Neurocomputing 441 (2021) 260–271

Contents lists available at ScienceDirect

Neurocomputing

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

Automated assessment of Parkinsonian finger-tapping tests through a vision-based fine-grained classification model

Hao Li [a], Xiangxin Shao [b], Chencheng Zhang [c], Xiaohua Qian [a,*]

[a] School of Biomedical Engineering, Shanghai Jiao Tong University, China
[b] School of Electrical and Electronic Engineering, Changchun University of Technology, China
[c] Department of Functional Neurosurgery, Ruijin Hospital, Shanghai Jiao Tong, University School of Medicine, China

ARTICLE INFO

Article history:
Received 7 September 2020
Revised 9 January 2021
Accepted 1 February 2021
Available online 2 March 2021
Communicated by Zidong Wang

Keywords:
Parkinson's disease
Finger-tapping tests
Fine-grained action recognition
Imbalanced data learning
Three-stream model

ABSTRACT

Movement disorder of Parkinson's disease (PD) is usually quantified by the Movement Disorders Society-sponsored Revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS) to evaluate its severity. However, the lack of well-trained experts and subjective inter-rater variability often limit an effective and objective assessment in clinical practice. Hence, developing an automated assessment method for movement disorders in PD is crucial. Here, we present a novel vision-based fine-grained action recognition model to cope with one of the most critical and challenging tasks in clinical scales: the finger-tapping test. Specifically, we establish a three-stream fine-grained classification network with a Markov chain fusion model to aggregate multi-stream information of the skeleton sequence from finger-tapping test videos. Then, we develop a spatial–temporal attention mechanism to capture rich spatial and temporal long-range dependencies from skeleton data and introduce a symmetric bilinear pooling layer to enrich the local feature representation of each stream's output. Besides, a mini-batch-based balanced algorithm is designed to ensure that the samples in each mini-batch are inter-class balanced, thus mitigating the effect of imbalanced data on neural networks. Finally, our three-stream fine-grained classification network achieved an accuracy of 72.4% and an acceptable accuracy of 98.3% on 157 patients and 744 videos. Extensive experiments further confirm our approach's effectiveness and reliability. This method does not require any wearable device and has excellent potential for remote monitoring of PD patients in the future.

1. Introduction

Parkinson's disease (PD) is the second most common neurodegenerative disorder and its cardinal motor dysfunctions are bradykinesia, tremor, rigidity, and postural instability [1,2]. Levodopa [3] and deep brain stimulation (DBS) [4] are conventional and effective methods for the treatment of PD. Generally, PD patients and neurologists need regularly adjust the treatment regimens according to the clinical symptoms. However, only well-trained experts can assess the severity of PD motor symptoms based on the standardized clinical rating scales, which cause the limitation of low efficiency and subjective inter-rater variability. Motor diaries recorded by patients are also beneficial to the clinical judgment on disease stage, but this tool is often unreliable due to the patient's low compliance, recall bias, and weak self-assessment skill [5]. Moreover, the recent outbreak of the novel coronavirus (COVID-19) [6] has caused extensive travel restrictions, making it more difficult for PD patients to obtain timely clinical evaluation and treatment. Therefore, an automated and objective assessment method for motor evaluation in PD is urgently required for clinical practice and healthcare delivery.

The Movement Disorders Society-sponsored Revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS) [7] is a standard rating scale for measuring the severity of PD motor symptoms and follow-ups. MDS-UPDRS includes various motor examinations such as gait, finger-tapping, and leg agility tests. The severity of each movement is scored on a five-Likert scale: 0 = normal, 1 = slight, 2 = mild, 3 = moderate, and 4 = severe (Fig. 1). Among items in MDS-UPDRS, the finger-tapping test is a meaningful method to assess the disturbances of rhythm formation and bradykinesia of the extremity in PD patients, and it is also the early sign of motor dysfunction in PD [8]. During the finger-tapping test, the patient is instructed to tap the index finger on the thumb ten times, and then experts rate the score by evaluating the speed,

* Corresponding author at: School of Biomedical Engineering, Shanghai Jiao Tong University, China.
E-mail address: xiaohua.qian@sjtu.edu.cn (X. Qian).

https://doi.org/10.1016/j.neucom.2021.02.011
0925-2312/

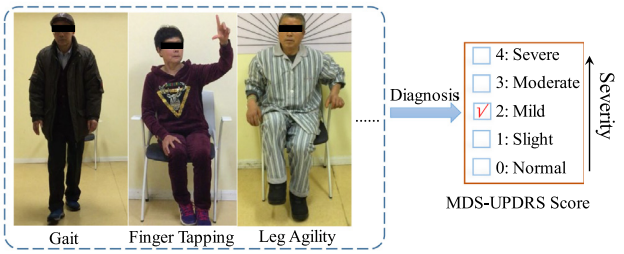


Fig. 1. Examples of motor evaluation for PD patients.

amplitude, and delay [7]. However, the finger-tapping test has been reported as one of the most challenging items to be assessed [9]. Thus, this work aims to construct an automated assessment method for measuring the score of the finger-tapping test, one of the most critical and challenging examinations in the MDS-UPDRS.

The finger tapping test provides conceptual simplicity and highly valuable information, making it a widely used protocol in numerous diseases [10]. As reported in [10,11], the kinematics of the finger tapping test has a strong correlation with bradykinesia and rhythm formation; thus, it is generally used to assess motor dysfunction in dementia research [12], PD research [8], and neurophysiological examinations [13]. For instance, Strauss et al. [14] utilized the finger-tapping task to detect and localize cerebral lesions and further evaluated cerebral lateralization by distinguishing each hand's performance. Shimoyama et al. [15] applied the finger tapping test as a sensitive marker to assess the rhythm formation caused by aging. Hence, it is of great clinical application value to realize automated assessment of the finger-tapping test.

The sensor-based assessment scheme is one of the mainstream approaches in the automated assessment of the motor dysfunction in PD [16]. For instance, Manzanera et al. [17] deployed a nine-degrees-of-freedom sensor to capture the movement trajectory of the finger-tapping test and achieved 68.5% accuracy through an SVM classifier. In [18,19], the body sensor network was successfully deployed to automated assess the severity of leg agility, sit-to-stand, and gait tasks in UPDRS. Recent studies [20,21] proposed several smartphone sensor-based assessment methods to measure motor symptoms in PD, demonstrating the potential of smartphone sensor data in clinical practice. Besides, Pérez-Ibarra et al. [22] and Hssayeni et al. [23] presented a more effective linear classification algorithm and hybrid feature extraction method based on wearable sensor data of PD patients. Sensor-based approaches show promising performance thanks to the precision dynamic signals. However, the deployment of these methods generally required professional skills and specific sensors, making universal wearable device-based assessments infeasible.

