Skeleton-Based Detection of Abnormalities in Human Actions Using Graph Convolutional Networks

Bruce X. B. Yu
Department of Computing
The Hong Kong Polytechnic University
Hong Kong, China
csxbyu@comp.polyu.edu.hk

Yan Liu
Department of Computing
The Hong Kong Polytechnic University
Hong Kong, China
csyliu@comp.polyu.edu.hk

Keith C. C. Chan
Department of Computing
The Hong Kong Polytechnic University
Hong Kong, China
cskcchan@comp.polyu.edu.hk

***Abstract*—Human action abnormality detection has been attempted by various sensors for application domains like rehabilitation, healthcare, and assisted living. Since the release of motion sensors that ease the human body skeleton retrieval, skeleton-based human action recognition has recently been an active topic in the area of artificial intelligence. Unlike human action recognition, human action abnormality detection is an emerging field that aims to detect the incorrect action from the same action class. Graph convolutional network has been widely adopted for human action recognition. However, to the best of our knowledge, whether it could be effective for the task of human action abnormality detection has not been attempted. To advance prior work in the emerging field of human action abnormality detection, we propose a novel method that uses graph convolutional network to detect abnormal actions in skeleton data. To validate the effectiveness of our proposed method, we conduct extensive experiments on a public dataset called UI-PRMD. Based on the experimental results, our proposed method achieved superior action abnormality detection performance comparing with existing deep learning methods.**

Keywords—abnormality detection, graph convolutional network, human action evaluation

I. INTRODUCTION

Traditionally, rehabilitation therapies are often conducted by experienced clinical staffs or physical therapists for the recovery and prevention of a broad array of musculoskeletal disorders like tendonitis, epicondylitis, mechanical back syndrome, etc. With the affected working ability, it is often unaffordable for patients to take such regular rehabilitation therapy episodes [1]. Accordingly, home-based rehabilitation programs initiated with the supervision of a therapist becomes a widely accepted and cost-effective alternative [2]. In home-based rehabilitation setting, patients need to follow tailored exercises and regularly report the outcome of their rehabilitation progresses to their physical therapists. However, patients are usually failed to adhere to the exercise regimens recommended by their therapists in the home-based setting, which usually leads to even higher healthcare expenditure [3]. There remains a lack of practical methods for increasing adherence to home-based rehabilitation exercises, but providing patient-identified barriers is likely to increase the adherence [4]. Patient-identified barriers is intuitively working as a professional therapist to motivate the patients to adhere their rehabilitation exercises.

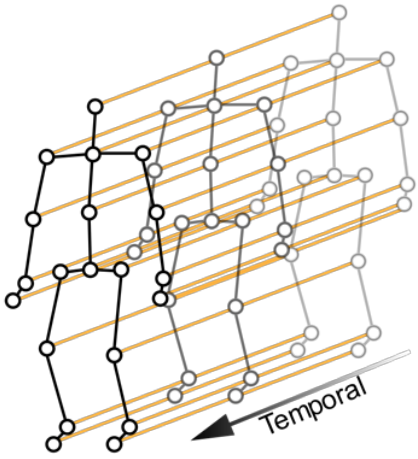

Fig. 1. A sequence of the skeleton graph representing the skeleton modality.

Recently, home-based physical therapy systems [5, 6] have been proposed to enable therapists remotely monitoring the rehabilitation process of patients and even providing real-time feedbacks to patients. In existing virtual rehabilitation systems [5, 6], the therapeutic exercises performed by patients are recorded with a motion sensor known as Kinect. The Kinect sensor could provide multiple data modalities like skeleton, depth and RGB [7]. Since the release of such affordable motion sensors, automatic detecting abnormality or incorrectness in actions like physical exercises and daily activities becomes one of the emerging topics of transdisciplinary Artificial Intelligence (AI). The depth modality of the Kinect sensor has been attempted for abnormal action detection by [8]. Nonetheless, automatic monitoring and evaluating the rehabilitation exercises in the skeleton data modality are still not tackled well.

