtained results.

Chapter 5: In this concluding chapter, we discuss the effectiveness of our method, its limitations, its usability for real-time applications, and future work.

Chapter 2

RELATED WORK

2.1 Discrete Movement Score Approaches

Discrete movement score approaches in rehabilitation exercises focus on categorizing each exercise repetition into distinct classes. Typically, these classifications are binary, delineating correct and incorrect movements. Consequently, the outcome is usually a binary class value, indicating either a correct (1) or incorrect (0) repetition.

Several machine learning techniques have been employed to differentiate these classes. Techniques such as the Adaboost classifier, k-nearest neighbours, Bayesian classifier, and an ensemble of multi-layer perceptron neural networks have proven effective. For example, in one study [10], a k-nearest neighbors classifier was used to classify exercises. This process involved filtering out noise from the data and reducing dimensionality through Principal Component Analysis (PCA), achieving a classification accuracy of 95.6.

Machine learning classifiers have also been extensively used in movement classification within standard clinical tests. This includes approaches like Support Vector Machines (SVM), random forest, and artificial neural networks. These methods have been applied for automated Functional Movement Analysis (FMA), with naïve Bayes classifiers implemented for the Wolf Motor Function Test (WMFT) and random forest for Functional Ability Scale (FAS) evaluations. All these techniques are mentioned in detail in a review work by [4].

In a notable study by Um et al. [11], an ensemble of convolutional neural networks was utilized for detecting states of Parkinson's disease using data from a wrist-worn wearable sensor. This method could distinguish between different states such as OFF (with Parkinson's syndrome symptoms), DYS (with dyskinesic symptoms), and ON (no salient symptoms).

Despite the high accuracy reported in distinguishing between correct and incorrect movement sequences, a common limitation of these discrete movement score approaches is their inability to track continuous changes in movement quality or to quantify the progression of a

patient's performance throughout the rehabilitation program.

2.2 Rule-based Approaches

Rule-based methods for evaluating rehabilitation exercises utilize established rules and criteria to measure and analyze a patient's performance. These methods are grounded in clinical standards and best practices, with the rules being specifically adapted to the patient's individual needs and condition. For instance, the Functional Independence Measure (FIM) is used for neurological impairments, while the Knee Injury and Osteoarthritis Outcome Score (KOOS) is applied for knee-related disorders [12]. These methods involve standardized procedures to evaluate patient movements, such as tracking the angles of knees and ankles [13] or establishing specific kinematic guidelines [14]. Nonetheless, these approaches can be time-intensive and might not always align with the unique goals of individual patients. While these methods are effective for straightforward exercises, their utility tends to decrease with the complexity of the exercise. Furthermore, they are not particularly flexible in accommodating new types of exercises.

2.3 Template-based Approaches

Template-based approaches in rehabilitation are centred around the comparison of measured patient movements with a standard template. Typically, this template represents the correct execution of exercises by healthy individuals. The core of these approaches lies in assessing the resemblance between the patient's performed movements and the template, facilitating a quantitative evaluation of the patient's performance.

One prominent method within these approaches involves the utilization of distance functions [15]. These functions—Euclidean, Mahalanobis, and Dynamic Time Warping being the most prevalent—are employed to compute a similarity score. This score quantitatively expresses how closely a patient's movement mirrors the reference template. Further, Gaussian Mixture Models [16] and Hidden Markov Models [17] have been instrumental in motion modelling. In these models, the quality of a movement is determined based on the probability of the observed movement sequence originating from a predefined model. This probabilistic approach is particularly beneficial in accommodating the inherent stochastic nature of human movements.

2.4 Deep Learning-based Approaches

Deep learning-based approaches are increasingly popular in various fields for their ability to automatically learn feature representations without necessitating specific domain expertise. These models are adept at understanding complex, nonlinear relationships between observed and hidden variables, and their capacity to process data features across multiple levels of abstraction makes them highly suitable for motion analysis and modeling. Liao et al. [3] proposed a log-likelihood-based performance metric to train their DL framework for the assessment of rehabilitation exercises. They trained and compared the performances of CNN, RNN and Hierarchical Neural Network (HNN), where the log-likelihood-based performance metric was used as a label to regress the network for predicting deviations from normal actions.

Since DeepPose's introduction in 2014, CNN-based techniques for human pose estimation have seen remarkable advancements in accuracy. A notable example is the work by Li et al. [18], who utilized Convolutional Pose Machines (CPM) to extract skeleton data for studying Levodopa-Induced Dyskinesia (LID) in Parkinson's Disease (PD). In one study, a deep recurrent network was specifically developed to learn co-occurrence features from skeletal data, implementing an innovative regularization technique [19]. Song et al. [9] introduced a deep learning model that incorporated spatial and temporal attention mechanisms, enhancing the model's focus on key joints and critical time frames.

While there is extensive research in the domain of deep learning for motion analysis, such as recognition, classification, etc., its application in the evaluation of movement in rehabilitation exercises remains relatively under-explored. As mentioned earlier, skeleton-based approaches became more feasible with the advent of Kinect. These approaches utilize the detailed skeletal data provided by Kinect for tracking and assessing movements, offering a more accurate and comprehensive analysis of physical activities.

Chapter 3

PROPOSED METHOD

3.1 An Overview of the Approach

In this section, we will introduce our proposed method and its overall architecture. An outline of our proposed method is illustrated in Figure 3.1.

The pipeline for our proposed method can broadly be divided into three sections. Firstly, the data is collected using acquisition devices, such as a Kinect or a RGB-D camera. the acquired data consists of skeleton sequences of the patients performing exercises. Further, the joint coordinates are extracted from these sequences using skeleton extraction algorithms. Lastly, the graph is constructed in the spatial and temporal domain using the joint coordinates obtained after skeleton extraction.