The vision-based pose estimation scheme, which can extract the human skeleton position from images and videos through deep learning (DL) networks, has attracted considerable attention in automated assessment of PD patients' movement disorder. For example, Li et al. [24,25] employed the Convolutional Pose Machine to obtain the movement trajectories of individual joints and trained classifiers and regressors to evaluate drinking, communication, leg-agility, toe-tapping tasks, levodopa-induced dyskinesia of PD patients. Liu et al. [26] developed a lightweight pose estimation algorithm and verified their method on finger-tapping, hand-clasping, and hand pro/supination tasks in UPDRS. Guo et al. [27] combined an adaptive graph convolutional network with OpenPose [28] pose estimation algorithm and obtained accurate prediction results of leg agility task in PD. Vision-based pose estimation schemes do not need any additional wearable devices and are convenient for clinical application and remote monitoring of PD patients.

Currently, the skeleton-based action recognition using convolutional neural networks (CNNs) has shown great potential and achieved outstanding results in many action recognition tasks [29,30]. For instance, Du et al. [31] fed skeleton matrices into a CNN and achieved the precision of 91.16% on the ChaLearn gesture recognition dataset. Li et al. [32] proposed an end-to-end two-stream CNN framework to achieve excellent performance in action recognition on the NTU RGB + D dataset with accuracy of 83.2% in cross-subject and 89.3% in cross-view. Based on [32], Hu et al. [30] introduced synchronous local and non-local blocks combined with an attention mechanism in the frequency domain to improve the classification performance. Unlike traditional feature engineering algorithms, CNNs can adaptively learn critical information from skeleton data rather than extract parameters manually. Hence, CNNs are widely used in action recognition tasks like sports scenarios and daily activities [33,34], where the difference in the movements between various categories is prominent, such as tennis and bowling. However, the difference of finger-tapping tests between adjacent scores (e.g., score-0 and score-1 class) is indeterminate for non-professionals. This fine-grained classification problem poses a severe challenge in developing an accurate CNNs based classification model for the finger-tapping test.

In this study, we proposed a novel three-stream fine-grained CNN model for automated assessment of the finger-tapping test in the MDS-UPDRS for PD patients. In particular, we address the following aspects: i) a state-of-the-art pose estimation algorithm OpenPose [28] is used to extract the hand skeleton sequence from the finger-tapping test videos; ii) a skeleton-based three-stream network is constructed to obtain the most discriminative features from the pose, motion, and geometry information; iii) a sequential feature fusion scheme with the Markov chain model is developed to aggregate the multi-stream features effectively; iv) the spatial–temporal attention mechanism and symmetric bilinear pooling technique are embedded into the three-stream network to further improve the ability of fine-grained action recognition; v) a mini-batch-based balanced algorithm is proposed to alleviate the effect of imbalanced video dataset.

The contributions of our proposed method can be summarized as follows:

- We developed a skeleton-based three-stream fine-grained classification network with a Markov chain fusion model. Multi-view features, including location, velocity, and inter-joint relationship information, were designed and input into the three-stream network and then fused by the Markov chain model to predict the MDS-UPDRS score of the finger-tapping test.
- We presented a spatial–temporal attention mechanism to capture rich global spatial and temporal dependencies from skeleton data. And we combined it with symmetric bilinear pooling layers to further improve the fine-grained classification performance of our approach.
- We proposed a mini-batch-based balanced algorithm, which applies a novel sampling method to ensure that each mini-batch is inter-class balanced, thus reducing the effect of imbalanced PD video data.

2. Related work

2.1. Multimodal and Multi-view learning

Multimodal and multi-view learning methods explore the consistency and complementary properties among multiple data sources or feature subsets, thereby obtaining better performance

than single-modality learning methods [35]. Ngiam et al. [36] presented a group of tasks for multimodal learning and successfully demonstrated the advantages of cross-modality feature learning. Yu et al. [37] designed a multi-view hypergraph-based learning method and obtained state-of-the-art image reranking performance. Further, Yu et al. [38] proposed a novel multimodal sparse coding click prediction method for image reranking and finally verified its effectiveness on a large-scale database and real-world datasets. Multimodal and multi-view learning methods have also attracted increasing attention in progression prediction and chronic disease diagnosis [39,40]. For instance, Nie et al. [41] proposed an adaptive multimodal multi-task regression model to predict chronic disease progression of Alzheimer's disease on a real-world dataset. Vasquez-Correa et al. [42] established a novel multi-view learning method by generalized canonical correlation analysis to predict the UPDRS and m-FDS scores of PD patients.

2.2. Classification of imbalanced data

Class-imbalanced data is widespread in clinical practice like finger-tapping tests; it poses a challenge in developing a reliable automated assessment system. Currently, there are three classical approaches to cope with imbalanced data classification: under-sampling, over-sampling, and cost-sensitive learning. Specifically, the under-sampling method aims to balance the sample number in different classes by discarding redundant samples in the majority class; the Easy Ensemble [43] and Balance Cascade [44] are the representative algorithms. By contrast, the over-sampling method attempts to balance the sample number in the classes by creating new samples in the minority class, such as the SMOTE algorithm [45]. Additionally, cost-sensitive learning strategies, like focal loss [46], address the imbalanced data problem at the algorithm level by considering the misclassification of minority class samples to cause higher costs than that of majority classes.

3. Method

Fig. 2 presents the schematic view of our proposed three-stream network. First, the hand skeleton sequences are extracted from the finger-tapping test videos of PD patients by the pose estimation algorithm. Based on the skeleton data, the pose feature S, motion feature M, and geometry feature G of the hand movement are designed and fed into a three-stream network. The three-stream network is composed of CNN extractors, spatial–temporal attention modules, and symmetric bilinear pooling layers, with the purpose of achieving competitive fine-grained classification accuracy. The Markov chain model is applied to fuse the multiple features from the multi-stream network effectively. Finally, a mini-batch-based balanced algorithm is established for multi-classification learning on inter-class imbalanced data. In the following subsections, we expand the proposed method in detail.