The skeleton modality of Kinect could be represented as a sequence of human joint locations formed as 2D or 3D coordinates. By analyzing the joint movement patterns thereof, abnormal actions can then be detected. Prior work of using skeleton data for assessing rehabilitation exercise utilize handcrafted geometrical features or Deep Learning (DL) models [9]. The capability of separating incorrect actions from correct ones are limited in existing methods as they do not consider the spatial relationships among the skeleton joints that are essential for human action understanding. In the human activity

978-1-7281-8699-3/20/$31.00 ©2020 IEEE
DOI 10.1109/TransAI49837.2020.00030

Authorized licensed use limited to: Hong Kong Polytechnic University. Downloaded on June 30,2021 at 07:36:54 UTC from IEEE Xplore. Restrictions apply.

recognition domain, Graph Convolutional Network (GCN) have achieved encouraging performance on classifying different human activities [10] [11] [12]. Yet, it has not been adopted for the task of abnormal action detection.

In this article, we propose a novel neural representation approach by using the GCN for skeleton-based action abnormality detection. As illustrated in Fig. 1, the skeleton modality is represented as a sequence of skeleton graph. Since GCN shows effective representation ability in skeleton-based action recognition, we adopted three graph convolution kernels proposed in [10] as the convolution traversal rules for the proposed method. For abnormal action detection, inspired by the work of that the probability distribution before the SoftMax layer contains more information than the result of SoftMax classifier [13], we utilize the probability distribution before the SoftMax layer to generate evaluation scores. The evaluation scores are then used for abnormality detection with a proper threshold value. The main contributions of this paper are as below:

- For action abnormality detection, we propose a novel representation method that uses a GCN model to classify if an exercise is abnormal or normal.
- To verify the representation ability of the proposed method, we test it on the UI-PRMD dataset [14] for the task of abnormality detection.
- Our proposed method is one the first attempts utilizing GCN for the skeleton-based action abnormality detection, which could have plenty of real-world applications given the superior abnormality detection performance over the existing methods.

II. Related Work

We now review state-of-the-art work on skeleton-based action abnormality detection from both perspectives of algorithms and datasets. We propose our algorithm and conduct experiments based on the previous literature.

A. Action Abnormality Detection Algorithms

Action abnormality detection is closely related to the field of Human Action Evaluation (HAE) that aims to design computation models and evaluation methods to automatically assess the quality of human motions. Qing et al. [9] surveyed the potential applications in domains like physical rehabilitation, assistive living for elderly people, skill training, and sports activity scoring. It turns out that HAE relies on human tracking, human motion recognition, action segmentation, and efficient methods for evaluate the quality of the action performance. As far as we know, there are very few works that investigate the standard evaluation methods of action abnormality detection algorithms. Following the review of [9], we investigate existing action evaluation methods from two categories: handcrafted feature representation methods and deep learning feature representation methods.

Handcrafted feature representation methods reply on constructing effective geometrical feature vectors that encode the skeleton data of rehabilitation exercises. Traditionally, Hidden Markov Model (HMM) was popularly utilized as in the work of [15], which evaluates human actions based on geometric features of the skeleton data. The training process of the HMM model in [15] is supervised by the abnormality degree (on the scale of 1 to 5) evaluated by a professional physiatrist. Various traditional methods based on HMM were compared in [16] and their performance have been surpassed by DL models according to the experimental results on various datasets [17]. For example, a DL framework was proposed to encode the skeleton exercise data, which is supervised by a quality score function [18]. Common DL models like Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) have been attempted in [19] for gesture correctness estimation. However, the dataset in [19] has not yet been publicly accessible by the time of this study. For HAE, existing methods are either supervise by human labels [15] or an arbitrary score function [18]. On one hand, training with the subjective human labels will make the evaluation results remain subjective and unacceptable for patients. One the other hand, supervised the training process with a score function has a redundant issue as the results could already be delivered by the evaluation function. Our method is different from these existing methods that are supervised either by the arbitrary function score or the abnormality degree.

B. Skeleton-Based HAE Datasets

Very few works have reviewed the datasets relating to HAE. Ahad et al. [20] briefly collected some HAR datasets, but most of the surveyed datasets are focusing on action recognition instead of abnormality detection. Due to the recent popularity of the Kinect sensor for human action analysis, we further investigate the evaluation methods of representative benchmarks that use Kinect for skeleton retrieval.