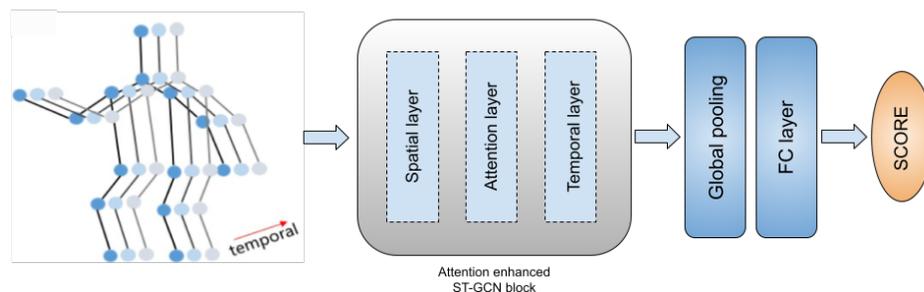


Figure 3.1: An overview of the proposed methodology for the rehabilitation exercise assessment.

For our experiment, we have considered the extracted joint coordinates. We have constructed the graph as proposed by [1]. We have provided a detailed information on the graph

construction method in section 3.3.1. After the graph construction, the next step is feature extraction. The constructed graph is fed into our proposed attention-enhanced spatial-temporal graph convolutional network to extract the spatial and temporal features along with joint-specific contribution. The block consists of three layers: spatial layer, attention layer and temporal layer.

Spatial layer: The spatial layer performs convolution operations in the spatial domain. It takes each joint and looks at its immediate neighbours which are the other joints it's directly connected to. It then combines information from a joint and its neighbours to create a new representation of that joint. This new representation captures not just the joint's own properties but also how it relates to its connected neighbours. By applying the same filter across all joints in the skeleton, the spatial convolution operation produces a new set of features that reflect the spatial structure of the skeleton at that particular time step.

Attention layer: This layer helps to calculate a self-attention map which makes the adjacency matrix dynamic, assigning weights based on their importance for a given exercise sequence. For this purpose, we have utilized the non-local attention mechanism adopted from the work by [2]. The advantage of this network is that it is flexible and can be added to the model without the need to significantly alter the existing structure of the network. The non-local attention mechanism takes into account the attributes of every joint in the body, not limiting its focus to those that are immediately connected. By doing so, it assigns unique importance to each joint based on its role in a particular action. For instance, let's consider the action of a shoulder rehabilitation exercise where a patient is required to lift their arm. While the movement primarily involves the shoulder and arm, the correct posture may also depend on the position of the spine and hips to maintain balance. The non-local attention mechanism can effectively discern the importance of these non-adjacent body parts and their contribution to the execution of the exercise. By giving more importance to the spine and hip positions along with the shoulder and arm, the mechanism can provide a comprehensive assessment of the patient's form and technique, which is critical for effective rehabilitation.

Temporal layer: Following attention processing, the data undergoes temporal analysis where features are extracted in the temporal domain. Temporal layers look at the sequence of movements over time, capturing the dynamics of how joints move and how their movements are related across the duration of the exercise.

The output from the spatial, attention, and temporal layers is then aggregated using a global pooling operation. This step synthesizes the information across the entire sequence and all joints to form a unified representation of the movement. Lastly, the pooled data is further processed by one or more fully connected layers, where the high-level features extracted from the previous steps are combined and transformed to aid in the final prediction score.

3.2 Problem Formulation

Assuming we have videos V_i capturing various instances of a specific exercise, each video being a sequence of frames X_t with a variable length n_i . These videos are labelled with a score y_i that ranges from 0 to 1, with higher scores indicating better execution of the exercise. Our collection of training data consists of pairs of videos and their corresponding scores for the exercise, totalling T examples.

We train a model F with parameters θ_e for the exercise, which aims to predict a score \hat{y}_j that closely matches the actual quality score y_j^* for any given test video V_j . The model's prediction for a test video is expressed as $\hat{y}_j = F(V_j; \theta_e)$, with the goal of \hat{y}_j being a close approximation of y_j^* .

In our setup, each RGB-D frame is assumed to feature a single, prominent individual. To model the human form, we turn to skeleton data that outline the coordinates of various joints. Such skeletal data are conveniently acquired in real-time from advanced RGBD cameras like the Kinect, or through algorithms like BlazePose or VideoPose3D.

Consider that each joint is defined by a C -dimensional coordinate vector, the dimensions of a frame from the i -th video are denoted by $X_t \in \mathbb{R}^{N \times C}$, and the video itself by $V_i \in \mathbb{R}^{T \times N \times C}$. When assessing the quality of the exercise performed, different joints have varying levels of importance. Our observations reveal that this importance varies depending on the specific exercise.

We also introduce $M_j \in \mathbb{R}^{T \times N \times N}$ as the self-attention map for the j -th video, highlighting the significance of each joint during the exercise. By studying M_j across a broad spectrum of users, including both patients and experts, we can gain valuable insights into the elements that influence the final score prediction. When provided with an RGBD video V_j , our objectives are twofold: first, to assess the exercise by predicting the score \hat{y}_j , and second, to calculate the self-attention map, M_j , which reflects the importance of each joint.

3.3 Attention-enhanced Spatial-Temporal Graph Convolutional Network

The quantitative rehabilitation assessment is typically a regression task that predicts a performance score from an input landmark sequence. Figure 3.2 (left) shows the overall structure of the proposed attention-enhanced spatial-temporal graph convolutional network, inspired by the architecture proposed by [20] for the action recognition task.