3.1. Pose estimation

In this study, we use OpenPose [28] to estimate the hand skeleton sequence from finger-tapping tests videos of PD patients. OpenPose is the state-of-the-art pose estimation DL framework, and numerous studies have been carried out based on OpenPose. The specific process of hand pose estimation in OpenPose is as follows: 1) identify the skeleton of the human body, 2) generate the hand bounding box proposal with the arm joints, and 3) estimate the hand skeleton based on the hand bounding box. As shown in Fig. 3, the hand skeleton data obtained from OpenPose include 21 joints, which will be used as the input of the proposed DL model to predict the MDS-UPDRS score of finger-tapping tests.

3.2. Skeleton-based Three-stream classification network

3.2.1. Three-stream feature design

The pose and motion features are the most basic input information in popular skeleton-based DL networks [29,32]. Pose features represent the location of joints in hand skeleton data, that is, the spatial coordinate of each joint. Suppose that each video data has T frames and hand skeleton data has V joints. Then, the pose feature of t^{th} frame can be denoted as $S_t = \{J_1^t, J_2^t, \cdots, J_V^t\}$, where J_i^t is the spatial coordinate of the i^{th} joint. In addition, a motion feature is the temporal inter-frame variation of the skeleton data, that is, the velocity of hand movement. The motion feature of frame t can be formulated as $M_t = S_{t+1} - S_t = \{J_1^{t+1} - J_1^t, J_2^{t+1} - J_2^t, \cdots, J_V^{t+1} - J_V^t\}$. Then, the pose and motion features can be formulated as,

$$X_s = (S_1, S_2, \cdots, S_T) = \left(\left\{J_1^1, J_2^1, \cdots, J_V^1\right\}, \cdots, \{J_1^1, J_2^1, \cdots, J_V^1\}\right) \tag{1}$$

$$X_m = (M_1, M_2, \cdots, M_T, 0) = (S_2 - S_1, \cdots, S_T - S_{T-1}, 0) \tag{2}$$

$$= \left(\left\{J_1^2 - J_1^1, \cdots, J_V^2 - J_V^1\right\}, \cdots, \left\{J_1^T - J_1^{T-1}, \cdots, J_V^T - J_V^{T-1}\right\}, 0\right)$$

where $\{X_s, X_m\} \in \mathbb{R}^{T\times V\times C}$ and C is the coordinate dimension of joints (e.g., $C = 2$ for 2D coordination). To achieve a uniform temporal length of X_s and X_m, the zero-padding operation is introduced to fill the last frame of the motion feature.

Inspired by [47], we introduce the geometry feature as the third stream input of our network. The geometry feature represents the inter-joint relationship of skeleton data, which is location-invariant and viewpoint-invariant. Fig. 4 shows the extraction process of the geometry feature. Specifically, we calculate the Euclidean distance between each joint pair and record the result as a Euclidean distance matrix, defined as $E \in \mathbb{R}^{V\times V}$. As E is a symmetric matrix (i.e., $E = E^T$), its information is redundant, so we only take the lower triangular matrix, denoted as G, as the valid information. Similarly, G_t represents the geometry feature of frame t, which can be formulated as,

$$G_t = \begin{bmatrix} \|J_2^t J_1^t\| & & \\ \vdots & \ddots & \\ \|J_V^t J_1^t\| & \cdots & \|J_V^t J_{V-1}^t\| \end{bmatrix} \tag{3}$$

$$X_g = \{G_1^*, G_2^*, \cdots, G_T^*\} \tag{4}$$

where $\|\cdot\|$ is the Euclidean distance between two joints and $X_g \in \mathbb{R}^{T\times[v(v-1)/2]}$ is the combination of geometry features of all T frames. During combination, the geometry feature of each frame is flattened into a one-dimension vector G^*; then, these one-dimension vectors of all T frames are combined into X_g.

3.2.2. Sequential feature fusion method with Markov chain

Multiple features are typically combined by summation or concatenation operations. Instead, in this study, we employ a sequential fusion method with a Markov chain. The Markov chain model introduces an implicit regularization, rendering the network more robust to the over-fitting problem [48].

Specifically, given the input sequence $X = \{X_g, X_s, X_m\}$ and the output sequence $Y = \{Y_g, Y_s, Y_m\}$, where $Y_g, Y_s, Y_m \in \mathbb{R}^5$ are the probability map of 5 MDS-UPDRS scores predicted from geometric, pose, and motion stream, respectively. In a Markov chain, the probability of correct prediction can be formulated as,