One of the public HAE datasets is the SPHERE dataset [16] that includes three sub-datasets: Staircase2014, Walking2015 and SitStand2015. SPHERE is originally collected for a competition and just provide the body center data instead of the whole skeleton. Another dataset called EJMQA including four simple actions was collected by [15], which is a similar dataset with SPHERE [16]. A fitness exercise dataset called UI-PRMD [14] that includes both correct and incorrect actions is collected for evaluating HAE algorithms. Since the UI-PRMD dataset [14] did not provide a standard evaluation method, [18] proposed the quality score function, which makes it meaningless to train a representation model as the results already could be inferred by the quality score function. The lack of standard evaluation methods also makes UI-PRMD hard to be compared by a similar work in [21]. Unlike the 10 incorrect exercises in UI-PRMD that are simulated by the ones that perform the other correct motion sequences, the AHA-3D dataset [22] is performed by both elderly and young people but the AHA-3D dataset is not publicly accessible by the date of this work. The evaluation method in [22] is per frame but not in terms of the whole action sequence.

III. Proposed Method

In this section, we introduce our skeleton-based action abnormality detection method. As Fig. 2 shows, we adopt a graph convolutional network to learn a representation of the skeleton data and then infer the severity level of abnormality based on the feature distribution before the SoftMax Layer determined by the GCN model.

Authorized licensed use limited to: Hong Kong Polytechnic University. Downloaded on June 30,2021 at 07:36:54 UTC from IEEE Xplore. Restrictions apply.

Fig. 2. Illustration of the training model. There are 9 ST-GCN blocks (B1-B9). The three numbers of each block represent the number of input channels, the number of output channels and the stride, respectively. GAP represents the global average pooling layer.

A. Data Structure and Notation

For a particular exercise, using the skeleton retrieval sensor, we could record a sequence of skeleton frames corresponding to the exercise performed as shown in Fig. 1. Give N repetitions of an exercise performed by all the subjects in a dataset, it could be denoted as $S = \{S^{(i)} \mid i = 1, \dots, N\}$, where $S^{(i)}$ is a sequence of skeleton frames that characterize the exercise. A skeleton frame is consisted of a set of skeleton joints which represent body parts like head, spine, hands, etc. For a skeleton frame with J skeleton joints observed at time t, let us represent it as $S_t^{(i)} = (S_{t1}^{(i)}, \dots, S_{tj}^{(i)}, \dots, S_{tJ}^{(i)})$, where $S_{tj}^{(i)}$ has some attributes corresponding to its position and orientation features. The position of joint $S_{tj}^{(i)}$ usually has 3 attribute features denoted as *(x, y, z)*, which indicates the 3D cartesian coordinates of the joint. Whereas the orientation of joint $S_{tj}^{(i)}$ is represented as (X, Y, Z), where X, Y and Z could be transformed to corresponding angular values of yaw, roll and pitch, respectively. In the experimentational setting, we will investigate the contribution of these features of different sensors.

B. Graph Convolutional Network

1) Graph construction: The raw skeleton data in one frame is always streamed as an ordered sequence of vectors. Each vector represents the position and orientation attributes of the corresponding human joint. A complete exercise repetition contains multiple frames with varied lengths for different repetitions. We adopt a spatiotemporal graph to model the structured information among these joints along both the spatial and temporal dimensions. The structure of the graph is similar with that of ST-GCN [10]. Fig. 1 give an example of the constructed spatiotemporal skeleton graph, where the joints are represented as vertexes and their natural connections in the human body are represented as spatial edges (the black lines in Fig. 1). For the temporal dimension, the corresponding joints between two adjacent frames are connected with temporal edges (the orange lines in Fig. 1). The position and orientation features of each joint are set as the attribute of the corresponding vertex. The skeleton graph at time t could be symbolized as $\vartheta_t = \{v_t, \varepsilon_t\}$, where v_t denotes the skeleton joints and ε_t demotes the skeleton bones, respectively. In this graph, the node set $v_t = \{v_{tj} | v_{tj} = S_{tj}^{(i)}, j = 1, \dots, J\}$ contains all joints in the skeleton sequence.

2) Convolutional operation: To represent the sampling area of convolutional operations, a neighbor set of a node v_{ti} is defined as $B(v_{tj}) = \{v_{tj} | d(v_{tj}, v_{tk}) \le D\}$, where D is the minimum path length of $d(v_{tj}, v_{tk})$. The right sketch in Fig. 3 shows this strategy, where × represents the center of gravity of the skeleton. The sampling area $B(v_{tj})$ is enclosed by the curve. In detail, the strategy empirically uses 3 spatial subsets: the vertex itself (the green node in Fig. 3); the centripetal subset, which contains the neighboring vertexes that are closer to the center of gravity (the blue node in Fig. 3); the centrifugal subset, which contains the neighboring vertexes that are farther from the gravity center (the yellow node in Fig. 3).

