A Contrastive Graph Convolutional Network for Toe-Tapping Assessment in Parkinson's Disease

Rui Guo, Jie Sun, Chencheng Zhang, and Xiaohua Qian

***Abstract*— One of the common motor symptoms of Parkinson's disease (PD) is bradykinesia. Automated bradykinesia assessment is critically needed for helping neurologists achieve objective clinical diagnosis and hence provide timely and appropriate medical services. This need has become especially urgent after the outbreak of the coronavirus pandemic in late 2019. Currently, the main factor limiting the accurate assessment is the difficulty of mining the fine-grained discriminative motion features. Therefore, we propose a novel contrastive graph convolutional network for automated and objective toe-tapping assessment, which is one of the most important tests of lower-extremity bradykinesia. Specifically, based on joint sequences extracted from videos, a supervised contrastive learning strategy was followed to cluster together the features of each class, thereby enhancing the specificity of the learnt class-specific features. Subsequently, a multi-stream joint sparse learning mechanism was designed to eliminate potentially similar redundant features of joint position and motion, hence strengthening the discriminability of features extracted from different streams. Finally, a spatial-temporal interaction graph convolutional module was developed to explicitly model remote dependencies across time and space, and hence boost the mining of fine-grained motion features. Comprehensive experimental results demonstrate that this method achieved remarkable classification performance on a clinical video dataset, with an accuracy of 70.04% and an acceptable accuracy of 98.70%. These results obviously outperformed other existing sensor- and video-based methods. The proposed video-based scheme provides a reliable and objective tool for automated quantitative toe-tapping assessment, and is expected to be a viable method for remote medical assessment and diagnosis.**

***Index Terms*— Parkinson's disease, toe tapping, video-based assessment, contrastive learning, graph convolutional network.**

I. Introduction

BRADYKINESIA, or the slowness of motion, is one of the main characteristics of Parkinson's disease (PD) [1] that can seriously affect a patient's quality of life, especially the ability to move naturally. Accurate bradykinesia assessment is vital for PD diagnosis and treatment. In clinical practice, neurologists assess the severity of the bradykinesia symptoms using clinical rating scales. The most widely used scale is the Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS) [2], which is also a standard diagnostic tool [3], [4]. Part III of the MDS-UPDRS, i.e., the motor examination part, defines a series of procedures to assess the PD bradykinesia symptoms.

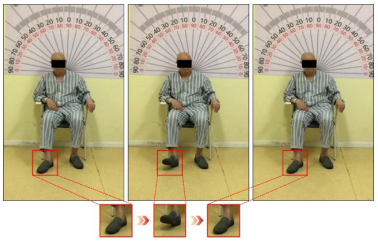

Fig. 1. Execution process of a toe-tapping action.

However, this clinical rating method has two limitations. First, the MDS-UPDRS ratings vary widely among radiologists based on the individual clinical experience of each neurologist [5], and this leads to low rating consistency. Second, the number of neurologists certified for assessing the bradykinesia symptoms is quite limited, while the number of PD patients is steadily increasing, a situation that is not conducive to meeting the increasing demand for bradykinesia assessment. Accordingly, these two limitations may preclude accurate and timely bradykinesia assessment. In particular, after the outbreak of the coronavirus (COVID-19) pandemic in late 2019, automated bradykinesia assessment became more critically needed in order to facilitate remote PD diagnosis and home monitoring, and thereby avoid infection by eliminating unnecessary direct human contact during the pandemic.

The toe-tapping task for lower-extremity bradykinesia assessment is an essential item of the motor examination in the MDS-UPDRS [2], thus constituting one of the vital components of PD diagnosis. The foot angular velocity and other indicators can effectively quantify lower-extremity bradykinesia [6]. In this task, patients are required to tap the ground with their toes at the largest amplitude and fastest speed (Fig. 1). Neurologists then rate the patient response on a scale from 0 to 4 according to the tapping speed, amplitude, pause and change trend. In this paper, we propose an accurate and objective method for automated toe-tapping assessment, in order to improve the reliability and efficiency of PD diagnosis.

Manuscript received 25 April 2022; revised 2 July 2022 and 26 July 2022; accepted 27 July 2022. Date of publication 1 August 2022; date of current version 6 December 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 62171273, in part by the Natural Science Foundation of Shanghai under Grant 22ZR1432100, and in part by the Med-Engineering Crossing Foundation from Shanghai Jiao Tong University under Grant AH0820009 and Grant YG2022QN007. This article was recommended by Associate Editor Y. J. Jung. *(Corresponding author: Xiaohua Qian.)*

Rui Guo and Xiaohua Qian are with the School of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai 200030, China (e-mail: xiaohua.qian@sjtu.edu.cn).

Jie Sun is with the College of Engineering, Carnegie Mellon University, Pittsburgh, PA 15213 USA.

Chencheng Zhang is with the Department of Functional Neurosurgery, Ruijin Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai 200025, China.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCSVT.2022.3195854.

Digital Object Identifier 10.1109/TCSVT.2022.3195854

According to the source of the motor signals, the studies on automated bradykinesia assessment can be divided into three categories: 1) Sensor-based methods: Samà *et al.* [7] extracted gait-related signals measured by a waist-worn triaxial accelerometer, and then estimate bradykinesia severity through regression. Dai *et al.* [8] utilized the parameters of a 6-axis electromagnetic tracking system to train classifiers for quantifying bradykinesia. 2) Optical-device-based methods: Bank *et al.* [9] quantified the motor components of upper-extremity bradykinesia through optical hand tracking. Růžička *et al.* [10] analyzed the motor parameters of finger tapping through a contactless 3D motion-capture system for effectively identifying PD patients. 3) Video-based methods: Liu *et al.* [11] designed a pose estimation algorithm to extract kinematic features and then trained a support vector machine (SVM) to rate three symptoms of upper-extremity bradykinesia. Pang *et al.* [12] extracted features of finger joints through wavelet analysis, and these features were used for effective identification of PD patients.

To the best of our knowledge, there are only three studies on automated toe-tapping assessment. Two of these studies used sensor-based methods, while the third one used a video-based method. Martinez-Manzanera *et al.* [13] collected signals through a nine degrees-of-freedom sensor, and finally achieved a classification accuracy of 62.5-64.5% through a SVM. Kim *et al.* [14] used a gyrosensor to measure ankle movements, and then the correlation coefficient between the motion features and the clinical ratings was found to be in the range of 0.72-0.81. Li *et al.* [15] followed a video-based approach where ankle joint coordinates were extracted to infer task bounding boxes, and dense optical flow was used to capture motion signals. Regression analysis was conducted to predict ratings, and the Pearson correlation between these ratings and the clinician ratings was finally found to be 0.372.

Although the above-mentioned studies provide good performance outcomes or valuable insights, the following difficulties still exist: 1) Sensor placement inconvenience: The sensors need to be fixed on the patients' insteps [13], [14], and this placement may affect patients' actions. Moreover, sensor deployment and calibration require available professional skills. This requirement impedes the popularization and remote implementation of automated assessment tools. 2) Hand-crafted feature engineering: The hand-crafted features were extracted from motor signals, and then the ratings were predicted through machine learning or statistical correlation analysis methods. However, manual design of discriminative features for fine-grained classification is challenging, and even impractical. Alternatively, deep learning schemes can automatically extract task-specific complex features.

In this work, we propose a deep learning scheme for automated quantitative assessment of the toe-tapping task. Nevertheless, the use of deep learning for this automated fine-grained toe-tapping assessment faces the challenge of mining the fine-grained discriminative motion features, which mainly includes the following three aspects:

1) **Low class specificity**: The toe-tapping movement range is relatively small, and the movement differences among different classes are also subtle and highly imperceptible. This results in a low feature specificity for each class.

2) **Low discriminability**: Both the joint position and motion features are important for toe-tapping assessment. However, full exploitation of the discriminability of these features and reducing their redundancy still pose an open problem.

3) **Difficulty in mining motion features**: The toe-tapping task involves repeated execution of a same action. Thus, the spatial-temporal motion feature variability is a primary criterion for neurologists' ratings. Nevertheless, explicit modeling of the remote dependencies and the associated motion features of each joint in the spatial-temporal domain is quite challenging.