$$P(Y|X) = P(Y_g|X)P(Y_s|X, Y_g)P(Y_m|X, Y_g, Y_s) \tag{5}$$

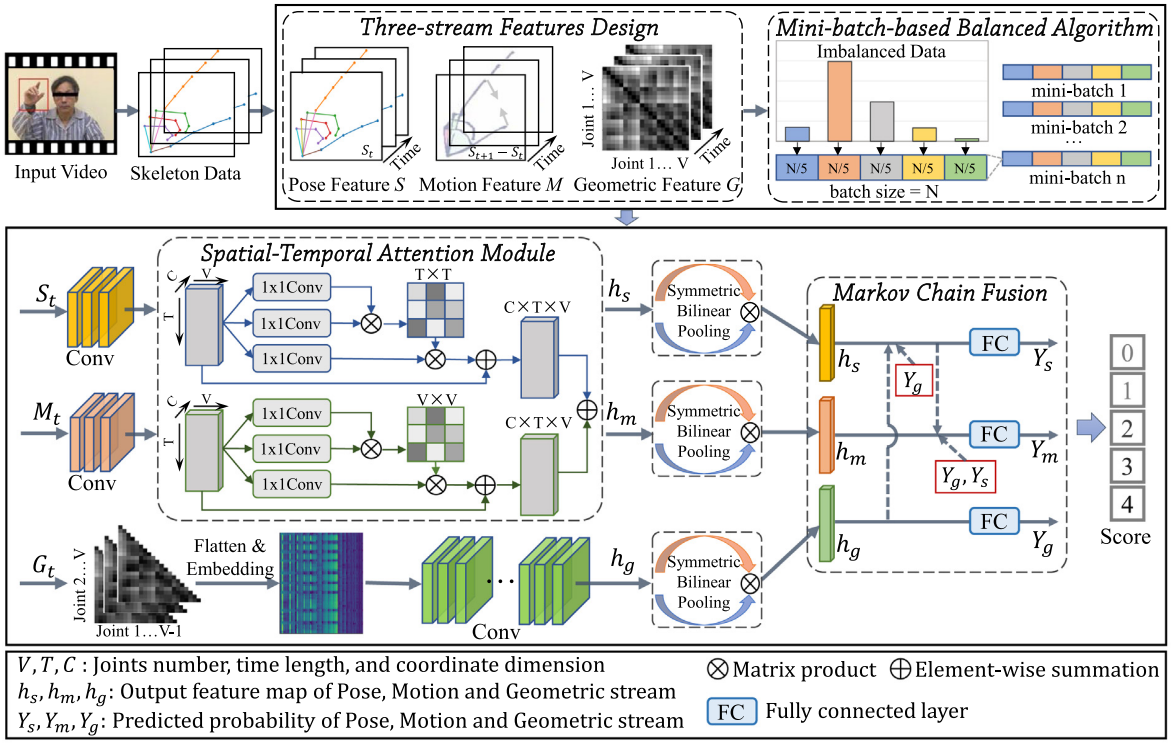


Fig. 2. Overall process flow of our proposed approach. Given pose, motion, and geometric features extracted from raw skeleton data, the three-stream fine-grained classification network will fuse the multiple flows by the Markov chain fusion model to obtain the final prediction. Besides, with the mini-batch-based balanced algorithm, spatial–temporal attention module, and bilinear pooling layer, our model can effectively cope with the challenges of data imbalance and fine-grained classification.

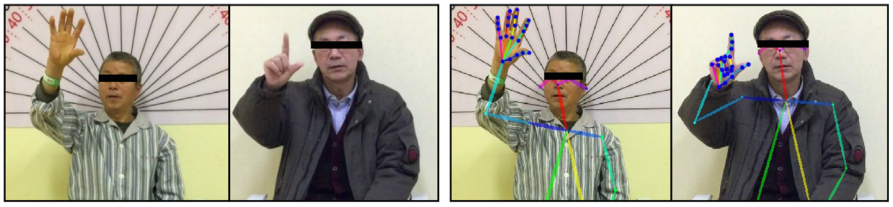


Fig. 3. Performance of OpenPose algorithm on Parkinsonian finger-tapping video data.

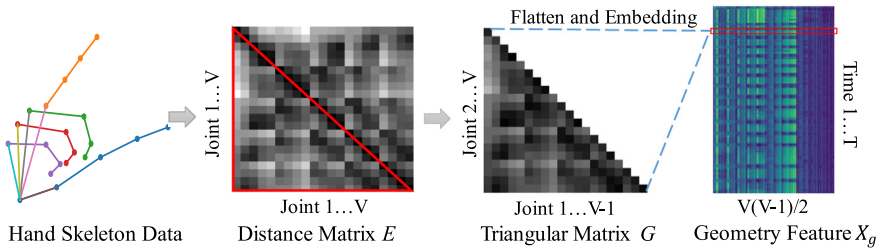


Fig. 4. Process of construction of a geometry feature. Euclidean distance matrix E is generated to capture the inter-joint relationship of hand skeleton data, and then its lower triangular matrix G is flattened and merged in the time dimension to obtain the geometry feature X_g.

Our purpose is to optimize the output sequence Y to the maximum correct predicted probability $P(Y|X)$. Fig. 5 illustrates the process of Markov chain feature fusion in the network architecture, in which the prediction process of geometry, pose, and motion feature streams are formulated as,

$$P(Y_g|X) = \sigma(f_g(X_g))$$

$$P(Y_s|X, Y_g) = \sigma(f_s[X_s, f_g(X_g)], P(Y_g|X)) \quad (6)$$

$$P(Y_m|X, Y_g, Y_s) = \sigma(f_m[X_m, f_g(X_g)], f_s(X_s)P(Y_s|X))$$

where $f_{(\cdot)}$ is the CNN feature extractor, and σ is the softmax activation function. The pseudocode of the feature fusion method with the Markov chain is shown in Algorithm 1. In the pseudocode,

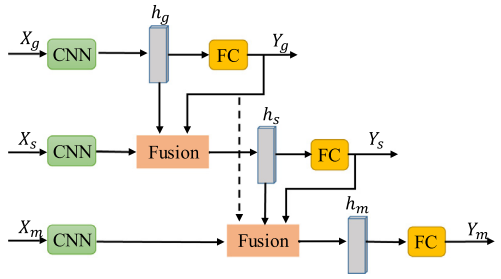


Fig. 5. Multi-stream feature fusion with a Markov chain.

$\mathrm{CNN}_{(\bullet)}$ refers to the feature learning module based on CNNs, ReLU refers to the nonlinear unit, $\mathrm{FC}_{(\bullet)}$ refers to the fully connected layer, and the softmax activation function is applied to obtain the predicted probability map of the output. For three predicted outputs of the Markov chain model, we calculate three loss values separately and then aggregate them for the back-propagation of our proposed three-stream network. The final automated classification result is obtained by considering the probability maps of the three outputs.

Algorithm 1: *Multi-Stream Fusion with Markov Chain*

Input: Geometry Feature X_g; Pose Feature X_s; Motion Feature X_m;
Output: prediction probability $P(Y|X)$;
$h_g = ReLU(CNN_{geometry}(X_g))$;
$P(Y_g|X) = \mathrm{softmax}(FC_{geometry}(h_g))$;
$h_s = ReLU([CNN_{pose}(X_s), h_g, Y_g])$;
$P(Y_s|X, Y_g) = \mathrm{softmax}(FC_{pose}(h_s))$;
$h_m = ReLU([CNN_{motion}(X_m), h_s, (Y_g, Y_s)])$;
$P(Y_m|X, Y_g, Y_s) = \mathrm{softmax}(FC_{motion}(h_m))$;
$P(Y|X) = P(Y_g|X)P(Y_s|X, Y_g)P(Y_m|X, Y_g, Y_s)$;
Return: $P(Y|X)$;