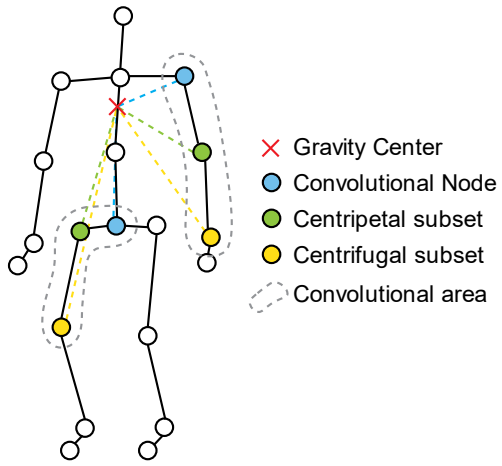

Fig. 3. Illustration of the spatial mapping strategy. Different colors nodes denote different subsets.

Suppose there are fixed number of L subsets in the $B(v_{tj})$, every neighbor set will be labelled numerically with a mapping $h_{tj}: B(v_{tj}) \to \{0, \dots, L-1\}$. Temporally, the neighborhood concept is extended to sequentially connected joints as $B(v_{tj}) = \{v_{qj} | d(v_{tj}, v_{tk}) \le K, |q - t| \le \Gamma/2\}$, where Γ is the temporal kernel size that controls the temporal range of the neighbor set. Then the graph convolution could be computed as:

$$f_{\text{out}}(v_{tj}) = \sum_{v_{tk} \in B(v_{tj})} \frac{1}{Z_{tj}(v_{tk})} f_{in}(v_{tk}) \text{w}(h_{tj}(v_{tk})) \quad (1)$$

Authorized licensed use limited to: Hong Kong Polytechnic University. Downloaded on June 30,2021 at 07:36:54 UTC from IEEE Xplore. Restrictions apply.

where $f_{in}: v_{tk} \rightarrow R^c$ is the feature map that gets the attribute vector of v_{tk}, $w(h_{tj}(v_{tk}))$ is a weight function $w\left(v_{tj}, v_{tk}\right): B\left(v_{tj}\right) \rightarrow R^c$ that could be implemented by indexing a tensor of (c, L) dimension. $Z_{tj}(v_{tk}) = \left|\{v_{tm} | h_{tj}(v_{tm}) = h_{tj}(v_{tk})\}\right|$ is a normalization term that equals to the cardinality of the corresponding subset.

3) Implementation: The implementation of graph-based convolution is not as straightforward as 2D or 3D convolution. The feature map of the network could be represented by a tensor of (C, T, J) dimensions, where C denotes the number of attributes of the joint vertex. With the specific partitioning strategy determined, it could be represented by a $J \times J$ adjacency matrix $\boldsymbol{A}$ with its elements indicating if a vertex v_{tj} belongs to a subset of $B\left(v_{tj}\right)$. The graph convolution is implemented by performing a $1 \times \Gamma$ classical 2D convolution and multiplies the resulting tensor with the normalized adjacency matrix $\boldsymbol{\Lambda}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{\Lambda}^{-\frac{1}{2}}$ on the second dimension. With L distance partitioning strategies $\sum_{l=1}^{L} \boldsymbol{A}_l$, Equation 1 could be transformed into:

$$f_{\text{out}}(v_{tj}) = \sum_{l=1}^{L} \boldsymbol{\Lambda}_l^{-\frac{1}{2}} \mathbf{A}_l \boldsymbol{\Lambda}_l^{-\frac{1}{2}} f_{\text{in}} \mathbf{W}_l \odot \mathbf{M}_l \tag{2}$$

where $\boldsymbol{\Lambda}_l^{jj} = \sum_k (\mathbf{A}_l^{jk}) + \alpha$ is a diagonal matrix with α set to 0.001 to avoid empty rows in the diagonal matrix. $\mathbf{W}_l$ 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. $\mathbf{M}_l$ is an attention map with the same size of $\mathbf{A}_l$, which indicates the importance of graph nodes. $\odot$ denotes the element-wise product between two matrixes.