To meet the above challenges and deal with them satisfactorily, we constructed class-specific feature spaces, promoted the similarity and discriminability of joint position and motion features, and hence obtained specific and discriminative features for different fine-grained classes. Besides, the spatial-temporal joint relationships during action execution were constructed and captured. Essentially, we propose a contrastive sparse learning framework with a spatial-temporal interaction graph convolutional network (CS-STIGCN) for automated and objective fine-grained toe-tapping assessment. Specifically, our work can be divided into the following stages: 1) We extracted coordinates of human joints from videos using an advanced human pose estimator, and then built a two-stream framework composed of position and motion data; 2) We designed a supervised contrastive learning scheme in which limited prior knowledge is used for aggregating together class-specific features, while setting features of different classes far apart, thereby improving the feature specificity for each class; 3) We proposed a multi-stream joint sparse learning mechanism to reduce the weights of redundant features, and thus strengthen the feature discriminability for all streams; 4) We developed a spatial-temporal interaction graph convolutional module in order to mine long-range spatial-temporal joint dependencies based on feature similarity, and hence achieve effective motion feature modeling through information propagation.

In short, the key goal of this paper is to create a GCN-based architecture for toe-tapping assessment based on PD patient videos. This approach leads to automated MDS-UPDRS rating using only videos captured by consumer-level cameras. Also, this approach enables the realization and popularization of remote and mobile PD assessment and diagnosis. In fact, our technical contribution is to propose a spatial-temporal interaction GCN architecture that combines supervised contrastive learning with multi-stream joint sparse learning, for mining the fine-grained discriminative motion features in toe tapping, as follows:

1) A supervised contrastive learning strategy is designed to strengthen the feature specificity of each class during feature learning. This results in high-quality fine-grained features.

2) A multi-stream joint sparse learning mechanism is proposed to reduce the redundancy and adaptively improve the discriminability of joint position and motion features.

3) A spatial-temporal interaction graph convolutional module is developed to directly model the remote spatial-temporal dependencies for mining fine-grained motion features.

II. Related Work

A. Applications of Deep Learning in PD Motor Assessment

The applications of deep learning in PD motor assessment can be mainly divided into two categories:

1) Video-based joint coordinate extraction [11], [12], [15], [16]: Human pose estimators are typically utilized to extract the coordinates of human joints from videos. Then, relevant features are obtained through feature engineering. Finally, machine learning algorithms are applied for prediction.

2) Motion feature extraction and classification [17]–[23]: Hu *et al.* [17] proposed a vision-based automated freezing of gait (FoG) detection system which includes a novel GCN architecture. This system led to an area-under-the-ROC-curve (AUC) of 0.887. Subsequently, the same authors proposed a graph sequence recurrent neural network, and this achieved an AUC value of 0.90 [18]. Also, a vision-based automated system for rating leg agility was proposed by Guo *et al.* [19], who created a GCN model for adaptive mining of discriminative human spatial relationships. This system had an accuracy of 70.34%. Sabo *et al.* [22] proposed a two-stage training method for GCN models to estimate parkinsonism severity in gait. In addition, a GCN with self-supervised learning was designed for automated arising-from-chair classification with an accuracy of 70.60% [23].

To the best of our knowledge, the toe-tapping classification problem hasn't been addressed with deep learning methods. Such methods can be highly beneficial for exploiting automatically extracted discriminative features (instead of hand-crafted ones) for toe-tapping assessment.

B. Fine-Grained Human Action Recognition

Fine-grained human actions have great inter-class similarities, while different participants and backgrounds show large intra-class variations. Consequently, although the studies of human action recognition have achieved remarkable results [24]–[29], fine-grained human action recognition is still one of the most challenging problems in pattern recognition [30]. Existing methods seek to tackle this problem from multiple perspectives. For example, Zhou *et al.* [31] proposed an interactive-component mining method to model interactions between humans and objects. Singh *et al.* [32] used a multi-stream convolutional neural network (CNN) to create representations of the full-frame and person-centric streams. Then, the long-term context was learned through a bidirectional long short-term memory layer. To utilize prior knowledge on the positions of the human parts, Ma *et al.* [33] used image patches of these parts as CNN inputs, and then encoded the output of the last CNN pooling layer for generating discriminative action descriptors. To reduce the effects of the environmental bias in datasets, Munro *et al.* [34] combined adversarial learning with a self-supervised method, and thus proposed a multimodal unsupervised domain-adaptation strategy for improving the adaptability to unlabeled data.

Although the above-mentioned studies have generally achieved good performance outcomes, the inter-class similarities and intra-class variations are still not fully accounted for during feature extraction. Therefore, we proposed a supervised contrastive learning strategy to constrain the feature extraction process for fine-grained classification. Thereby, features of enhanced specificity were obtained for each class. Besides, a multi-stream joint sparse learning mechanism was developed to enhance the discriminability of the two-stream features through sparse feature selection.

III. Methods

Figure 2 shows the architecture of the proposed contrastive sparse learning framework with the spatial-temporal interaction GCN (CS-STIGCN). First, the skeleton joint coordinates are extracted from videos. Then, a two-stream network with joint position and motion streams is constructed. Each stream is embedded with a spatial-temporal interaction graph convolutional module (ST-IGCM). A supervised contrastive learning (SCL) strategy is applied to the output features of the global average pooling layer of each stream. The weights of two fully connected (FC) layers are constrained by a multi-stream joint sparse learning (MJSL) mechanism.

A. Model Input and Output

Each skeleton joint extracted from videos is represented as a vector with two elements, i.e., x and y coordinates. The position sequences are composed of the two-dimensional (2D) coordinates of these skeleton joints. For example, the position of joint v at time t is expressed as $p_{v_t} = \left(x_{v_t}, y_{v_t}\right)$. The motion sequences refer to the motion (or speed) coordinates obtained by position coordinate differencing among adjacent frames. Formally, the motion of joint v at time t is calculated as $m_{v_t} = \left(x_{v_{t+1}} - x_{v_t}, y_{v_{t+1}} - y_{v_t}\right)$. These two sequences both constitute the model input. Finally, the model outputs the predicted probability values of five classes, and the class with the highest probability value is the assessment result of the model input.

B. Supervised Contrastive Learning (SCL)

In supervised classification tasks, the prior class-specific knowledge can provide rich information for learning discriminative features, and hence improving the model performance. Contrastive learning methods have been developed in many related studies on self-supervised learning [35], where specific pretext tasks are designed without using prior class knowledge to construct positive and negative pairs of input samples. The model learning process evolves to get the positive pairs closer to each other, while setting the negative pairs clearly apart in the embedding space. This leads to representative features for each class.

Based on [35] and [36], we proposed a supervised contrastive learning (SCL, Fig. 3) strategy to fully utilize supervised signals, maximize intra-class feature consistency, and maximize the separation of the feature embeddings of different classes. This strategy further enhances the learning outcomes

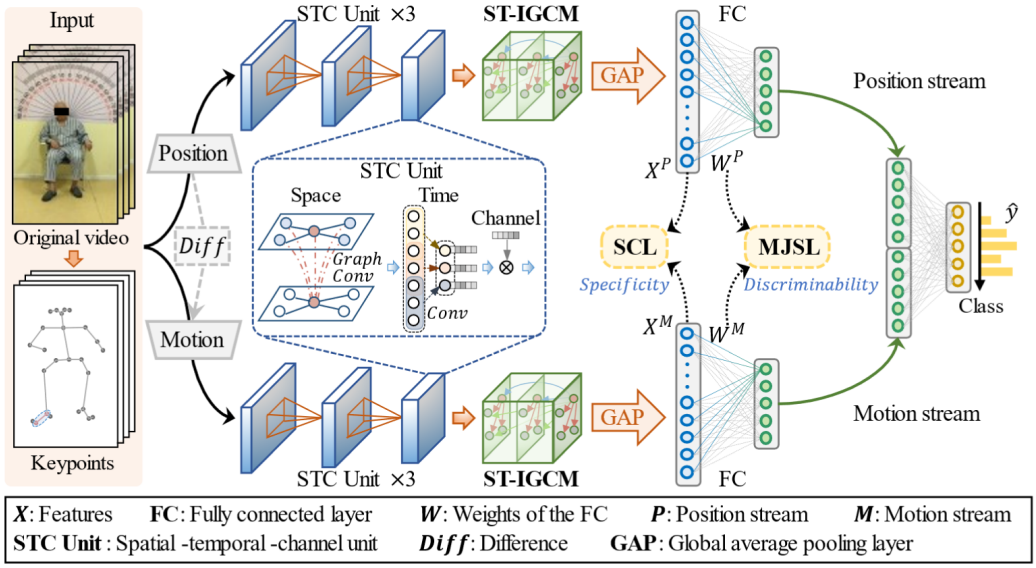

Fig. 2. Architecture of the contrastive sparse learning framework with the spatial-temporal interaction graph convolutional network (CS-STIGCN). It contains three main components: supervised contrastive learning (SCL), multi-stream joint sparse learning (MJSL), and spatial-temporal interaction graph convolutional module (ST-IGCM).