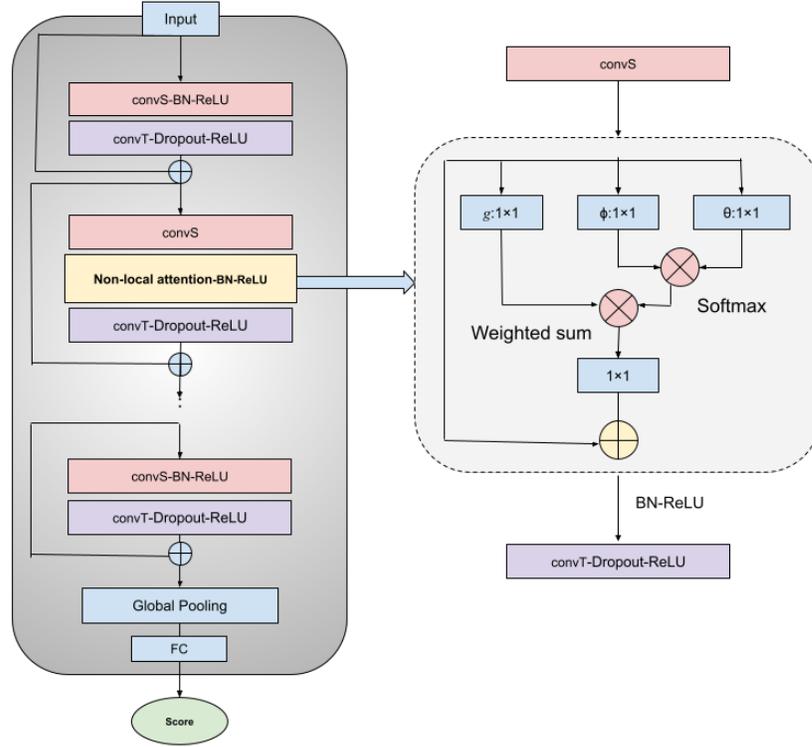


Figure 3.2: The architecture of the proposed attention-enhanced spatial-temporal graph convolutional network (left). Non-local attention block (right).

We perform spatial as well as temporal convolutions on the input data to learn the local patterns. Further, these features are concatenated with the original input to learn more complex patterns. This processed input is fed into the ST-GCN which includes the non-local attention module adapted from [2] between the spatial and temporal convolution layer. Spatial convolution is applied over the nodes in the same frame to capture the spatial relationships between different body parts. Afterwards, we employ an attention module as shown in Figure 3.2 (right), to learn long-range non-local dependencies in the spatial features. After processing through several layers of spatial and temporal blocks along with a non-local attention mechanism, a global pooling layer aggregates the features across the layers to predict the final assessment score.

3.3.1 Graph Construction

Our goal here is to understand the motion patterns by examining the relationship between the joint coordinates over time. We use a more complex method called a spatial temporal graph, denoted as $G = (V, E)$ which was initially proposed by Yan et al. [1] in 2018. In our graph, V represents all the joints across all frames of the video. So $V = \{v_{t,i} | t = 1, \dots, T, i =$

$1, \dots, N\}$ includes all these points. For each joint (or point) $v_{t,i}$, we consider not only its location (coordinates) but also how confident we are about its position (estimation confidence value).

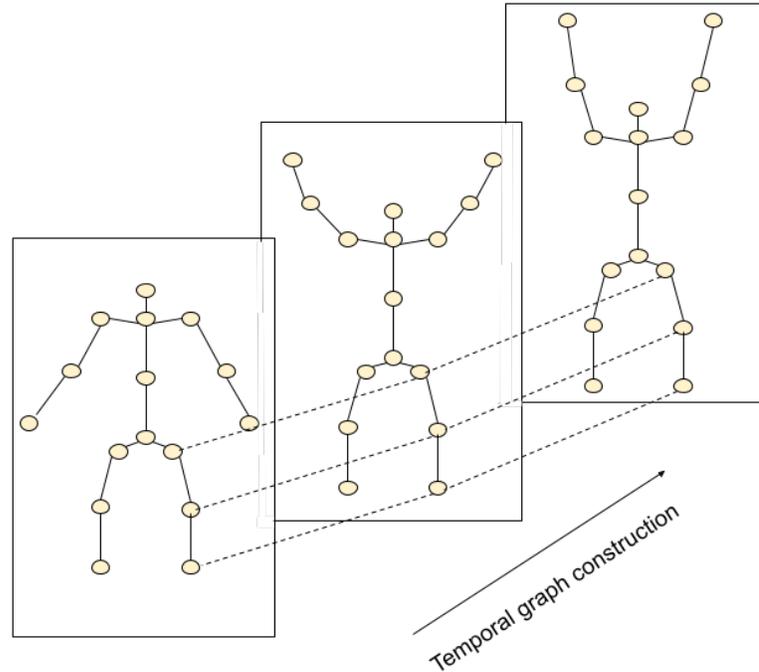


Figure 3.3: The spatial and temporal graph of a skeleton sequence proposed by [1]. Yellow dots denote the body joints and edges are connected in inter-frame and intra-body fashion.

Constructing the graph: The joints are connected in mainly two ways as depicted in Figure 3.3:

Intra-body Connections: Within each frame, we connect the points based on how joints are connected in the human body (like connecting elbow to shoulder). These connections are represented by $E_S = \{v_{t,i}v_{t,j} | (i, j) \in H\}$, where H contains pairs of joints that are naturally connected.

Inter-frame Connections: We also connect each joint to its corresponding joint in the next frame. This shows how each joint moves from one frame to the next and is represented by $E_F = \{v_{t,i}v_{(t+1),i}\}$. These connections help track the movement of each joint over time.

This approach allows us to analyze movements more comprehensively, regardless of the number of joints or how they're connected, making it adaptable to different datasets.

3.3.2 Partitioning Strategy

The ST-GCN model employs a technique to categorize the immediate neighbours (1-neighbour) of a joint into three distinct groups based on their spatial relation to the body's centre of gravity. This technique is depicted in Figure 3.4 and is described as follows:

a) Root Node: This is the central joint of interest, marked as a red dot. It is the reference point from which the neighbours are evaluated.

b) Centripetal Subset: These are the joints, shown as a green dot, that are situated closer to the body's centre of gravity, represented by a black cross. They are considered to have a pulling effect towards the centre.

c) Centrifugal Subset: Represented by blue dots, these joints are positioned further away from the body's centre of gravity. They are perceived to be influenced by a pushing effect away from the centre.

Each group is associated with a unique weight vector that can be adjusted during learning.

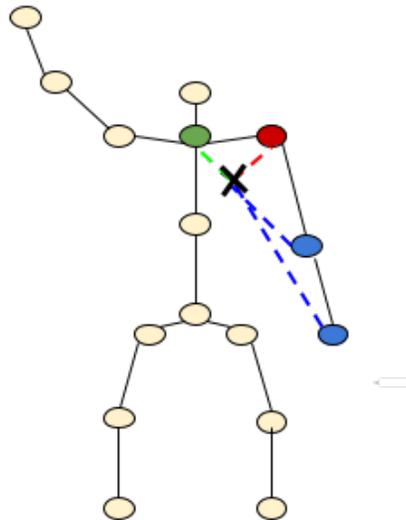


Figure 3.4: An illustration of the partitioning strategy adopted in ST-GCN architecture.