C. Network Architecture

The convolution for the temporal dimension is the same as ST-GCN, i.e., performing the $1 \times \Gamma$ convolution on the $C \times T \times J$ feature map. Both the spatial GCN and temporal GCN are followed by a batch normalization (BN) layer and a ReLU layer. As Fig. 4 shows, one basic ST-GCN block is the combination of one spatial GCN (Convs), one temporal GCN (Convt) and an additional dropout layer with the drop rate set as 0.5 to avoid overfitting. To stabilize the training, a residual connection is added for each block.

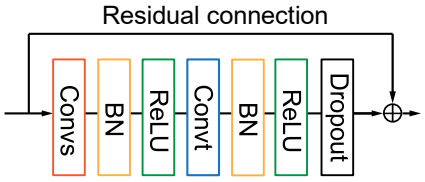

Fig. 4. Illustration of the ST-GCN block. Convs represents the spatial GCN, and Convt represents the temporal GCN, both of which are followed by a BN layer and a ReLU layer. Moreover, a residual connection is added for each block.

The ST-GCN model is a stack of these basic blocks, as shown in the middle of Fig. 2. There are 9 blocks in total. The first three layers have 64 channels for output. The follow three layers have 128 channels for output. And the last three layers have 256 channels for output. These layers have 9 temporal kernel size. The strides of the 4-th and the 7-th temporal convolution layers are set to 2 as pooling layer. data BN layer is added at the beginning to normalize the input data. A global average pooling layer is performed at the end of the ST-GCN blocks to pool feature maps of different samples to a 256-dimension feature vector. The last layer of the model is a 2D convolutional layer that transfer the 256-dimension feature vector to a 2D vector as we labelled the data as normal or abnormal. Finally, we feed it to a SoftMax classifier that will classify an action as normal or abnormal, before which we use the trained model to generate an evaluation score.

D. Optimization

The learning process of the weights Θ of the ST-GCN model G is supervised by the binary clinical label y with a cross-entropy loss as:

$$\arg\min_{\Theta} \sum_{i=1}^{N} - \sum y^{(i)} log(\sigma(G(\Theta, S^{(i)}))) \tag{3}$$

where $G(\Theta, S^{(i)})$ represents the graph convolutional operation defined in Equation 2, which is expanded to temporal dimension with the kernel size set to 1. σ is the Softmax function that transfer the recognition results to human understandable format.

We use the stochastic gradient descent to optimize the model with a base learning rate set to 0.1. The learning rate will be decayed by 0.1 at the epochs of 10, 50 and 100 throughout the total 200 epochs. The training process will be terminated when the model converges at the accuracy of 100%. To infer the action evaluation score, we retrieve the first dimension of the output before the SoftMax layer of the model and transfer it to a range of [0,1] where 0 refers to the worst exercise quality and 1 indicates the best exercise quality by using a sigmoid function defined as:

$$f_{\text{score}}(S^{(i)}) = \frac{1}{1 + e^{-fout(S^{(i)})}} \tag{4}$$

IV. Experiments

In this section, we introduce the detailed implementation of our proposed method on the UI-PRMD dataset in terms of the human action abnormality detection. We also compare features which feature and which motion sensor could be more effective for the abnormality detection task.

A. Dataset

Based on the review of existing benchmark dataset in Section II, we use the UI-PRMD dataset [14] in our experiment. UI-PRMD consists of skeletal data collected from 10 healthy subjects with every subject performing 10 repetitions of 10 rehabilitation exercises like deep squat, hurdle step, and sit to stand as illustrated in Table I. Subjects performed every exercise in both correct and incorrect manners, i.e., simulating performance by patients with musculoskeletal constraints. The data were collected with two sensors namely Kinect v2 and Vicon optical tracking system. Both sensors provide skeleton data with position (3D cartesian coordinates) and orientation (angular data) features. As Fig. 5 shows, the skeleton structures of Kinect v2 and Vicon tracking system has 22 and 39 joint

Authorized licensed use limited to: Hong Kong Polytechnic University. Downloaded on June 30,2021 at 07:36:54 UTC from IEEE Xplore. Restrictions apply.

nodes, respectively. Since the dataset has inconsistent samples caused by measurement errors and subjects performing the exercise with incorrect limbs, the dataset was transferred to a consistent version by [18]. We use the consistent data of the Kinect and Vicon sensors that have 1326 repetitions. The first purpose of using this dataset is to compare the HAE ability of our model with the one proposed by [18]. Kinect v2 has been considered as less accurate than Vicon optical tracking system. Given data are available from both sensors, the second purpose is to compare whether Kinect v2 could be better than Vicon optical tracking system for the human action analysis.