3.3. Spatial-Temporal attention module and bilinear pooling layer for Fine-grained classification

The attention mechanism is often used to capture the rich long-range dependencies from image data to improve the accuracy of classification or segmentation. In this work, we propose a spatial–temporal attention mechanism composed of two parts: a spatial attention module (SAM) and a temporal attention module (TAM). The SAM and TAM are, respectively, used to capture the spatial and temporal global dependencies of pose and motion features. Fig. 6 illustrates the structure of the spatial–temporal attention module, where $H_{in} \in \mathbb{R}^{C\times T\times V}$ is the feature map generated by the CNN extractor, and H_{in} is fed into the SAM and TAM separately. Then, these features from the two attention modules are aggregated by an element-wise summation operation to obtain the attention feature map output H_{out}, as follows,

$$H_{out} = SAM(H_{in}) + TAM(H_{in}) \tag{7}$$

In the TAM, the feature map H_{in} is first fed into two convolution layers to generate new feature maps A and $B \in \mathbb{R}^{C\times T\times V}$, where C is the number of feature channels. Then, A and B are reshaped and transposed to $\mathbb{R}^{CV\times T}$, and then a matrix multiplication is applied between the transpose of A and B. Next, a softmax activation function is performed to acquire the temporal attention map $\beta \in \mathbb{R}^{T\times T}$, which can be formulated as,

$$\beta_{ji} = exp(A_j^{\mathrm{T}}\cdot B_i)/\sum_{i=1}^{T} exp(A_j^{\mathrm{T}}\cdot B_i) \tag{8}$$

where β_{ji} is the impact of the i^{th} frame on the j^{th} frame in the temporal dimension. The larger value of β_{ji}, the stronger is the correlation between the i^{th} and j^{th} frames. Meanwhile, H_{in} is fed into another convolution layer to obtain feature map D, and it is reshaped to $\mathbb{R}^{CV\times T}$. Then, D is multiplied with the transpose of β, and the result is reshaped to $\mathbb{R}^{C\times T\times V}$. Finally, it is summed with H_{in} to obtain the output of the TAM by the following formula,

$$o_j = \lambda\sum_{i=1}^{T} exp(\beta_{ji}\cdot D_i^{\mathrm{T}}) + H_{in,j} \tag{9}$$

where λ is the coefficient of the attention feature map, which can be updated adaptively. As Eq. (9) indicates, the output of each frame is a weighted sum of temporal attention features of all frames and original features. Thus, the TAM can adaptively capture the global dependencies of the time dimension.

Similarly, in the SAM, analogous calculations are performed to capture the long-range spatial dependencies from different hand joints of skeleton data, as shown in the lower part of Fig. 6. Different from the TAM, in the SAM, the new features A, B, and D generated through the convolution layers are reshaped to $\mathbb{R}^{CT\times V}$ to obtain the spatial attention map $\beta \in \mathbb{R}^{V\times V}$. The formulas of the SAM are as follows,

$$\beta_{ji} = exp(A_j^{\mathrm{T}}\cdot B_i)/\sum_{i=1}^{V} exp(A_j^{\mathrm{T}}\cdot B_i) \tag{10}$$

$$o_j = \lambda\sum_{i=1}^{V} exp(\beta_{ji}\cdot D_i^{\mathrm{T}}) + H_{in,j} \tag{11}$$

where β_{ji} is the impact of the i^{th} joint on the j^{th} joint of the hand skeleton data and the output of each joint is a weighted sum of the spatial attention features of all joints and original features. We take advantage of the combination of the TAM and SAM to improve the performance of our proposed three-stream hand action recognition network.

In addition, a symmetric bilinear pooling layer [49] is adopted to enrich the local feature representation of each stream in the three-stream network, as shown in Fig. 7. In this study, the symmetric bilinear pooling layer is placed at the end of each stream. Specifically, the output of the spatial–temporal attention module H_{out} is reshaped to $\mathbb{R}^{C\times VT}$, and it is multiplied with the transpose of itself to obtain the bilinear vector. This process can be formulated as $B = H_{out}H_{out}^{\mathrm{T}}$, where $B \in \mathbb{R}^{C\times C}$ represents the output bilinear vector. Finally, the bilinear vector is flattened as the input of the following Markov chain fusion module.

3.4. Mini-batch-based balanced algorithm

The classical mini-batch-based training process of DL networks is as follows: 1) All training data are shuffled randomly and assigned to multiple mini-batches. 2) The DL network considers one mini-batch once to update the network parameters with a back-propagation algorithm until all mini-batches are executed. This process is defined as an epoch. 3) A DL network usually needs to be trained for many epochs for convergence to the ideal state. However, the classical mini-batch-based training method cannot tackle the imbalanced data problem. When the training dataset is imbalanced, each mini-batch is inevitably dominated by the majority class, enabling networks to easily ignore the minority class, resulting in biased classification results.

Fig. 8 illustrates our proposed novel mini-batch-based balanced algorithm. This algorithm employs an inter-class balanced sampling strategy to generate mini-batches, ensuring an equal sample size of each class in a mini-batch. Specifically, we first group all training data according to the five MDS-UPDRS scores and then randomly shuffle each group's samples. Then, for each

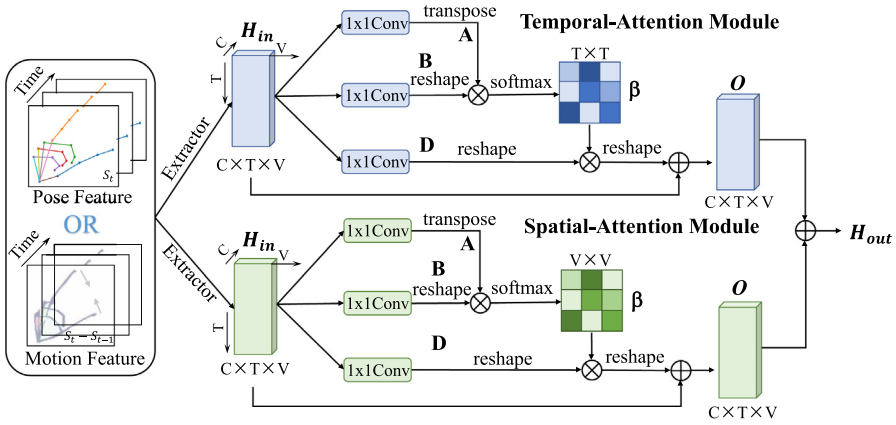


Fig. 6. Schematic of the spatial–temporal attention module. The upper branch is the detailed architecture of the temporal attention module, and the lower branch is the architecture of the spatial attention module.

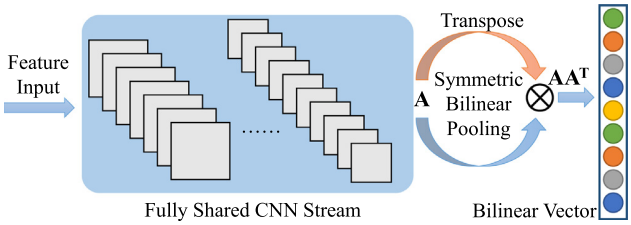


Fig. 7. Schematic of a fully shared symmetric bilinear pooling module.

mini-batch, we randomly select an equal number of samples from each class group and place them into mini-batch; the proportion of each class is one-fifth of the batch size. We use non-repeated sampling to avoid duplicate data in a mini-batch. When a class group is empty, we restore it to the original size and reshuffle the samples in it. This mini-batch-based balanced algorithm does not use the traditional epoch process. Instead, it continuously generates mini-batches with balanced data for the back-propagation algorithm until the parameters converge to the optimal state.