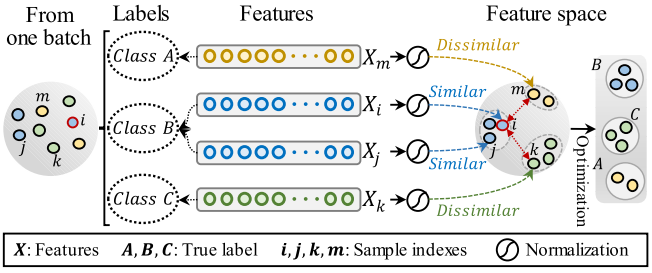

Fig. 3. The process of the supervised contrastive learning (SCL).

for fine-grained class-specific features. Theoretically, the SCL strategy should satisfy the following condition:

$$S(f(X_i), f(X_j)) \gg S(f(X_i), f(X_k)) \tag{1}$$

where X_i and X_j are samples of the same class, which form a positive pair, while X_i and X_k are samples of different classes, which form a negative pair. The symbol f denotes a mapping function, and $S(\cdot,\cdot)$ is a score function to calculate the sample feature similarity. Under ideal conditions, positive pairs show high similarity scores, whereas negative samples have low similarity scores. Therefore, the SCL strategy can guide the model to maximize the intra-class feature consistency.

Specifically, let the high-level feature of the sample extracted by the main network component be represented by the matrix $X^{init} \in \mathbb{R}^{N\times C}$, where N is the batch size, and C is the number of channels. First, the feature vector of each sample (such as $X_i^{init} \in \mathbb{R}^C$, $i \in \{1,\ldots,N\}$) is normalized by the Euclidean norm in the channel dimension:

$$X_{i,c} = \frac{X_{i,c}^{init}}{\max(\sqrt{\sum_{c=1}^{C}\left|X_{i,c}^{init}\right|^2}, \varepsilon)} \tag{2}$$

where ε is a small non-zero value to avoid division by zero. Then, the similarity between sample i and other samples is calculated through an inner product, and the SCL loss function value $(\mathcal{L}_{SCL}^{(i)})$ of the sample i is obtained as:

$$\mathcal{L}_{SCL}^{(i)} = -\frac{1}{N_{y_i}-1}\sum_{j=1}^{N}\delta_{[i\neq j]}\cdot\delta_{[y_i=y_j]} \cdot \log\frac{\exp\left(\langle X_i, X_j\rangle\right)}{\sum_{k=1}^{N}\delta_{[i\neq k]}\cdot\exp\left(\langle X_i \cdot X_k\rangle\right)} \tag{3}$$

where y is the class label of the sample. N_{y_i} is the number of samples in the current batch who are in the same class as sample i, while $\delta_{[\cdot]}$ is an indicator function. If the condition in $[\cdot]$ holds, $\delta_{[\cdot]}$ is equal to 1. Otherwise, it is equal to 0. In (3), only the positive pair similarity is retained, and the minimization of $\mathcal{L}_{SCL}^{(i)}$ corresponds to the maximization of the positive pair similarity. Afterwards, the SCL loss of N samples in the same batch is given by:

$$\mathcal{L}_{SCL} = \frac{1}{N}\sum_{i=1}^{N}\mathcal{L}_{SCL}^{(i)} \tag{4}$$

Because the proposed model adopts a two-stream framework, the SCL strategy is applied to each stream, then the overall SCL loss is the sum of the stream-specific loss terms (P: joint position stream; M: joint motion stream):

$$\mathcal{L}_{SCL} = \mathcal{L}_{SCL(P)} + \mathcal{L}_{SCL(M)} \tag{5}$$

In summary, we designed a SCL strategy for the two-stream framework, where we utilized class label information to construct spaces of discriminative class-specific features for fine-grained classification. This design draws features of the same class closer together, and features of different classes to be far apart, thus enhancing the feature specificity for each class.

C. Multi-Stream Joint Sparse Learning (MJSL)

As mentioned above, high-level spatial-temporal features have been extracted from the main network component. However, since the joint position and motion sequences are obtained from the same video, the two streams can be

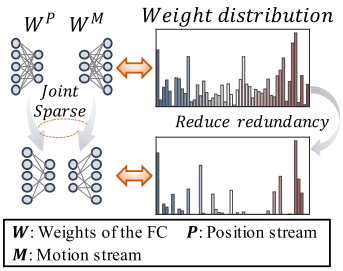

Fig. 4. The diagram of the multi-stream joint sparse learning (MJSL).

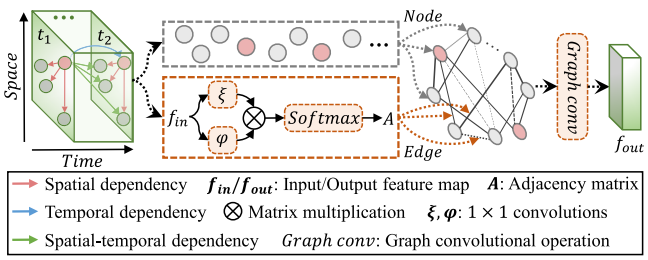

Fig. 5. The structure of the spatial-temporal interaction graph convolutional module (ST-IGCM).

potentially correlated. Moreover, feature redundancy exists in each stream. We propose here a mechanism to eliminate this redundancy and enhance the similarity and discriminability of the joint position and motion features.

In particular, we propose a multi-stream joint sparse learning (MJSL) mechanism, as shown in Fig. 4. The joint sparsity constraint is integrated into the model loss function to constrain the weight matrix of the FC layer of each stream. This mechanism promotes the similarity between different streams and the sparsity within each stream. Let the weight matrices of the FC layers of the S streams be $W_1, W_2, \ldots, W_S \in \mathbb{R}^{C\times K}$, respectively, where C is the number of channels with the FC input feature map, and K is the number of classes. Then, the joint sparsity constraint can be expressed as:

$$\begin{aligned}\mathcal{L}_{MJSL} &= \frac{1}{K}\sum_{k=1}^{K}\sum_{c=1}^{C}\left\|W_{c:}^{(k)}\right\|_2 \\ &= \frac{1}{K}\sum_{k=1}^{K}\sum_{c=1}^{C}\sqrt{\sum_{s=1}^{S}W_{cs}^{(k)2}}\end{aligned} \tag{6}$$

where $W^{(k)} = \left[W_1^{(k)}, W_2^{(k)}, \ldots, W_S^{(k)}\right] \in \mathbb{R}^{C\times S}$ is the result of concatenating the FC weight matrices associated with the $k-$th class for the S streams.

For the FC weights of each class, the MJSL mechanism calculates the ℓ_2-norm in the stream dimension to promote the similarity between the multiple streams. Then, the ℓ_1-norm is calculated in the channel dimension to promote feature sparsity. Finally, the computational results are summarized and averaged for each class. Through the model optimization process, the weights of the connections between the redundant features and the class prediction nodes will be maximally reduced to 0. If the weight values associated with the $c-$th channel and the $k-$th class are reduced to 0, then the associated features are considered redundant. Hence, the prediction outputs of the $k-$th class will not be affected by these features. Formally, the constraining process of each class is similar to the $\ell_{2,1}-$norm method whose effectiveness as a robust feature selection method has been verified in many studies [37].