TABLE I. EXERCISES IN THE UI-PRMD DATASET

Order	Exercise
E1	Deep squat
E2	Hurdle step
E3	Inline lunge
E4	Side lunge
E5	Sit to stand
E6	Standing active straight leg raise
E7	Standing shoulder abduction
E8	Standing shoulder extension
E9	Standing shoulder internal–external rotation
E10	Standing shoulder scaption

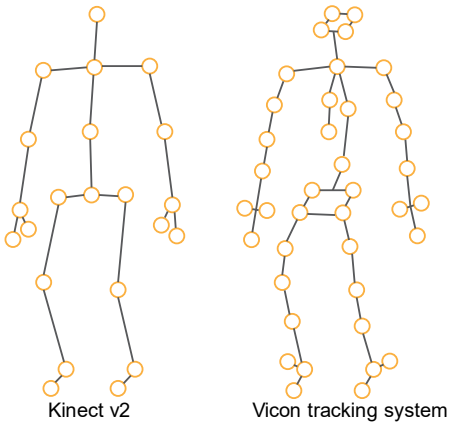

Fig. 5. The skeleton structures of Kinect v2 (22 joints) and Vicon optical tracking system (39 joints).

B. Evaluation Criterion

To test the representation power of the proposed GCN model, we adopt the concept of separation degree (SD) that is proposed in [18]. For a pair of positive numbers x and y, their SD could be defined as $S_D(x,y) = \frac{x-y}{x+y} \in [-1,1]$. Then the separation degree between two positive sequences $\boldsymbol{x} = (x_1, x_2, \dots, x_m)$ and $\boldsymbol{y} = (y_1, y_2, \dots, y_m)$ could be defined by:

$$S_D(\boldsymbol{x},\boldsymbol{y}) = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} S_D(x_i, y_j) \tag{5}$$

We define the action abnormality detection as a binary classification problem to classify exercise repetitions to correct and incorrect groups. To evaluate the detection performance, we examine the classification accuracy derived from the confusion matrix as shown in Table II.

TABLE II. CONFUSION MATRIX OF BINARY CLASSIFICATION

	1 (Predicted)	**0 (Predicted)**
1 (Actual)	True Positive (TP)	False Negative (FN)
0 (Actual)	False Positive (FP)	True Negative (TN)

Based on the confusion matrix, the overall accuracy could be calculated as:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{6}$$

C. Results and Analysis

Following the experimental setting in [18], we record the training accuracy of all exercises from the UI-PRMD dataset as shown in Table III. The training accuracy indicates that although Kinect v2 sensor might have more noise than the Vicon sensor, it still could be good enough for exercise quality evaluation. From the average accuracy, the accuracy of Kinect v2 (99.59%) is even better than that of Vicon (97.21%). It is also worth to note that the results of angular orientation attributes on both two sensors outperform the results of their 3D position attributes.

TABLE III. TRAINING ACCURACIES OF EXERCISES ON UI-PRMD

Order ID	**Training Accuracy (%)**			
	Kinect v2		**Vicon Tracking System**	
	3D Position	***Angular***	***3D Position***	***Angular***
E1	100.00	100.00	82.78	100.00
E2	100.00	100.00	96.36	90.00
E3	99.02	98.04	86.27	100.00
E4	93.57	100.00	56.43	92.86
E5	92.86	99.40	54.17	100.00
E6	100.00	100.00	100.00	97.95
E7	98.41	98.41	99.21	100.00
E8	100.00	100.00	90.48	91.27
E9	92.50	100.00	67.50	100.00
E10	50.00	100.00	85.19	100.00
Average	92.64	**99.59**	81.84	97.21

To validate the effectiveness of proposed method, we calculate the SD of each exercise that use different data attributes as shown in Table IV, which is compared with that of the DL framework proposed in [18]. The average separation degree of [18] for the inter-subject case is 0.515, while our method achieved a separation degree of 0.768 by using the same data that is the angular attributes of the Vicon sensor. From the

Authorized licensed use limited to: Hong Kong Polytechnic University. Downloaded on June 30,2021 at 07:36:54 UTC from IEEE Xplore. Restrictions apply.