4. Experimental result

4.1. Video dataset of Finger-tapping tests

4.1.1. Dataset collection

This retrospective study was approved by the IRB of Ruijin Hospital, Shanghai Jiao Tong University School of Medicine. The clinical video data of finger-tapping tests were collected from 2017 to 2019 from the department of neurology of Ruijin Hospital.

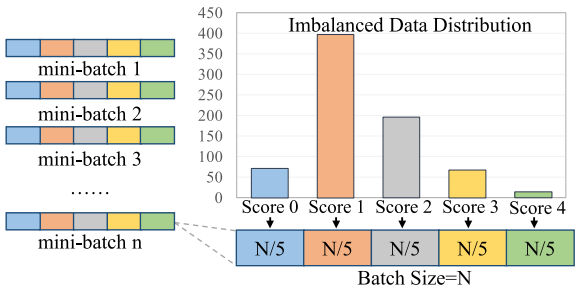


Fig. 8. Illustration of mini-batch-based balanced algorithm.

As shown in Table 1, for 157 participants (103 male and 54 female) in the dataset, the mean age was 63.8 ± 8.7 years. The mean disease duration was 9.4 ± 3.9 years. The MDS-UPDRS motor score was 55.4 ± 15.9 and 35.9 ± 13.6 in off- and on-medication state, respectively. The median value of the modified Hoehn and Yahr Scale was 2.5 (interquartile range: 2.0–3.0). We excluded the participants with severe levodopa-induced dyskinesia during motor evaluation. Generally, all video recordings consisting of finger tapping tests could be included in the analysis; we only excluded those poor video samples in two situations: 1) the target was out of camera range; 2) there was interference from other people except for participants.

Table 2 shows the sample distribution of the finger-tapping video dataset, including numbers and proportions of different MDS-UPDRS score classes. During video recording, each subject was required to sit on a chair and face the camera. The frame rate of the recorded video is 30 frames per second (FPS), and the resolution is 1280 × 720 (720P). The video data of left and right hands were clipped separately, leading to a total of 744 video data. Most of the subjects completed the finger-tapping test in around 6 s. Considering the balance between the classification performance and computational efficiency, we empirically determined the first 5 s (150 frames) as the basic input unit. To the best of our knowledge, this video dataset is the largest clinical dataset of the finger-tapping test in MDS-UPDRS.

4.1.2. Data preprocessing

In the data preprocessing stage, the video frames of left hands were transformed into those of right hands through a horizontal mirror flipping operation to realize a uniform output of the pose estimation algorithm. Then, the pose estimation algorithm was used to acquire hand skeleton data from the videos. The Savitzky-Golay filter [50] was adopted to smooth the raw skeleton data to alleviate noise interference. Subsequently, we set the wrist joint as the coordinate origin to normalize and standardize the

Table 1
Demographics and clinical information of PD patients in our finger-tapping dataset.

Characteristics	Value
Age (mean ± SD)	63.8 ± 8.7
Sex, Male/Female	103/54
Disease duration, y (mean ± SD)	9.4 ± 3.9
Off-medication state MDS-UPDRS score (mean ± SD)	55.4 ± 15.9
On-medication state MDS-UPDRS score (mean ± SD)	35.9 ± 13.6
Modified Hoehn and Yahr Scale [median (IQR)]	2.5 (2.0–3.0)

Table 2
Data distribution on five MDS-UPDRS scores (0–4).

Score	0	1	2	3	4	Sum
Number	70	397	196	67	14	744
Proportion	9.4%	53.4%	26.3%	9.0%	1.9%	100%

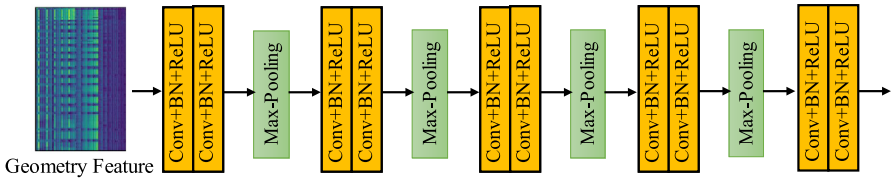


Fig. 9. CNN feature extractor for the geometry feature stream.

coordinates of other hand joints such that the effect of different shooting distances could be eliminated.

4.2. Evaluation criteria

Experimental results were verified by 4-fold cross-validation (CV). The testing data and training data were completely independent, and the hyper-parameter setting of 4-fold CV was the same. The results were obtained as the average of the 4-fold CV. The accuracy (ACC), acceptable accuracy, precision, recall, and F1-score values were used to evaluate the effectiveness and robustness of our proposed method.

Acceptable accuracy refers to the percentage of those cases whose absolute classification bias to label is no more than 1 [18,51], which is calculated as follows,

$$\text{AcceptableACC} = \frac{N_{|\widehat{y}-y|\le 1}}{N} \times 100\% \tag{12}$$

where N is the sample number, $\widehat{y}$ is the predicted score and y is the manual label. Due to the subjectivity and inter-neurologist variability, different raters' MDS-UPDRS score is often non-homogeneous, especially for adjacent score classes. Thus, this criterion can be acceptable and comparable with the inter-rater variability in clinical practice [51]. In addition, the receiver operating characteristic (ROC) curve, area under the curve (AUC), cumulative distribution functions (CDFs), and confusion matrix were adopted to evaluate the predicted results.

4.3. Implementation details

The proposed method was implemented on PyTorch with an NVIDIA GeForce GTX 1080Ti (11 GB). Two types of CNN feature extractors were used in the architecture of the three-stream network: one for the pose feature stream and motion feature stream, and the other for the geometry feature stream. The extractors of pose and motion feature streams consisted of a transform module and convolutional layers [32], and the transform module was implemented by a fully connected layer. The extractor of the geometry stream was a simplified version of the VGG net [52]. We cropped the VGG net to decrease the network parameters while ensuring high classification accuracy, as shown in Fig. 9. We set the batch size to 8, the number of iterations to 3.2×10^5 and the learning rate to $1e^{-4}$. The training time for each CV of this end-to-end three-stream network was about 4.5 h.