Therefore, the MJSL-based optimization of the FC weights leads to adaptive elimination of redundant features, and achieves sparse selection of discriminative features, thereby making the learning process more robust and stable.

D. Spatial-Temporal Interaction Graph Convolutional Module (ST-IGCM)

Current methods of skeleton-based human action recognition with GCN mostly follow the scheme proposed by Yan *et al.* [26]. In this scheme, skeleton sequences are modeled with spatial-temporal graphs. Then, high-level spatial-temporal features are gradually extracted through a spatial-temporal graph convolutional network (ST-GCN). The basic ST-GCN unit performs two types of operations: a spatial graph convolution operation and a temporal convolution operation. However, this arrangement hinders the direct exchange of joint information across time and space.

Therefore, we proposed the ST-IGCM to explicitly model complex spatial-temporal remote dependencies in a video (Fig. 5). Let the spatial-temporal graph be expressed as $G = \{V, E\}$. For the given video, all extracted spatial-temporal joints constitute the node set V of that graph. Each edge in the edge set E is weighted by the similarity of the attributes (or features) of the associated nodes. Specifically, the input feature map is denoted by $f_{in} \in \mathbb{R}^{C_{in}\times V}$, where C_{in} and V are the number of channels and the number of nodes, respectively. Then, f_{in} is mapped to the low-dimensional feature space through two 1×1 convolutions (ξ and φ). This yields the intermediate feature maps f_ξ, $f_\varphi \in \mathbb{R}^{C'_{in}\times V}$:

$$f_\xi = W_\xi f_{in} \tag{7}$$

$$f_\varphi = W_\varphi f_{in} \tag{8}$$

Finally, the spatial-temporal similarity adjacency matrix $A \in \mathbb{R}^{V\times V}$ is obtained through the matrix multiplication and $softmax$ normalization as follows:

$$A = softmax(f_\xi^T f_\varphi) \tag{9}$$

where A_{ij} in A represents the interdependency of the nodes v_i and v_j in the spatial-temporal domain. We adopted the graph convolution proposed by Kipf & Welling [38], and further defined the graph convolution operator in the ST-IGCM by

$$f_{out} = A f_{in} W \tag{10}$$

where W is the weight function obtained through a 1×1 convolution. $f_{out} \in \mathbb{R}^{C_{out}\times V}$ denotes the output feature map.

Consequently, the ST-IGCM can calculate and utilize the spatial-temporal similarities of joint features. This leads to the adaptive discovery of remote spatial-temporal dependencies, direct information exchange, and finally extracting discriminative spatial-temporal features.

E. Overall Framework

Based on the joint position and motion sequences, the proposed model consists of joint position and motion streams. In each stream, first, three spatial-temporal-channel units are stacked in each stream, where each unit is composed of a spatial graph convolution module with learnable edge importance weighting [26], a temporal convolution module [26], and a channel-based squeeze-and-excitation block [39] (Fig. 2). Afterwards, the input features of all preceding spatial-temporal-channel units are concatenated along the channel dimension, in order to promote effective feature transmission through dense connections [40]. Next, each stream is equipped with an ST-IGCM, an average pooling layer, and an FC layer. Finally, an additional FC layer is used to integrate the output scores from the two streams.

The total loss of the proposed model is composed of the cross-entropy term ($\mathcal{L}_{CE}$), the SCL term ($\mathcal{L}_{SCL}$), and the MJSL term ($\mathcal{L}_{MJSL}$):

$$\begin{aligned}\mathcal{L} &= \mathcal{L}_{CE} + \lambda_1 \mathcal{L}_{SCL} + \lambda_2 \mathcal{L}_{MJSL} \\ &= -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} y_{i,k} \log \hat{y}_{i,k} \\ &\quad + \lambda_1 \left(\frac{1}{N}\sum\nolimits_{i=1}^{N} \mathcal{L}_{SCL(P)}^{(i)} + \frac{1}{N}\sum\nolimits_{i=1}^{N} \mathcal{L}_{SCL(M)}^{(i)} \right) \\ &\quad + \lambda_2 \left(\frac{1}{K}\sum\nolimits_{k=1}^{K}\sum\nolimits_{c=1}^{C} \sqrt{\sum\nolimits_{s=1}^{S} W_{cs}^{(k)2}} \right) \end{aligned} \tag{11}$$

where y and $\hat{y}$ represent the true label and the predicted label probability for a given sample, respectively, while λ_1 and λ_2 are hyper-parameters to balance the contributions of $\mathcal{L}_{SCL}$ and $\mathcal{L}_{MJSL}$, respectively.

IV. Experiments and Results

A. Dataset Description

The study is approved by the Institutional Review Board of Ruijin Hospital, Shanghai Jiao Tong University School of Medicine in China. We used a video dataset for clinical assessment of PD motor symptoms. This dataset was provided by the Neurosurgery Department of Ruijin Hospital, and included videos collected between 2017 and 2019 for 144 PD patients. The patients were in different PD states and had different motor performance ratings due to differences in the follow-up times, medication conditions, and surgical conditions. Therefore, patient videos were independently captured and rated for each PD state. Each video was shot from a frontal view at a frame rate of 30 frames per second (FPS) with a resolution of 720P (1280×720) or 1080P (1920×1080). We identified short video clips of the patient's left and right feet while performing the toe-tapping task. We then discarded of the video clips that exhibited non-acceptable characteristics, such as camera shaking, nonstandard task execution, and interference during recording. Finally, 691 video clips were available for model construction and validation. Table I shows the class-specific distribution of the MDS-UPDRS scores of these videos.

TABLE I
Class-Specific Distribution of the MDS-UPDRS Scores of Samples

Score	0	1	2	3	4	Total
Number	101	281	165	105	39	691
Proportion	14.6%	40.7%	23.9%	15.2%	5.6%	100.0%

To standardize the model input, the left-foot video frames were horizontally mirrored to right-foot frames. Then, the state-of-the-art human pose estimator, the official BODY_25 model of the OpenPose [41] (version 1.5.1) with its default hyper-parameter settings, was used to extract the joint coordinates in each frame. We empirically set the first 150 frames of each video as the time-series length, and experimentally selected the y-coordinate sequences of the two joints closely related to the toe-tapping action (i.e., the right big toe and the right ankle) as the model input. These sequences were subjected to z-score normalization.

B. Implementation Details and Evaluation Metrics

We utilized the PyTorch library to implement the proposed model. The model training and testing were carried out on an NVIDIA GeForce 1080Ti GPU. During training, we set the batch size to 8, and used the stochastic gradient descent (SGD) algorithm with an initial learning rate of 0.001. The learning rate was set to drop to 1/10 of the previous value after the 80^{th}, 90^{th}, and 100^{th} epochs. A total of 310 epochs were performed in the training phase, which lasted for 0.3 hour.

A patient-based five-fold cross validation (CV) scheme was used to evaluate the proposed model. That is, the patients were randomly divided into five folds. Four folds were used for training while one was used for testing. The training and testing datasets were independent of each other, and the hyper-parameters used in the five experiments were the same. To quantitatively evaluate the model classification performance, six evaluation metrics were used: the accuracy (Acc), balanced accuracy (Balanced Acc) [42], acceptable accuracy (Acceptable Acc), precision (Prec), F1-score (F1), and the AUC value (i.e., the area under the receiver operating characteristic curve), where the balanced accuracy is often used to evaluate the imbalanced dataset by calculating the arithmetic mean of the proportion of correct predictions in each class [42]. Due to the subjective differences of the raters' opinions [43]–[45] and the natural attributes of the MDS-UPDRS [16], the proportion of the samples with a prediction error less than 1 in all samples is defined as the acceptable accuracy [19], [21], which is a widely acceptable metric in clinical practice [46], [47].