SD results, we could see that our method achieves a significant improvement over the best model named Log-likelihood GMM in [18]. We also achieves an even higher separation degree (0.808) by using the angular attributes of Kinect v2 sensor. These results indicate that Kinect v2 sensor could be effective for exercise evaluation as it is better than the Vicon in terms of both the 3D position and angular orientation attributes.

TABLE IV. SEPARATION DEGREE OF EACH EXERCISE ON UI-PRMD

Order ID	Separation Degree			
	Kinect v2		Vicon Tracking System	
	3D Position	*Angular*	*3D Position*	*Angular*
E1	0.745	0.895	0.310	0.926
E2	0.797	0.961	0.720	0.610
E3	0.639	0.867	0.406	0.806
E4	0.561	0.736	0.005	0.577
E5	0.527	0.830	0.230	0.842
E6	0.865	0.764	0.579	0.670
E7	0.734	0.701	0.681	0.895
E8	0.723	0.697	0.494	0.659
E9	0.033	0.808	0.209	0.737
E10	0.060	0.823	0.253	0.904
Average	0.584	**0.808**	0.378	0.768

To have a closer observation of the result, as Fig. 6 shows, we visualize the quality evaluation values for exercise E1 of UI-PRMD by using the 3D position features of Kinect v2. It is noticeable that the correct and incorrect (follow the description in [18]) repetitions are clearly classified by using the evaluation score transferred with Equation 4 from the abnormality detection results. While in the result of [18], most correct and incorrect pairs could not be clearly separated as most of the incorrect repetitions have an evaluation score of around 0.9 given that 1 is the fully correct score. From Fig. 6, with our proposed abnormality detection method, we could see that the scores of correct repetitions are all over 0.6, whereas the scores of incorrect repetitions are all below the threshold of 0.5.

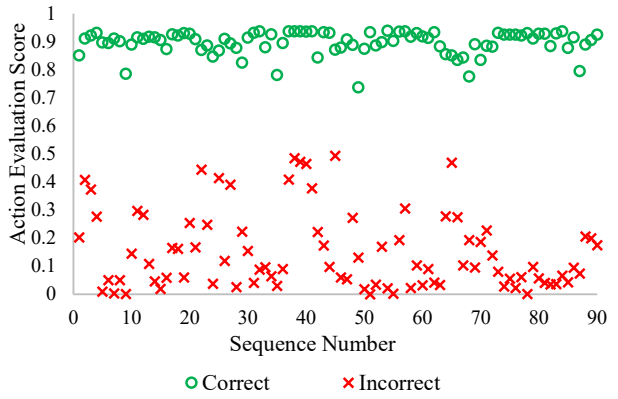

Fig. 6. Quality evaluation values for exercise E1 of UI-PRMD by using the 3D position features (S_D=0.745).

V. CONCLUSION AND FUTURE WORK

To conclude, this paper introduced an abnormality detection method that adopted a graph convolutional network to model the skeleton exercise data. Meanwhile, it also investigated the ability of different motion sensors for human action analysis. To infer the correctness of an action, a score of the abnormality was retrieved from the probability distribution before the SoftMax layer of the proposed GCN model. With a threshold selected, it uses the retrieved score to infer whether an exercise is correctly performed. According to the experimental results on the benchmark dataset named UI-PRMD [14], our method significantly improves the results of [18] in terms of the separation degree. While we also found that the Kinect v2 sensor performed better than the Vicon optical tracking system in terms of both position and angular features.

Although our method shows the potential for detecting abnormal actions with an evaluation score, it might be lack of modelling the expert knowledge in the field of rehabilitation therapy. It means that the involvement of domain knowledge could make such data-driven methods more explainable. In the future, we will focus on this issue by developing more interpretable methods and expand the experiments to real-word exercise data performed by real rehabilitation patients.

REFERENCES

[1] S. R. Machlin, J. Chevan, W. W. Yu, and M. W. Zodet, "Determinants of utilization and expenditures for episodes of ambulatory physical therapy among adults," *Physical therapy,* vol. 91, no. 7, pp. 1018-1029, 2011.