3.3.3 Spatial and Temporal Modeling

Every frame of a skeleton sequence is analyzed using a graph structure, which is made up of a set of nodes representing the joints (denoted by N_G), a set of edges that connect these joints (denoted by E_G), and an adjacency matrix (denoted by A) that maps out how each joint is connected. The adjacency matrix, A_k , is specifically adjusted to account for the connections at

various distances, or ‘k-steps’, between joints. It does this by incorporating self-connections through an identity matrix I and normalizing with the degree matrix D_k of the graph. This is expressed using equation 3.1 below:

$$A_k = D_k^{-1/2} \cdot (\tilde{A}_k + I) \cdot D_k^{-1/2}. \quad (3.1)$$

The k-adjacency matrix, $\tilde{A}^{(k)}$, is created to measure the connections between nodes that are ‘k’ steps apart. If two joints are connected by ‘k’ steps, the matrix entry is 1. If it’s the same joint (no steps apart), the entry is also 1. All other entries are 0.

To process the skeleton sequence, a graph convolution is performed. This involves taking the product of the processed input sequence P with each adjacency matrix A_k , then applying a set of weights W_k . This convolution takes into account each joint’s features and its neighbours, essentially averaging them based on their connectivity. All the features generated from the convolution are then combined to create a comprehensive feature vector for each joint. This can be expressed as equation 3.2, similar to the classical convolution operation for ST-GCN by [1]:

$$G(P) = \sum_k (P \cdot A_k) \cdot W_k. \quad (3.2)$$

Equations 3.1 and 3.2 collectively describe how the spatial features of the skeleton sequence are extracted using a graph-based approach, taking into account the natural connections between joints and the movement over time.

3.3.4 Attention Mechanism

In the standard ST-GCN model as expressed in Equation 3.2, the convolution layer’s receptive field considers only one-neighbourbor adjacent nodes for the calculation, which limits it to capture only local features. However, in real-life scenarios, a combination of multiple non-local nodes can interact for a given action or exercise sequence. For example, if the patient is performing a sidelong, the role of waist position and spine alignment can play as crucial a role as the knee joint positions. Therefore, to capture such non-local dependencies, a non-local attention block is incorporated into the network which directly focuses on the features of all joints, and gets more efficient features by attention operations. The updated graph convolution function is expressed as:

$$f_{\text{out}} = \sum_j \left(\Lambda^{-\frac{1}{2}} (A + I) \Lambda^{-\frac{1}{2}} \otimes M_j \right) f_{\text{in}} W_j, \quad (3.3)$$

Where; f_{out} represents the resulting feature map, and f_{in} is the input feature map that is a $C_{in} \times T \times V$ dimensional tensor, with V indicating the number of nodes, T the temporal length, and C_{in} the number of input channels. The adjacency matrix A , which is an $18 \times 18 \times 3$ matrix, captures the inter-node relationships within the skeleton. The matrix W_j is a weight vector applied through a 1×1 convolution process. Furthermore, we modify the adjacency matrix A_k by incorporating a $V \times V$ trainable self-attention map, denoted as The matrix M that adjusts the influence of each node based on its importance. The operation \otimes symbolizes the Hadamard product, meaning that the attention map is applied to the adjacency matrices in an element-wise manner. This ensures that the attention map only affects existing connections, as any non-connection represented by a zero in A remains zero regardless of M 's values. Therefore, M essentially affects only the immediate neighbours of a given node.

The non-local neural network is a highly adaptable component that can be seamlessly integrated into pre-existing 2D and 3D convolutional neural networks. This integration enables the fusion of both broad (global) and specific (local) data, creating a more complex hierarchical structure of information. The non-local operation for deep neural networks can be formulated as:

$$y_i = \frac{1}{C(x)} \sum_{\forall j} f(x_i, x_j) g(x_j). \quad (3.4)$$

In this context, 'i' denotes the index for a specific output location (which could be in space, time, or both), where we want to calculate the response. 'j' is an index that covers all possible positions. 'x' represents the input data (which could be an image, a sequence, or a video, and often includes their extracted features), while 'y' is the resultant output data, which has the same dimensions as 'x'. The function 'f' measures a scalar value that signifies the relationship, like similarity, between the position 'i' and every position 'j'. A separate function 'g' is responsible for translating the input data at position 'j' into a new representation. The overall response at each position 'i' is then scaled by a normalization factor 'C(x)'.

3.4 Network Architecture

The architecture of the model is structured into 9 layers, each designed to perform spatial-temporal graph convolution operations. Output channels increase from 64 in the first three layers to 128 and then 256 in the final layers. Each layer includes a residual connection for improved information flow.

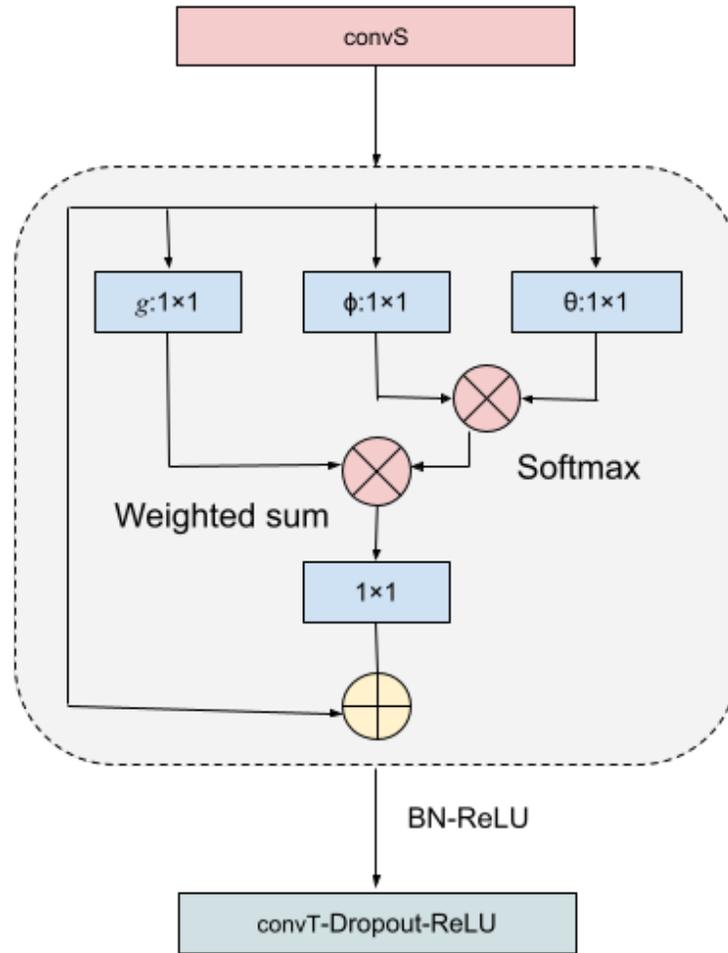


Figure 3.5: An illustration of the non-local neural network proposed by [2] used as the attention block.