4.4. Qualitative and quantitative analysis

4.4.1. Evaluation of classification results

We conducted extensive experiments to evaluate the effectiveness of our proposed approach. Fig. 10 compares our proposed three-stream CNN with backbone two-stream CNN [32]. Our proposed method achieved an accuracy of 72.4% and an acceptable accuracy of 98.3%, which is much higher than the backbone method. The Three-Stream network, Spatial-Temporal Attention module, and Bilinear pooling layer all play a critical role in improving accuracy and lowering the standard deviation. Table 3 lists the specific acceptable accuracy, AUC, precision, recall, and F1-score of the final predicted results on 5 MDS-UPDRS scores.

Fig. 11 illustrates the CDFs analysis, ROC curve, and confusion matrix of predicted results. The CDFs analysis verified our method's superiority over the backbone method, and we obtained an accuracy of above 98% when the cumulative error is 1. In addition, the ROC curves of the 5 MDS-UPDRS scores all have satisfactory AUC values, and the confusion matrix shows high accuracy for each score in the acceptable range. These results reflect the excellent fine-grained classification performance of our proposed method.

To measure the stability of our proposed model, we implemented a random 4-fold CV experiment by 10 times; this means randomly shuffling samples index to generate different 4-fold split at each time. Fig. 12 illustrates the results of 10 repetitions, the fluctuations of accuracy (from 71.37% to 73.25%) and acceptable accuracy (from 97.92% to 98.59%) are all small, and the standard deviations are 0.56% and 0.22%, which demonstrate the robustness of our approach.

4.4.2. Effectiveness of Three-stream network

In this study, we expanded the traditional two-stream network to a three-stream network by introducing the geometry feature; thus, the accuracy of automated assessment of finger-tapping tests

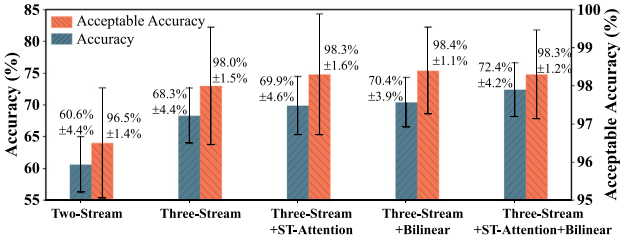


Fig. 10. Comparison our proposed method with backbone two-stream CNN.

Table 3
Classification results of the proposed method.

	Acceptable ACC (%)	AUC	Precision (%)	Recall (%)	F1 (%)
Score-0	95.7	0.7788	37.1	18.6	24.8
Score-1	100.0	0.8441	77.4	86.4	81.7
Score-2	99.5	0.8779	69.1	73.0	71.0
Score-3	89.6	0.9045	61.5	47.8	53.8
Score-4	85.7	0.9761	71.4	35.7	47.6

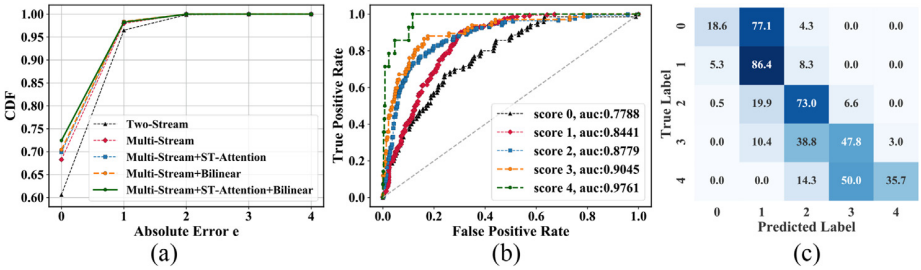


Fig. 11. CDFs analysis, ROC curve, and confusion matrix of predicted results.

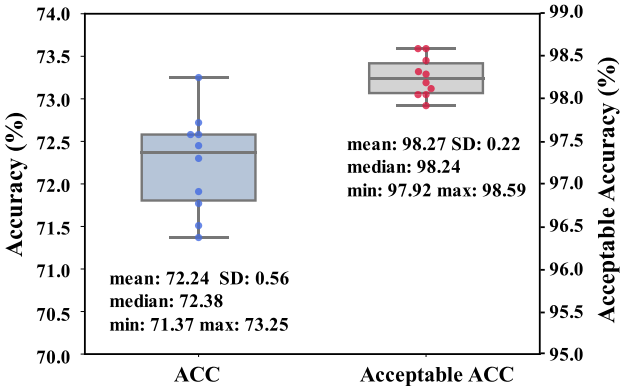


Fig. 12. Stability analysis of 10 time 4-fold CV experiments.

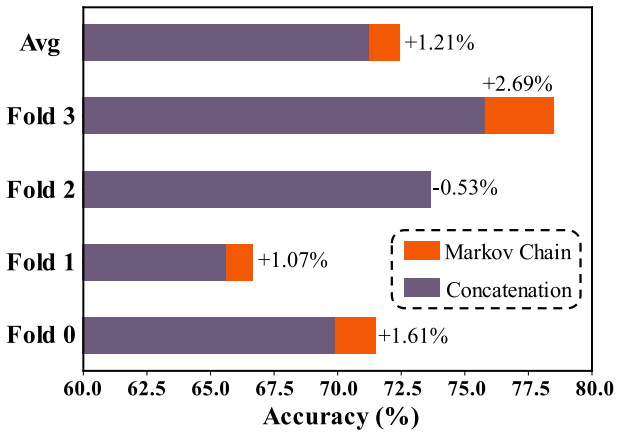


Fig. 13. Comparison between the traditional concatenation and Markov chain fusion method.

is significantly improved, as shown in Fig. 10. To demonstrate the effectiveness of our novel network, we compared the performance between the three-stream network and the two-stream network combined with the additional modules of the spatial–temporal attention mechanism and bilinear pooling layer. As shown in Table 4, the classification accuracy of the three-stream network in various situations was approximately 6%-8% higher than that of the two-stream network.

In addition, we compared feature fusion methods using the Markov chain and the traditional concatenation, as shown in Fig. 13. The Markov chain model increased the accuracy of the 4-fold CV of the three-stream network model by an average of 1.2%. Thus, the effectiveness of this new feature fusion scheme is verified.