[2] S. A. Jessep, N. E. Walsh, J. Ratcliffe, and M. V. Hurley, "Long-term clinical benefits and costs of an integrated rehabilitation programme compared with outpatient physiotherapy for chronic knee pain," *Physiotherapy,* vol. 95, no. 2, pp. 94-102, 2009.

[3] S. F. Bassett and H. Prapavessis, "Home-based physical therapy intervention with adherence-enhancing strategies versus clinic-based management for patients with ankle sprains," *Physical Therapy,* vol. 87, no. 9, pp. 1132-1143, 2007.

[4] K. N. Karmali, P. Davies, F. Taylor, A. Beswick, N. Martin, and S. Ebrahim, "Promoting patient uptake and adherence in cardiac rehabilitation," *Cochrane database of systematic reviews,* no. 6, 2014.

[5] R. Komatireddy, A. Chokshi, J. Basnett, M. Casale, D. Goble, and T. Shubert, "Quality and quantity of rehabilitation exercises delivered by a 3-D motion controlled camera: A pilot study," *International journal of physical medicine & rehabilitation,* vol. 2, no. 4, 2014.

[6] E. Saraee *et al.*, "Exercisecheck: remote monitoring and evaluation platform for home based physical therapy," in *Proceedings of the 10th International Conference on PErvasive Technologies Related to Assistive Environments*, 2017, pp. 87-90.

[7] J. K. Aggarwal and L. Xia, "Human activity recognition from 3d data: A review," *Pattern Recognition Letters,* vol. 48, pp. 70-80, 2014.

[8] C. Dhiman and D. K. Vishwakarma, "A Robust Framework for Abnormal Human Action Recognition Using $\boldsymbol {\mathcal {R}} $-Transform and Zernike Moments in Depth Videos," *IEEE Sensors Journal,* vol. 19, no. 13, pp. 5195-5203, 2019.

[9] Q. Lei, J.-X. Du, H.-B. Zhang, S. Ye, and D.-S. Chen, "A Survey of Vision-Based Human Action Evaluation Methods," *Sensors,* vol. 19, no. 19, p. 4129, 2019.

[10] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *32nd AAAI conference on artificial intelligence*, 2018.

[11] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Two-Stream Adaptive Graph Convolutional Networks for Skeleton-Based Action Recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12026-12035.

Authorized licensed use limited to: Hong Kong Polytechnic University. Downloaded on June 30,2021 at 07:36:54 UTC from IEEE Xplore. Restrictions apply.

[12] Z. Liu, H. Zhang, Z. Chen, Z. Wang, and W. Ouyang, "Disentangling and Unifying Graph Convolutions for Skeleton-Based Action Recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 143-152.

[13] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531,* 2015.

[14] 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.

[15] A. Elkholy, M. 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,* 2019.

[16] L. Tao *et al.*, "A comparative study of pose representation and dynamics modelling for online motion quality assessment," *Computer vision and image understanding,* vol. 148, pp. 136-152, 2016.

[17] L. L. Presti and M. La Cascia, "3D skeleton-based human action classification: A survey," *Pattern Recognition,* vol. 53, pp. 130-147, 2016.

[18] 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.

[19] N. Sadawi, A. Miron, W. Ismail, H. Hussain, and C. Grosan, "Gesture Correctness Estimation with Deep Neural Networks and Rough Path Descriptors," in *2019 International Conference on Data Mining Workshops (ICDMW)*, 2019: IEEE, pp. 595-602.

[20] M. A. R. Ahad, A. D. Antar, and O. Shahid, "Vision-based Action Understanding for Assistive Healthcare: A Short Review," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops 2019*, 2019, pp. 1-11.

[21] F. Sardari, A. Paiement, and M. Mirmehdi, "View-Invariant Pose Analysis for Human Movement Assessment from RGB Data," in *International Conference on Image Analysis and Processing*, 2019: Springer, pp. 237-248.

[22] J. Antunes, A. Bernardino, A. Smailagic, and D. P. Siewiorek, "AHA-3D: A Labelled Dataset for Senior Fitness Exercise Recognition and Segmentation from 3D Skeletal Data," in *BMVC*, 2018, p. 332.

Authorized licensed use limited to: Hong Kong Polytechnic University. Downloaded on June 30,2021 at 07:36:54 UTC from IEEE Xplore. Restrictions apply.