To build our attention module as shown in Figure 3.5, inspired by the non-local neural network concept [2], initially, a linear transformation is applied to the feature map produced by the spatial convolution (ConvS). This transformation, executed through a 1×1 convolution process, generates three sets of features: θ , φ , and g . Next, an element-wise multiplication of θ and φ is calculated, which allows us to determine the feature auto-correlation. Following this, the Softmax function is utilized to determine the self-attention coefficients, which indicate the relevance of each feature. These self-attention coefficients are then reapplied to the g feature matrix, enhancing the features with the calculated attention values. Lastly, a residual connection is established with the initial input feature map, resulting in an updated set of features. Additionally, we implement a 2×2 MaxPooling operation post-processing the θ and φ features that reduce the computational operations.

The pooled data is further processed by fully connected layers, where the high-level features extracted from the previous steps are combined and transformed to aid in the final prediction task. Finally, the output from the fully connected layers is passed to a scoring mechanism, which generates a performance score based on the analyzed movements.

3.5 Training Loss

Similar to [21], we use the Huber loss function for our experiment, since it is robust to outliers. Rehabilitation data may include noise or outliers due to inconsistent performance of patients or errors in measurement. Huber loss reduces the influence of these outliers compared to mean squared error (MSE), which can be disproportionately affected by them. The function is defined as:

$$\delta(a) = \begin{cases} \frac{1}{2}a^2 & \text{for } |a| \leq \delta, \\ \delta(|a| - \frac{1}{2}\delta) & \text{otherwise,} \end{cases} \quad (3.5)$$

where $L_\delta(a)$ is the Huber loss, a is the error term (difference between the predicted value and the actual value), δ is a threshold parameter that defines the point where the loss function changes from a quadratic function to a linear function.

The equation states that for errors smaller than δ , the loss is quadratic, making it similar to the mean squared error (MSE). For errors larger than δ , the loss is linear, which reduces the influence of outliers on the overall loss.

For small errors, the Huber loss acts like the mean squared error (MSE) and provides stable and small-error-sensitive updates. This is important in rehab exercises where small improvements are significant and should influence the model's learning. Since the Huber loss is less steep for large errors, it provides more stable gradient updates during training. By combining the best of both the mean squared error (MSE) and mean absolute error (MAE), the Huber loss function offers a compromise that can lead to better performance in the presence of outliers while still maintaining sensitivity to small errors.

Chapter 4

EXPERIMENTAL RESULTS

4.1 Experimental Setup

4.1.1 Dataset

For the evaluation, we use two publicly available datasets; UI-PRMD and KIMORE. We list down the features of these datasets in table 4.1

The University of Idaho – Physical Rehabilitation Movement Dataset (UI-PRMD), developed by Vakanski et al. [22], comprises 10 exercises commonly utilized in physical rehabilitation programs. These exercises include deep squat, hurdle step, inline lunge, side lunge, sit-to-stand, standing active straight leg raise, standing shoulder abduction, standing shoulder extension, standing shoulder internal-external rotation, and standing shoulder scaption. Each exercise was performed 10 times by 10 healthy subjects, demonstrating both correct and incorrect techniques. Data collection involved two sensors: a Vicon optical tracking system and Kinect v2. The Vicon system provided positions and orientation angles for 39 joints, while Kinect measured the positions and orientation angles for 22 joints during the exercise movements.

The KInect-based MOvement Rehabilitation dataset (KIMORE), introduced by Capecci et al. [23], utilized Kinect v2 and involved 78 participants performing five exercises recognized in low back pain physiotherapy. These exercises encompass lifting the arms, lateral tilt of the trunk with arms in extension, trunk rotation, pelvis rotations on the transverse plane, and squatting. The study included 44 healthy subjects and 34 patients with chronic motor disabilities. Clinicians established rules for feature extraction from raw data for each exercise, and these features are provided in the dataset.

Table 4.1: Summary table of KIMORE and UI-PRMD dataset.

Feature	UI-PRMD dataset (2018) [22]	KIMORE dataset (2019) [23]
Modality	Skeleton data	RGB-D and skeleton data
No. of Subjects	10	78
No. of Exercises	10	5
Score range	0 - 1	0 - 50

4.1.2 Implementation Details

We implemented the proposed attention-enhanced the ST-GCN model using the pyTorch framework and two Nvidia RTX 3060 GPUs, each with 12 GB RAM. We train our model using Adam optimizer for 1500 epochs with a 0.0004 learning rate. We choose a batch size of 10 for the KIMORE and 3 for the UI-PRMD dataset. We choose the best weight to assess the model performance on the test set. We set the dropout to 0.25 to avoid over-fitting and add the ReLU function after each temporal convolution operation. For the Huber loss function, after experimenting with different values, we set the value of δ to 0.1. For quantitative evaluation we use mean absolute deviation (MAD), root mean square error (RMSE) and mean absolute percentage error (MAPE) metrics as detailed in the next section.

4.1.3 Evaluation Metrics

Similar to [3] and [21], we have used MAD, RMSE and MAPE metrics to evaluate the performance of our model. The equations for the metrics are given as follows:

1. Mean Absolute Deviation (MAD):

$$MAD = \frac{1}{n} \sum_{i=1}^n |e_i| \quad (4.1)$$

where $e_i = \hat{y}_i - y_i$

2. Root Mean Square Error (RMSE):

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n e_i^2} \quad (4.2)$$

where $e_i = \hat{y}_i - y_i$

3. Mean Absolute Percentage Error (MAPE):

$$MAPE = \frac{100\%}{n} \sum_{i=1}^n \left| \frac{e_i}{y_i} \right| \quad (4.3)$$

where $e_i = \hat{y}_i - y_i$

In these equations, y represents the actual values (targets), \hat{y} represents the predicted values, n is the total number of observations, and e is the error or difference between the predictions and the actual values.