C. Classification Performance

Table II shows the classification performance metrics of the proposed method, including the per-class accuracy, balanced accuracy, acceptable accuracy, precision, F1-score, and AUC value, as well as the total accuracy and total acceptable accuracy of all samples. The proposed method achieves a 70.04% total accuracy and a 98.70% total acceptable accuracy, and also achieves acceptable accuracies of more than 89% in all classes. Additionally, Fig. 6(a) illustrates the ROC curve

TABLE II

CLASSIFICATION PERFORMANCE METRICS OF THE PROPOSED METHOD IN 5-FOLD CV EXPERIMENT

	Acc (%)	Acceptable Acc (%)	Prec (%)	Rec (%)	AUC
Score-0	20.79	98.02	72.41	20.79	0.85
Score-1	93.24	100.00	71.20	93.24	0.86
Score-2	78.79	100.00	69.15	78.79	0.90
Score-3	43.81	97.14	67.65	43.81	0.90
Score-4	64.10	89.74	65.79	64.10	0.95
Total Acc (%)		70.04			
Balanced Acc (%)		60.15			
Total Acceptable Acc (%)		98.70			

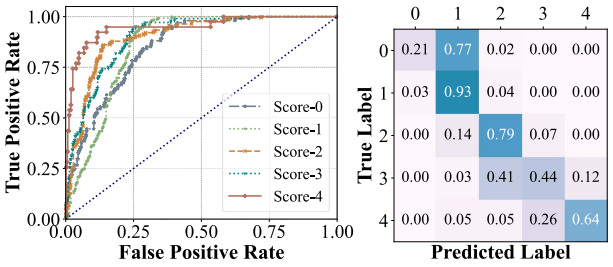

Fig. 6. (a) ROC curve and (b) confusion matrix of classification results.

of each class. Indeed, the AUC value of each class is high while the average AUC value reaches 0.89. The confusion matrix of the classification results is shown in Fig. 6(b). This matrix shows satisfactory performance for each of the scores 1, 2 and 4, while the performance for the scores 0 and 3 is not particularly good. Potential justifications of these results are discussed in Section V. Nevertheless, these results reflect excellent performance of the proposed method in the fine-grained toe-tapping classification task.

D. Comparisons With Other Methods

1) Comparisons With the State-of-the-Art Skeleton-based Action Recognition Methods: We compared the proposed model with five state-of-the-art skeleton-based action recognition models, including a CNN model (The two-stream CNN model [48]) and five GCN models (namely, ST-GCN [26], Js-AGCN [27], Bs-AGCN [27], 2s-AGCN [27], and Motif-STGCN [28]). The comparative results in Table III show that our proposed model achieves the highest accuracy ($p-$value< 0.05) and the best values for the other evaluation indexes. This further demonstrates the superiority of our method in fine-grained toe-tapping classification.

2) Comparisons With Existing Baselines for Video-Based PD Assessment: We also compared the proposed model with other existing video-based PD assessment baselines, including the method combining feature engineering and random forest [15], and the two-stage training method for ST-GCN models [22]. Joint coordinates or dense optical flow was utilized to track trajectories and calculate motion features in the former. We tested the performance of these existing baselines on the toe-tapping dataset, and used the Pearson correlation between predictions and clinical ratings as the evaluation metric. The results in Table IV show that our proposed method obviously outperforms other existing baselines on the toe-tapping dataset.

TABLE III

COMPARATIVE RESULTS WITH FIVE STATE-OF-THE-ART SKELETON-BASED ACTION RECOGNITION MODELS

Models	Total Acc (%)	Balanced Acc (%)	Average Prec (%)	Average F1 (%)
Two-stream CNN [48]	62.08[a]	55.96	59.23	57.28
ST-GCN [26]	55.28[b]	52.01	52.43	52.21
Js-AGCN [27]	58.61[c]	55.13	56.85	55.91
Bs-AGCN [27]	48.63[d]	44.74	46.04	45.26
2s-AGCN [27]	58.18[e]	52.04	56.97	53.92
Motif-STGCN [28]	61.79[f]	57.88	59.61	58.52
Ours	**70.04**[g]	**60.15**	**69.24**	**60.96**

$P_{ag} = 0.0055, P_{bg} = 2.88 \times 10^{-5}, P_{cg} = 0.0027, P_{dg} = 3.08 \times -5, P_{eg} = 0.0003, P_{fg} = 0.0022$ as derived from a T-test.

TABLE IV

COMPARATIVE RESULTS WITH EXISTING BASELINES FOR VIDEO-BASED ASSESSMENT IN PD

Methods	Pearson Correlation
Feature engineering (joints) + Random forest	0.63
Feature engineering (optical flow) + Random forest	0.67
Two-stage training method for ST-GCN models	0.82
Ours	**0.84**

3) Comparisons With Other Studies on Automated Toe-Tapping Assessment: We also investigated three other existing studies on automated toe-tapping assessment (See Table V), and made a comprehensive comparison. Since Kim *et al.* [14] and Li *et al.* [15] used correlation analysis in their evaluation, we also analyzed the correlation between our predicted results and the clinical scores provided by neurologists, and found the Pearson correlation coefficient to be 0.84. Compared with sensor-based methods [13], [14], our scheme achieves the highest accuracy and correlation coefficient in the classification and correlation analysis, respectively. All video-based methods typically utilize a human pose estimator to extract joints from videos. However, Li *et al.* [15] extracted artificially defined features, and then performed regression analysis. In our work, we directly designed a deep learning model to achieve end-to-end feature extraction and classification. The results demonstrate that our scheme has significant advantages and outperforms other related studies on the largest toe-tapping dataset.

E. Ablation Study

To comprehensively evaluate the effectiveness and significance of each component of the proposed method, we systematically performed ablation experiments, and demonstrated the statistical significance of the accuracy improvements brought by each component through a one-tailed paired sample t-test. The experimental results are shown in Table VI, where the baseline architecture is a two-stream network with three spatial-temporal-channel units for each stream. The results demonstrate that the $p-$values are all less than 0.05, and this indicates that the accuracy improvement brought by each component is statistically significant.

1) Supervised Contrastive Learning (SCL): The SCL component can significantly improve the classification performance, and increase the total accuracy of the baseline model by 3.04% (Table VI). To visualize the SCL-induced

TABLE V
COMPARATIVE RESULTS WITH THREE OTHER EXISTING STUDIES ON AUTOMATED TOE-TAPPING ASSESSMENT

Author	Data Sources	Dataset Size	Models	Performance	
				Correlation	Classification Accuracy (%)
Kim *et al*. [14]	Sensors (A gyrosensor)	39 patients, 14 healthy control subjects	Spearman's rank correlation coefficient	0.72-0.81	-
Martinez-Manzanera *et al*. [13]	Sensors (A nine degrees-of-freedom sensor)	25 patients, 10 controls	SVM	-	62.5-64.5
Li *et al*. [15]	Videos	9 patients	Random forest	0.372	-
Ours	Videos	144 patients	CS-STIGCN	0.84	70.04

TABLE VI
ABLATION EXPERIMENTAL RESULTS (SCL: SUPERVISED CONTRASTIVE LEARNING; MJSL: MULTI-STREAM JOINT SPARSE LEARNING; ST-IGCM: SPATIAL-TEMPORAL INTERACTION GRAPH CONVOLUTIONAL MODULE)

Method	Total Acc (%)	Balanced Acc (%)	Average Prec (%)	Average F1 (%)
Baseline	64.98[a]	55.31	62.14	53.03
Baseline + SCL	68.02[b]	58.47	66.39	57.59
Baseline + MJSL	68.16[c]	59.81	65.23	60.90
Baseline + ST-IGCM	68.45[d]	59.86	66.18	60.96
ST-IGCN (Baseline + SCL + MJSL + ST-IGCM)	**70.04**[e]	**60.15**	**69.24**	**60.96**

'a' represents the baseline model. 'b', 'c', and 'd' introduce SCL, MJSL and ST-IGCM into the baseline model, respectively, and 'e' represents the proposed model. $P_{ab} = 0.0417, P_{ac} = 0.0187, P_{ad} = 0.0100, P_{ae} = 0.0046$ as derived from T-tests.

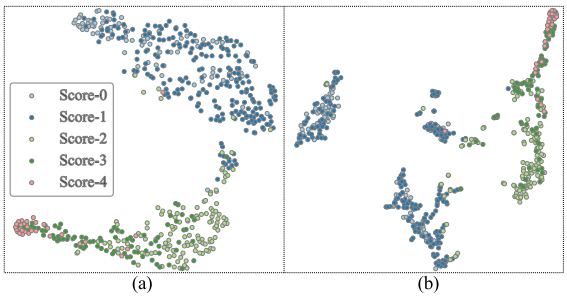

Fig. 7. T-SNE embeddings of sample features: (a) without the SCL; (b) with the supervised contrastive learning.

enhancement of the feature specificity of each class, we used the t-distributed stochastic neighbor embedding (t-SNE) method [49] to analyze the features of each class with or without the SCL component, and finally visualized the feature embedding results (Fig. 7). In comparison with the results obtained without SCL (Fig. 7(a)), to a certain extent, the SCL-equipped model can visually aggregate the feature embeddings of each class (Fig. 7(b)). Thus, the SCL component can aggregate features of the same class and set apart features of different classes. This in turn strengthens the feature specificity of each class, and boosts the fine-grained classification performance.

2) Multi-Stream Joint Sparse Learning (MJSL): The total accuracy of the baseline method increased from 64.98% to 68.16% with the inclusion of the MJSL component, and the balanced accuracy, average precision, and F1-score of each class were also improved (Table VI). To visualize the positive effect of the MJSL component on the FC weights, we created a FC weight distribution map for each class (Fig. 8). Obviously, the weight distribution is more dispersed without the MJSL component, whereas the MJSL inclusion increases the number of weight values close to 0 (indicted in orange in Fig. 8) and thus reduces the weights of redundant features. This illustrates that the MJSL mechanism can effectively promote the FC weight sparsity, and hence adaptively eliminate potential irrelevant features.

3) Spatial-Temporal Interaction Graph Convolutional Module (ST-IGCM): The ST-IGCM improves the total accuracy to 68.45%. Also, the $p-$value obtained by the t-test is $0.0100 (< 0.05)$, and this demonstrates the statistical significance of the improvement made by the ST-IGCM. To illustrate the positive effect of the ST-IGCM on the toe-tapping classification performance, the spatial-temporal similarity matrices of two samples are shown in Fig. 9. Because the toe-tapping task involves repeated execution of the same action, the movement of each joint has an obvious correlation across time and space, that is, the similarity of the joint features is high in the spatial-temporal dimension. The regular patterns of the adjacency matrices in Fig. 9(a) and (b) show that the ST-IGCM directly mines the spatial-temporal toe-tapping data, and further promotes effective modeling and propagation of spatial-temporal features.

F. Reliability and Sensitivity Analysis

1) Reliability Analysis: To demonstrate the stability of the proposed method, we performed 10 repetitions of the 5-fold CV scheme, where the samples were randomly shuffled and divided into folds in each repetition. The distributions of the accuracy and acceptable accuracy metrics are shown as box plots in Fig. 10. The ranges of the accuracy and acceptable accuracy metrics are relatively small, which are 1.16% and 0.43%, respectively. The accuracy reported in this paper is the fifth highest accuracy (70.04%) over the ten repetitions. These results further demonstrate the high reliability of the proposed model.

2) Parameter Sensitivity Analysis: To verify the robustness of the proposed method, we also analyzed the trade-off parameters of the loss function (λ_1 and λ_2), and carried out sensitivity analysis. We determined the optimal parameter settings to be $\lambda_1 = 0.80$ and $\lambda_2 = 0.0080$ through experiments, and used these settings as benchmark values. Then we varied these λ_1 and λ_2 values by 10%, and calculated the variation ratio

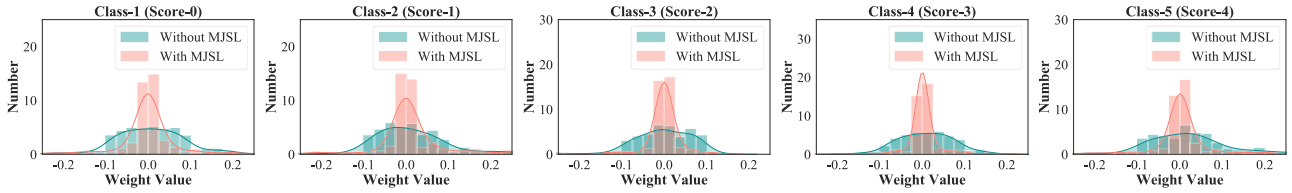

Fig. 8. Sparsity effects of the multi-stream joint sparse learning component.

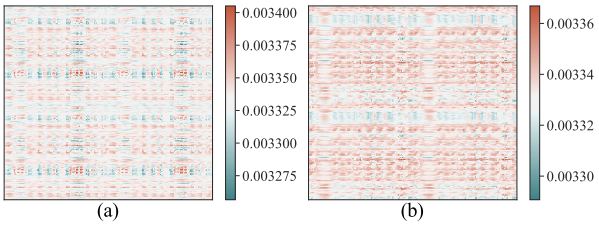

Fig. 9. Spatial-temporal similarity adjacency matrices in the spatial-temporal interaction graph convolutional module. (a) and (b) are from different samples.

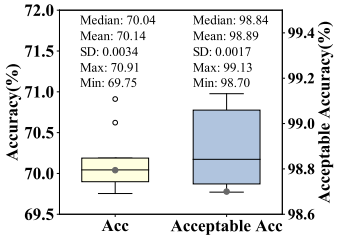

Fig. 10. The distribution of the accuracy and acceptable accuracy metrics. The dots represent the values reported in this paper, and SD is the calculated standard deviation.

of the evaluation indexes as follows:

$$Variation\ Ratio = \frac{Metric - Metric_{[\lambda_1=0.80,\lambda_2=0.0080]}}{Metric_{[\lambda_1=0.80,\lambda_2=0.0080]}} \times 100\% \quad (12)$$

where *Metric* is the selected evaluation index, i.e., accuracy and acceptable accuracy. The results of parameter sensitivity analysis are shown in Fig. 11. More than half (i.e., 13/25) of the absolute accuracy variation ratios are less than 1%, while all (i.e., 25/25) of the absolute acceptable accuracy variation ratios are less than 0.5%. This analysis shows that the proposed method is highly insensitive to the trade-off parameters, and that the SCL and MJSL schemes are clearly robust.

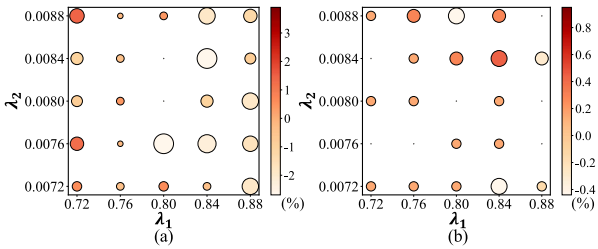

Fig. 11. Parameter sensitivity analysis in loss function (λ_1 and λ_2). (a) and (b) represent the results of accuracy and acceptable accuracy, respectively.

V. Discussion

The fine-grained quantitative assessment of the toe-tapping task is vital for automated bradykinesia assessment in PD patients. Sensor-based motion-capture schemes are unable to achieve large-scale deployment and application of such assessment tools. Also, discriminative features can't be effectively extracted from skeleton sequences with traditional feature engineering methods. Therefore, we designed a skeleton-based GCN method that combines contrastive learning and multi-stream joint sparse learning to realize video-based automated toe-tapping assessment, finally offering one of the indispensable components for PD diagnosis.

The ablation study results (Table VI), reliability analysis (Fig. 10) and parameter sensitivity analysis (Fig. 11) all illustrated the effectiveness and stability of the proposed method. There are three main reasons for the remarkable performance of the proposed method: 1) High class specificity: the supervised contrastive learning strategy promotes the grouping of features associated with the same class and the separation of features of different classes (Fig. 7). This provides class-specific features for fine-grained toe-tapping classification. 2) High discriminability: the multi-stream joint sparse learning mechanism adaptively reduces the weights of the redundant features (Fig. 8), and eventually eliminate such features. As a result, similar discriminative features are obtained in the joint position and motion streams. 3) Mining motion features: the spatial-temporal interaction graph convolutional module explicitly models spatial-temporal joint relationships and remote dependencies in the toe-tapping videos (Fig. 9). This leads to effective fine-grained feature extraction.