4.2 Ablation Study

In this section, we report the results of various ablation studies on the KIMORE dataset, to find out the best-performing architecture for evaluation. Later, we perform the experiments on UI-PRMD to verify the generalization capabilities of the selected architecture.

4.2.1 Effect of Attention Block Incorporated with the ST-GCN

We investigate the effect of the attention block in the ST-GCN architecture as shown in Table 4.2. Results show that for our particular assessment task, the performance of the ST-GCN model with attention block is better for all the metrics functions with lower values of MAD, RMSE, MSE, and MAPE scores, as compared to the traditional ST-GCN model.

Table 4.2: Comparison of our proposed model with and without Attention block on KIMORE dataset.

	MAD	RMSE	MAPE
Without attention	1.127	2.384	2.402
With attention	1.008	2.228	2.204

4.2.2 Effect of Position of Attention Block

We experimented with the position of the attention block by placing it after different spatial convolution layers of the ST-GCN model. We found out that placing the attention block in the initial layers of the model yields better performance rather than placing it in the last layers. The comparative results are given in Table 4.3 below.

Table 4.3: Comparison table of the position of one attention block in different spatial convolution (convS) layers of the ST-GCN model.

ST-GCN's Layers	ConvS ₁	ConvS ₂	ConvS ₃	ConvS ₄	ConvS ₅	ConvS ₆	ConvS ₇
MAD score	1.058	1.055	1.061	1.071	1.110	1.106	1.114

4.2.3 Effect of Adding Multiple Attention Blocks

We evaluated the performance of our model by adding multiple attention blocks instead of one and found out that it achieved a lower metrics score as shown in Table 4.4. Specifically, adding two attention blocks after the second and third spatial convolution (convS). Hence, in both the layers we added two attention blocks. This is justified because adding multiple attention blocks can reinforce the correlation information learned in the previous attention block. Therefore it assigns each node a more appropriate weight.

Table 4.4: Comparison table of adding multiple attention blocks in the second and third spatial layers of the ST-GCN.

No. of attention blocks	ConvS ₂ +2	ConvS ₂ +3	ConvS ₃ +2	ConvS ₃ +3
MAD score	1.031	1.047	1.043	1.059

4.3 Results

4.3.1 Quantitative Results of Rehabilitation Exercise Assessment on the UI-PRMD and KIMORE dataset

Table 4.5 contains the comparative result of our proposed method and the existing methods on the UI-PRMD dataset. As can be seen from the table, our model achieves improved results with lower metric scores. However, the improvement is not very significant. A possible reason could be that the UI-PRMD dataset does not include such exercise sequences that involve interactions between different body parts. The exercises are rather simple, so the movement does not require the involvement of non-local joints.

In Table 4.6, we report the results for Mean Absolute Deviation (MAD), Root Mean Square Error (RMSE), and Mean Absolute Percentage Error (MAPE) across five exercises from the KIMORE dataset and compare it with Liao et al. [3]. Liao et al. introduced a temporal pyramid network that handles various scales of movement repetitions. The early layers of their model are designed to capture spatial relationships in human movement, which are then processed through successive LSTM layers to discern temporal patterns within the captured

Table 4.5: Comparison of mean absolute deviation (MAD) scores of existing methods and our proposed method on the UI-PRMD dataset. A lower value indicates better result.

Exercise	Ours	Chen et al. [19]	Liao et al. [3]	Deep CNN [3]	Deep LSTM [3]	Hier. LSTM [22]
E1	0.016	0.009	0.011	0.014	0.017	0.030
E2	0.026	0.020	0.028	0.030	0.049	0.077
E3	0.034	0.036	0.040	0.041	0.094	0.138
E4	0.006	0.014	0.012	0.016	0.016	0.036
E5	0.009	0.015	0.019	0.013	0.025	0.064
E6	0.023	0.020	0.018	0.023	0.022	0.047
E7	0.023	0.022	0.038	0.033	0.041	0.193
E8	0.030	0.023	0.023	0.029	0.046	0.073
E9	0.018	0.026	0.023	0.025	0.044	0.065
E10	0.031	0.026	0.046	0.037	0.052	0.160
Average	0.020	0.021	0.025	0.026	0.041	0.088

Table 4.6: Comparison of mean absolute deviation (MAD), Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE) scores of the model by Liao et al. [3] and our proposed method on the KIMORE dataset.

Metric	Exercise	Ours	Liao et al. [3]
MAD ↓	Ex1	0.870	1.141
	Ex2	1.126	1.528
	Ex3	0.660	0.845
	Ex4	0.525	0.468
	Ex5	0.739	0.663
RMSE ↓	Ex1	2.329	2.534
	Ex2	2.877	3.738
	Ex3	1.436	1.561
	Ex4	0.640	0.792
	Ex5	1.996	1.931
MAPE ↓	Ex1	2.166	2.589
	Ex2	3.272	3.976
	Ex3	1.017	1.023
	Ex4	2.112	2.333
	Ex5	2.307	2.312

spatial data. However, their model does not take into account the structural information of the human skeleton, which may result in a loss of significant spatial feature details that are crucial for comprehensive movement analysis. On the contrary, our model is built to learn such correlations which contributes to achieving improved results.

4.3.2 Feature Visualization

In Figure 4.1 and Figure 4.2, we display the correlation matrix obtained from our method which depicts the interaction of local as well as non-local joints on particular activities within the KIMORE dataset. The map represents correlation values between joints depending on their interaction, which serves to highlight the differential significance of each joint in relation to a variety of rehabilitative movements.

In Figure 4.1 the correlation map for exercise 1 in the KIMORE dataset is visualized. Exercise 1 is a lifting arm that involves four key joints; elbow, spine, thumb, and wrist. In the attention map, these key joints are in darker shades, which represents a stronger correlation between these joints for a given exercise sequence. Whereas, the remaining joints are in lighter shades which represent less important interactions. Similarly, for exercise 4, which is pelvis rotation, the spine base and wrist play more important roles. It’s worth noting that the spine and wrist are distant joints (more than 1-neighbour apart) and the model is successfully able to capture such non-local features.

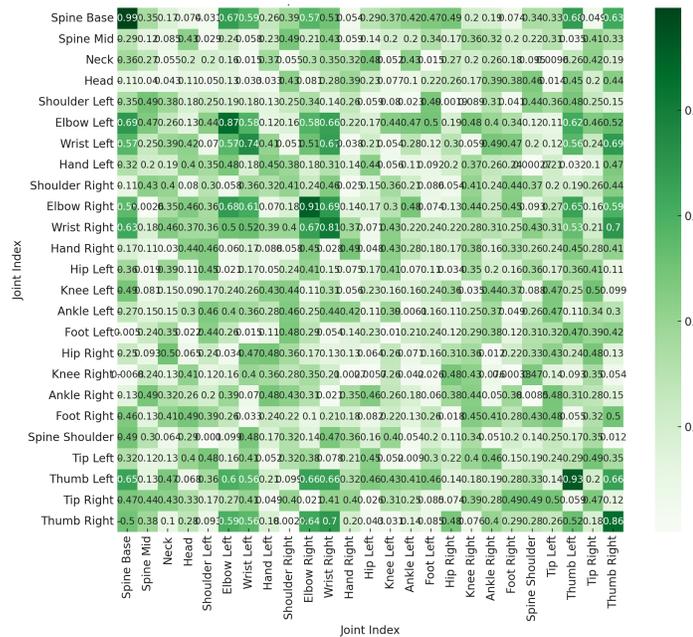


Figure 4.1: A visualization of the attention map for exercise 1 (lifting arms) in the KIMORE dataset. Darker shades represent a stronger correlation and vice-versa.

Figure 4.2 represents the correlation matrix for all five exercise sequences in the KIMORE dataset which consists of 25 joint co-ordinates. The attention value calculated by our approach shows the involvement of the joints depending on the corresponding activities. These findings closely align with the KIMORE dataset paper [23] that provides the details on the variation of the role of joints in individual exercises as suggested by clinical experts.

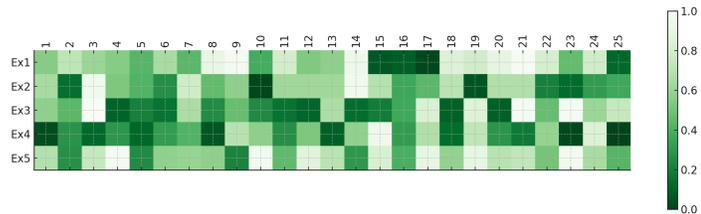


Figure 4.2: A visualization of how the role of joints varies with different exercises as determined by the attention value computed by our method for the KIMORE dataset. Darker shades represent stronger correlation and vice-versa.

Chapter 5

CONCLUSION

This thesis presents an innovative approach to assessing rehabilitation exercises through an attention-enhanced graph convolutional network. The research addresses the need for automated, accurate, and comprehensive analysis of physical rehabilitation exercises, particularly in a home-based setting. The traditional methods for physical therapy assessment, though effective, often face challenges like the need for continuous professional supervision, high costs, and the subjectivity of manual assessments. To overcome these challenges, the thesis proposes a vision-based technique leveraging the advancements in spatial and temporal feature analysis of skeleton data.

The core contribution of this research is the development of an attention-enhanced graph convolutional network model. Our model captures both spatial and temporal dependencies among body joints and applies an attention mechanism to understand the roles and interrelations of these joints more effectively. The network architecture, based on the spatial-temporal graph convolutional network (ST-GCN), is adapted to include a non-local attention module that enables the model to consider global information and relationships between non-adjacent joints, essential for a nuanced assessment of rehabilitation exercises.

5.1 Limitations

Data Dependency: The effectiveness of the model is dependent on the quality and quantity of the training data. As there is a scarcity of publicly available medical and healthcare datasets, it is a challenging task to improve the performance of the model. The UI-PRMD and KIMORE datasets that is used for training and evaluation are relatively small in size and do not include the subjects from all group. This limits the model's capability to generalize on to a population of a wider age group.

Handling variable-length input: In a real-life setting the patients performing rehabilitation

exercises could be of different age groups. They tend to perform exercises with varying paces. Therefore, it is important for the network to deal with such varying-length input. Currently, our model is not capable of dealing with a wide-range of input length. We plan to work on this in future.

Attention mechanism in the temporal domain: Our research deals with attention mechanism in only the spatial domain. Its scope in the temporal domain hasn't been explored yet. They are particularly useful in dealing with long sequences. These temporal attention weights can offer insights into which time steps the model considers important, thereby increasing the interpretability of the model's decisions.

5.2 Application for Real-time Assessment and Feedback

To effectively implement this model for real-time application in rehabilitation exercises, several work need to be done to ensure a reliable and convenient assessment and feedback:

Generalization Capability: The model's versatility should be enhanced to handle a wide range of exercises and adapt to diverse patient profiles, ensuring it's effective for various rehabilitation needs.

Expert Collaboration: Regular feedback should be incorporated from healthcare professionals to align the model with clinical standards and address practical healthcare needs.

Execution Efficiency: The model needs to be optimized to decrease time complexity, ensuring fast processing for real-time feedback. This could involve leveraging parallel computing and efficient algorithms.

User-Friendly Interface: An intuitive interface is needed for patients' convenient interaction, focusing on simplicity and clarity to enhance user experience and accessibility.

Scalable Deployment: It is crucial to ensure that the system is scalable and can be integrated with existing healthcare IT infrastructure, supporting a range of settings from individual homes to clinical environments.

Focusing on these areas will be crucial for the successful real-time implementation of the model, ensuring it is not only technologically sound but also clinically relevant and user-centred.

5.3 Future Work

For future work, we want to improvise the model to effectively process inputs of varying lengths, which is essential for accommodating patients with different exercise paces and duration. This could involve implementing dynamic sequence adjustment techniques, such as

adaptive padding or segmentation, and potentially integrating more flexible neural network architectures like LSTM and GRU that are better suited for variable-length data.