On the toe-tapping dataset in this paper, we experimentally determined that the proposed GCN-based method achieved optimal performance with only two key-points. Nevertheless, the spatial graph convolution in this method is essentially different from the standard convolution for the following reasons: first, the spatial graph convolution is performed on the directed graph; thus, its information aggregation is different from that of the standard convolution. Second, the proposed method includes two graph convolution schemes: spatial graph convolution and spatial-temporal graph convolution. Third, in other application scenarios, the number of key-points in the proposed method needs to be determined according to the experiments.

Compared with the existing studies on automated toe taping assessment, our work has mainly the following three characteristics: 1) To the best of our knowledge, our work is the first to address the toe-tapping assessment problem using a GCN scheme in order to achieve end-to-end feature extraction and classification, with supervised contrastive learning and joint sparse learning schemes. In particular, our work solves the challenge of mining the fine-grained discriminative motion

features in this task. 2) Our toe-tapping dataset is the largest one in the open literature. Besides, the accuracy of our proposed method is the highest one ever reported, and the predicted results have the highest correlation with the clinical scores. Thus, the present results are relatively more reliable than those of the existing sensor-based methods. 3) Our data acquisition setup is simple and cheap where we need only to use consumer-level cameras to capture videos of patients as they perform the toe-tapping task. This setup is enough to obtain objective rating results. Although the filtering procedure is performed on our dataset according to the instructions of the MDS-UPDRS, this would not affect the clinical use of the proposed method, as long as the tester follows the standard task execution instructed by the MDS-UPDRS.

Although our proposed model achieves excellent classification results, there is still one major limitation, i.e., the classification accuracies for samples with scores of 0 and 3 are low, while the acceptable accuracies are more than 97%. These results indicate that most of the incorrectly-classified samples were classified into adjacent classes. This can be mainly ascribed to the following three reasons: 1) Our toe-tapping videos come from real clinical data, and they fit real clinical distributions. Hence, there is a certain deviation in the number of samples for each class (in this paper, score 0: 14.6%, score 3: 15.2%). 2) The movement differences between patients with different scores are small. Thus, the toe-tapping task itself belongs to the domain of fine-grained classification. As revealed by the analysis of Fig. 6(b) and Fig. 7, samples with scores of 0 or 3 are easily misclassified into adjacent classes, and this leads to low accuracies. 3) According to our investigation, relatively low accuracies for scores 0 and 3 are reported as well in many studies on the automated assessment of the PD motor function [19], [21], [46]. So, this problem is a common challenge in this kind of work. In conclusion, our proposed model is more suitable for the five-classification severity assessment of PD symptoms rather than the two-classification symptom detection.

Another limitation is that this work only focuses on the single-modal pattern for the convenience of clinical practice and remote assessment. Inspired by Hu *et al.* [50], we will attempt to collect cross-modal data and develop multimodal learning algorithms for improving the model performance in the future work.

VI. Conclusion

In this paper, we proposed a novel GCN-based method to tackle the challenge of mining the fine-grained discriminative motion features, finally achieve video-based automated and objective toe-tapping assessment. First, a supervised contrastive learning strategy was designed to enhance the feature specificity of different fine-grained classes. Then, a multi-stream joint sparse learning mechanism was proposed to improve the discriminability of joint features. Finally, a spatial-temporal interaction graph convolutional module was developed to promoting the modeling of fine-grained motion features across time and space. The comprehensive experimental results on a large clinical video dataset demonstrate the superior performance of our proposed method. Using the videos captured through multimedia devices, our method has a high potential for automated motor function assessment, and is one of the vital components of remote PD diagnosis.

References

[1] J. Jankovic, "Parkinson's disease: Clinical features and diagnosis," *J. Neurol., Neurosurg. Psychiatry*, vol. 79, no. 4, pp. 368–376, 2008.

[2] C. G. Goetz *et al.*, "Movement disorder society-sponsored revision of the unified Parkinson's disease rating scale (MDS-UPDRS): Scale presentation and clinimetric testing results," *Movement Disorders*, vol. 23, no. 15, pp. 2129–2170, 2008.

[3] J. C. Nienstedt *et al.*, "Predictive clinical factors for penetration and aspiration in Parkinson's disease," *Neurogastroenterol. Motility*, vol. 31, no. 3, Mar. 2019, Art. no. e13524.

[4] J. Siuda *et al.*, "Validation of the Polish version of the movement disorder society-unified Parkinson's disease rating scale (MDS-UPDRS)," *Neurolog I Neurochirurgia Polska*, vol. 54, no. 5, pp. 416–425, 2020.

[5] J. L. Palmer, M. A. Coats, C. M. Roe, S. M. Hanko, C. Xiong, and J. C. Morris, "Unified Parkinson's disease rating scale-motor exam: Inter-rater reliability of advanced practice nurse and neurologist assessments," *J. Adv. Nursing*, vol. 66, no. 6, pp. 1382–1387, Apr. 2010.

[6] J.-W. Kim *et al.*, "Analysis of angular velocity during toe tapping for the quantification of the lower limb bradykinesia in patients with idiopathic Parkinson's disease," *Trans. Korean Inst. Elect. Eng.*, vol. 59, no. 11, pp. 2114–2118, 2010.

[7] A. Samà *et al.*, "Estimating bradykinesia severity in Parkinson's disease by analysing gait through a waist-worn sensor," *Comput. Biol. Med.*, vol. 84, pp. 114–123, May 2017.

[8] H. Dai, G. Cai, Z. Lin, Z. Wang, and Q. Ye, "Validation of inertial sensing-based wearable device for tremor and bradykinesia quantification," *IEEE J. Biomed. Health Informat.*, vol. 25, no. 4, pp. 997–1005, Apr. 2021.

[9] P. J. M. Bank, J. Marinus, C. G. M. Meskers, J. H. de Groot, and J. J. van Hilten, "Optical hand tracking: A novel technique for the assessment of bradykinesia in Parkinson's disease," *Movement Disorders Clin. Pract.*, vol. 4, no. 6, pp. 875–883, Nov. 2017.

[10] E. Růžička, R. Krupička, K. Zárubová, J. Rusz, R. Jech, and Z. Szabó, "Tests of manual dexterity and speed in Parkinson's disease: Not all measure the same," *Parkinsonism Rel. Disorders*, vol. 28, pp. 118–123, Jul. 2016.

[11] Y. Liu *et al.*, "Vision-based method for automatic quantification of parkinsonian bradykinesia," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 27, no. 10, pp. 1952–1961, Oct. 2019.

[12] Y. Pang *et al.*, "Automatic detection and quantification of hand movements toward development of an objective assessment of tremor and bradykinesia in Parkinson's disease," *J. Neurosci. Methods*, vol. 333, Mar. 2020, Art. no. 108576.

[13] O. Martinez-Manzanera, E. Roosma, M. Beudel, R. W. K. Borgemeester, T. van Laar, and N. M. Maurits, "A method for automatic and objective scoring of bradykinesia using orientation sensors and classification algorithms," *IEEE Trans. Biomed. Eng.*, vol. 63, no. 5, pp. 1016–1024, May 2016.

[14] J.-W. Kim *et al.*, "Analysis of lower limb bradykinesia in Parkinson's disease patients," *Geriatrics Gerontol. Int.*, vol. 12, no. 2, pp. 257–264, Apr. 2012.

[15] M. H. Li, T. A. Mestre, S. H. Fox, and B. Taati, "Vision-based assessment of parkinsonism and Levodopa-induced dyskinesia with pose estimation," *J. NeuroEng. Rehabil.*, vol. 15, no. 1, pp. 1–13, Dec. 2018.

[16] A. Sabo, S. Mehdizadeh, K.-D. Ng, A. Iaboni, and B. Taati, "Assessment of parkinsonian gait in older adults with dementia via human pose tracking in video data," *J. NeuroEng. Rehabil.*, vol. 17, no. 1, pp. 1–10, Dec. 2020.

[17] K. Hu *et al.*, "Vision-based freezing of gait detection with anatomic directed graph representation," *IEEE J. Biomed. Health Informat.*, vol. 24, no. 4, pp. 1215–1225, Apr. 2020.

[18] K. Hu *et al.*, "Graph sequence recurrent neural network for vision-based freezing of gait detection," *IEEE Trans. Image Process.*, vol. 29, pp. 1890–1901, 2019.

[19] R. Guo, X. Shao, C. Zhang, and X. Qian, "Sparse adaptive graph convolutional network for leg agility assessment in Parkinson's disease," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 28, no. 12, pp. 2837–2848, Dec. 2020.

[20] R. Guo, X. Shao, C. Zhang, and X. Qian, "Multi-scale sparse graph convolutional network for the assessment of parkinsonian gait," *IEEE Trans. Multimedia*, vol. 24, pp. 1583–1594, 2021.

[21] H. Li, X. Shao, C. Zhang, and X. Qian, "Automated assessment of parkinsonian finger-tapping tests through a vision-based fine-grained classification model," *Neurocomputing*, vol. 441, pp. 260–271, Jun. 2021.
[22] A. Sabo, S. Mehdizadeh, A. Iaboni, and B. Taati, "Estimating parkinsonism severity in natural gait videos of older adults with dementia," *IEEE J. Biomed. Health Informat.*, vol. 26, no. 5, pp. 2288–2298, May 2022.
[23] R. Guo, J. Sun, C. Zhang, and X. Qian, "A self-supervised metric learning framework for the arising-from-chair assessment of parkinsonians with graph convolutional networks," *IEEE Trans. Circuits Syst. Video Technol.*, early access, Mar. 31, 2022, doi: 10.1109/TCSVT.2022.3163959.
[24] J. Cheng, Z. Ren, Q. Zhang, X. Gao, and F. Hao, "Cross-modality compensation convolutional neural networks for RGB-D action recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 3, pp. 1498–1509, Mar. 2021.
[25] C. Wu, X.-J. Wu, and J. Kittler, "Graph2Net: Perceptually-enriched graph learning for skeleton-based action recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 4, pp. 2120–2132, Apr. 2022.
[26] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proc. AAAI Conf. Artif. Intell.*, 2018, vol. 32, no. 1, pp. 7444–7452.
[27] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Two-stream adaptive graph convolutional networks for skeleton-based action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 12026–12035.
[28] Y.-H. Wen, L. Gao, H. Fu, F.-L. Zhang, and S. Xia, "Graph CNNS with motif and variable temporal block for skeleton-based action recognition," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 8989–8996.
[29] H. Wu, X. Ma, and Y. Li, "Spatiotemporal multimodal learning with 3D CNNs for video action recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 3, pp. 1250–1261, Mar. 2022.
[30] D. Shao, Y. Zhao, B. Dai, and D. Lin, "FineGym: A hierarchical video dataset for fine-grained action understanding," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2616–2625.
[31] Y. Zhou, B. Ni, R. Hong, M. Wang, and Q. Tian, "Interaction part mining: A mid-level approach for fine-grained action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3323–3331.
[32] B. Singh, T. K. Marks, M. Jones, O. Tuzel, and M. Shao, "A multi-stream bi-directional recurrent neural network for fine-grained action detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1961–1970.
[33] M. Ma, N. Marturi, Y. Li, A. Leonardis, and R. Stolkin, "Region-sequence based six-stream CNN features for general and fine-grained human action recognition in videos," *Pattern Recognit.*, vol. 76, pp. 506–521, Apr. 2018.
[34] J. Munro and D. Damen, "Multi-modal domain adaptation for fine-grained action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 122–132.
[35] A. Jaiswal, A. R. Babu, M. Z. Zadeh, D. Banerjee, and F. Makedon, "A survey on contrastive self-supervised learning," *Technologies*, vol. 9, no. 1, p. 2, Dec. 2020.
[36] P. Khosla *et al.*, "Supervised contrastive learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 18661–18673.
[37] F. Nie, H. Huang, X. Cai, and C. Ding, "Efficient and robust feature selection via joint $\ell_{2,1}$-norms minimization," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 23, 2010, pp. 1813–1821.
[38] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," 2016, *arXiv:1609.02907*.
[39] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.
[40] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *IEEE CVPR*, Jul. 2017, pp. 4700–4708.
[41] Z. Cao, G. H. Martinez, T. Simon, S.-E. Wei, and Y. A. Sheikh, "OpenPose: Realtime multi-person 2D pose estimation using part affinity fields," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 1, pp. 172–186, Jan. 2021.
[42] N. Japkowicz, "Assessment metrics for imbalanced learning," in *Imbalanced Learning: Foundations, Algorithms, and Applications*, H. He and Y. Ma, Eds. Hoboken, NJ, USA: Wiley, 2013, pp. 187–206.
[43] T. H. Turner and M. L. Dale, "Inconsistent movement disorders society-unified Parkinson's disease rating scale part III ratings in the Parkinson's progression marker initiative," *Movement Disorders*, vol. 35, no. 8, p. 1488, 2020.
[44] J. P. Giuffrida, D. E. Riley, B. N. Maddux, and D. A. Heldman, "Clinically deployable Kinesia technology for automated tremor assessment," *Movement Disorders*, vol. 24, no. 5, pp. 723–730, Apr. 2009.
[45] B. Post, M. P. Merkus, R. M. de Bie, R. J. de Haan, and J. D. Speelman, "Unified Parkinson's disease rating scale motor examination: Are ratings of nurses, residents in neurology, and movement disorders specialists interchangeable?" *Movement Disorders*, vol. 20, no. 12, pp. 1577–1584, 2005.
[46] F. Parisi *et al.*, "Body-sensor-network-based kinematic characterization and comparative outlook of UPDRS scoring in leg agility, sit-to-stand, and gait tasks in Parkinson's disease," *IEEE J. Biomed. Health Informat.*, vol. 19, no. 6, pp. 1777–1793, Nov. 2015.
[47] L. Borzi *et al.*, "Smartphone-based estimation of item 3.8 of the MDS-UPDRS-III for assessing leg agility in people with Parkinson's disease," *IEEE Open J. Eng. Med. Biol.*, vol. 1, pp. 140–147, 2020.
[48] C. Li, Q. Zhong, D. Xie, and S. Pu, "Skeleton-based action recognition with convolutional neural networks," in *Proc. IEEE Int. Conf. Multimedia Expo. Workshops (ICMEW)*, Jul. 2017, pp. 597–600.
[49] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 11, pp. 1–27, 2008.
[50] K. Hu *et al.*, "Graph fusion network-based multimodal learning for freezing of gait detection," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Aug. 31, 2021, doi: 10.1109/TNNLS.2021.3105602.

Rui Guo received the B.S. degree in CS from Lanzhou University. She is currently pursuing the Ph.D. degree with the Medical Image and Health Informatics Laboratory, School of Biomedical Engineering, Shanghai Jiao Tong University, under the supervision of Dr. Qian. Her research interests include computer vision and medical image analysis.

Jie Sun received the B.S. degree from the School of Electronic Information and Communication, Huazhong University of Science and Technology, China, and the University of Birmingham, U.K. She is currently pursuing the master's degree in biomedical engineering from Carnegie Mellon University. She was a Research Assistant at the Medical Image and Health Informatics Laboratory, School of Biomedical Engineering, Shanghai Jiao Tong University, under the supervision of Dr. Qian. Her research interests include signal processing and image processing.

Chencheng Zhang received the Ph.D. degree from the Shanghai Jiao Tong University School of Medicine in 2019. He is currently working with the Department of Functional Neurosurgery, Ruijin Hospital, Shanghai Jiao Tong University School of Medicine. He is also a Junior Principal Investigator with the Shanghai Research Center for Brain Science and Brain-Inspired Intelligence. His research interests include neuromodulation and brain–machine interface.

Xiaohua Qian received the Ph.D. degree in EE from Jilin University. He was an Assistant Professor at the School of Biomedical Informatics, The University of Texas Health Science Center at Houston. He was a Research Fellow at the Wake Forest University's School of Medicine. During his doctoral program, he studied Medical Physics at Duke University Medical Center. He is currently an Associate Professor with the School of Biomedical Engineering, Shanghai Jiao Tong University (SJTU), and is also the Founder and the Director of the Medical Image and Health Informatics Laboratory. His primary research interests include computer vision, medical image analysis, and health data mining.