Moreover, we want to extend the model to include an attention mechanism in the temporal domain. This enhancement aims to improve the model's ability to identify key time steps in long sequences. Exploring the integration of Transformer models or similar architectures could provide a pathway to capture long-term dependencies more effectively and increase the interpretability of the model's assessments.

Bibliography

- [1] S. Yan, Y. Xiong, and D. Lin, “Spatial temporal graph convolutional networks for skeleton-based action recognition,” in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, 2018.
- [2] X. Wang, R. Girshick, A. Gupta, and K. He, “Non-local neural networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7794–7803, 2018.
- [3] Y. Liao, A. Vakanski, and M. Xian, “A deep learning framework for assessing physical rehabilitation exercises,” *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 28, no. 2, pp. 468–477, 2020.
- [4] Y. Liao, A. Vakanski, M. Xian, D. Paul, and R. Baker, “A review of computational approaches for evaluation of rehabilitation exercises,” *Computers in biology and medicine*, vol. 119, p. 103687, 2020.
- [5] D. J. Gladstone, C. J. Danells, and S. E. Black, “The fugl-meyer assessment of motor recovery after stroke: a critical review of its measurement properties,” *Neurorehabilitation and neural repair*, vol. 16, no. 3, pp. 232–240, 2002.
- [6] C. G. Goetz, W. Poewe, O. Rascol, *et al.*, “Movement disorder society task force on rating scales for parkinson’s disease: the unified parkinson’s disease rating scale (updrs): status and recommendations,” *Mov Disord*, vol. 18, no. 7, pp. 738–750, 2003.
- [7] S.-k. Jun, S. Kumar, X. Zhou, D. K. Ramsey, and V. N. Krovi, “Automation for individualization of kinect-based quantitative progressive exercise regimen,” in *2013 IEEE International Conference on Automation Science and Engineering (CASE)*, pp. 243–248, 2013.
- [8] L. Tao, A. Paiement, D. Damen, M. Mirmehdi, S. Hannuna, M. Camplani, T. Burghardt, and I. Craddock, “A comparative study of pose representation and dynamics mod-

- elling for online motion quality assessment,” *Computer vision and image understanding*, vol. 148, pp. 136–152, 2016.
- [9] S. Song, C. Lan, J. Xing, W. Zeng, and J. Liu, “An end-to-end spatio-temporal attention model for human action recognition from skeleton data,” in *Proceedings of the AAAI conference on artificial intelligence*, vol. 31, 2017.
- [10] S.-k. Jun, S. Kumar, X. Zhou, D. K. Ramsey, and V. N. Krovi, “Automation for individualization of kinect-based quantitative progressive exercise regimen,” in *2013 IEEE International Conference on Automation Science and Engineering (CASE)*, pp. 243–248, IEEE, 2013.
- [11] T. T. Um, F. M. J. Pfister, D. C. Pichler, S. Endo, M. Lang, S. Hirche, U. Fietzek, and D. Kulić, “Parkinson’s disease assessment from a wrist-worn wearable sensor in free-living conditions: Deep ensemble learning and visualization,” *arXiv preprint arXiv:1808.02870*, 2018.
- [12] J. Nolan, E. Godecke, K. Spilsbury, and B. Singer, “Post-stroke lateropulsion and rehabilitation outcomes: a retrospective analysis,” *Disability and rehabilitation*, vol. 44, no. 18, pp. 5162–5170, 2022.
- [13] A. P. L. Bo, M. Hayashibe, and P. Poignet, “Joint angle estimation in rehabilitation with inertial sensors and its integration with kinect,” in *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 3479–3483, IEEE, 2011.
- [14] W. Zhao, R. Lun, D. D. Espy, and M. A. Reinthal, “Realtime motion assessment for rehabilitation exercises: Integration of kinematic modeling with fuzzy inference,” *Journal of Artificial Intelligence and Soft Computing Research*, vol. 4, no. 4, pp. 267–285, 2014.
- [15] R. Houmanfar, M. Karg, and D. Kulić, “Movement analysis of rehabilitation exercises: Distance metrics for measuring patient progress,” *IEEE Systems Journal*, vol. 10, no. 3, pp. 1014–1025, 2014.
- [16] A. Elkholy, M. E. Hussein, W. Gomaa, D. Damen, and E. Saba, “Efficient and robust skeleton-based quality assessment and abnormality detection in human action performance,” *IEEE journal of biomedical and health informatics*, vol. 24, no. 1, pp. 280–291, 2019.

- [17] W. Wei, C. McElroy, and S. Dey, “Towards on-demand virtual physical therapist: Machine learning-based patient action understanding, assessment and task recommendation,” *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 27, no. 9, pp. 1824–1835, 2019.
- [18] M. H. Li, T. A. Mestre, S. H. Fox, and B. Taati, “Vision-based assessment of parkinsonism and levodopa-induced dyskinesia with pose estimation,” *Journal of neuroengineering and rehabilitation*, vol. 15, no. 1, pp. 1–13, 2018.
- [19] C. Du, S. Graham, C. Depp, and T. Nguyen, “Assessing physical rehabilitation exercises using graph convolutional network with self-supervised regularization,” in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pp. 281–285, IEEE, 2021.
- [20] Y. Kong, L. Li, K. Zhang, Q. Ni, and J. Han, “Attention module-based spatial–temporal graph convolutional networks for skeleton-based action recognition,” *Journal of Electronic Imaging*, vol. 28, no. 4, pp. 043032–043032, 2019.
- [21] S. Deb, M. F. Islam, S. Rahman, and S. Rahman, “Graph convolutional networks for assessment of physical rehabilitation exercises,” *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 30, pp. 410–419, 2022.
- [22] A. Vakanski, H.-p. Jun, D. Paul, and R. Baker, “A data set of human body movements for physical rehabilitation exercises,” *Data*, vol. 3, no. 1, p. 2, 2018.
- [23] M. Capecci, M. G. Ceravolo, F. Ferracuti, S. Iarlori, A. Monteriu, L. Romeo, and F. Verdini, “The kimore dataset: Kinematic assessment of movement and clinical scores for remote monitoring of physical rehabilitation,” *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 27, no. 7, pp. 1436–1448, 2019.