4.4.3. Ablation analysis of Spatial-temporal attention module and bilinear pooling layer operation

To verify the effectiveness of the spatial–temporal attention (ST-Attention) module, we compared it with the self-attention module [53], separate temporal attention module (TAM), and spatial attention module (SAM). The self-attention module, TAM, and SAM have similar architecture but different in calculating attention map β. As shown in Supplementary Fig. S1, in the self-attention module, the attention map $\beta \in \mathbb{R}^{C\times C}$ means the correlation among feature channels, while $\beta \in \mathbb{R}^{T\times T}$ and $\beta \in \mathbb{R}^{V\times V}$ in TAM and SAM refer to the long-range global dependency on temporal and spatial dimensions. Table 5 lists the ablation results of attention modules, and the spatial–temporal attention module outperforms others in the finger-tapping action recognition task. The implementation details are in the Supplementary material. The spatial–temporal attention mechanism is a general module for skeleton-based CNNs; it improves the performance of not only the three-stream network but also the two-stream network, as shown in Table 4.

The symmetric bilinear pooling layer also plays a critical role in increasing the fine-grained classification performance. As listed in Table 4, the bilinear pooling layer improved the accuracy of both the three-stream and two-stream networks by 2.1% and 4.1%, and it also reduced the standard deviation of the classification result (Fig. 10). Finally, we combined the spatial–temporal attention mechanism with the bilinear pooling layer and achieved 72.4% accuracy and 98.3% acceptable accuracy.

Table 4
Comparison between three-stream and two-stream networks.

Additional Module	Three-Stream Model		Two-Stream Model	
	ACC (%)	Acceptable ACC (%)	ACC (%)	Acceptable ACC (%)
–	68.3	98.0	60.6	96.5
ST-Attention	69.9	98.3	61.7	96.4
Bilinear	70.4	98.4	64.7	96.8
ST-Attention + Bilinear	**72.4**	**98.3**	**65.6**	**97.2**

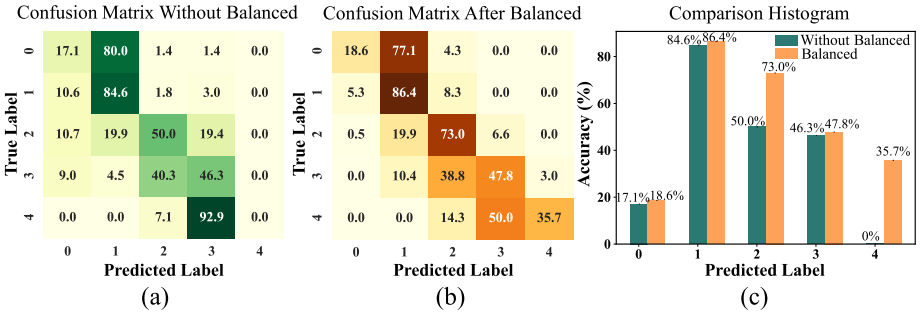


Fig. 14. Performance of Mini-batch-based balanced algorithm. (a) and (b) show the confusion matrix result without and with balanced processing, respectively; (c) is the comparison histogram.

Table 5
Performance of different attention modules.

Model	ACC (%)	Acceptable ACC (%)
Three-Stream CNN	70.4	98.4
Three-Stream CNN + Temporal-Attention	71.4	98.5
Three-Stream CNN + Spatial-Attention	71.2	98.2
Three-Stream CNN + Self-Attention	71.1	98.3
Three-Stream CNN + ST-Attention	**72.4**	**98.3**

4.4.4. Effect of Mini-batch-based balanced algorithm

To evaluate the effectiveness of our proposed balanced algorithm, we implemented the ablation study of our proposed balanced algorithm. Fig. 14 illustrates the confusion matrix and accuracy contrast histogram of the classification result with and without the balanced algorithm. Obviously, the novel mini-batch-based balanced algorithm improves the classification accuracy of each MDS-UPDRS score. In addition, our proposed balanced algorithm can avoid the phenomenon that minority classes cannot be recognized. Even for the score-4 class, whose sample number is only 14 out of 744, the balanced algorithm could considerably increase the accuracy from 0% to 35.7%. Detailed statistics and correlation analysis of score-0 and score-4 class is provided in the supplementary material. The improvement of the classification performance was not only for minority classes but also for the majority classes. For example, the predicted accuracy of the class with score-2 increased from 50.0% to 73.0%.

4.4.5. Comparison with State-of-the-art deep learning networks

To confirm the superiority of our method in fine-grained action recognition, we compared our approach with state-of-the-art skeleton-based DL networks, as shown in Table 6. Specifically, we implemented ST-GCN [29], Motif-GCNs [54], two-stream CNN [32], and 2 s-AGCN [2] on our finger-tapping dataset, and their classification accuracy was 53.4%, 57.1%, 60.6%, and 61.9%, respectively. In contrast, our proposed method obtained an accuracy of 72.4%, which is considerably higher than other networks.

Besides, we provided the computational complexity, including parameters (Params), floating-point operations (FLOPs), memory overhead (Memory), and inference time, of different state-of-the-art models in Table 6. The Params, FLOPs, and Memory were all calculated through an open-source tool (https://github.com/Swall0w/torchstat), and the inference time was computed by predicting the score of a single validation sample. All models in Table 6 were implemented on an NVIDIA GeForce 1080Ti GPU (11 GB). Compared to other methods, our model achieves the best balance between classification performance and computational burden. In addition, the OpenPose [28] hand pose estimation algorithm can achieve 6.4 FPS on our device, which is also suitable for real-time implementation.

5. Discussion

In this study, we proposed an automated assessment method for the finger-tapping test of PD patients; it is based on the vision-based action recognition Deep Learning network and finally achieved accurate and reliable results on a large dataset. Our approach provides an objective and quantitative tool for motor dysfunction evaluation, which can avoid the limitation of subjectivity and inter-neurologist variability in clinical practice. For future application prospects, our proposed method has excellent potential for clinical evaluation and remote monitoring of PD patients. In this way, doctors can measure the symptoms' severity of PD patients frequently and conveniently, draw up better treatment plans to determine the drug intake and DBS parameters.

5.1. Comparison with other automated assessment methods of Finger-tapping tests

We compared our proposed method with the existing sensor- and video-based automated assessment systems of finger-tapping tests in the MDS-UPDRS, as shown in Table 7. Manzanera et al. [17] applied a nine-degrees-of-freedom sensor and SVM clas-

Table 6
Comparison with state-of-the-art skeleton-based action recognition deep learning networks.

Model	Accuracy (%)		Computational Complexity			
	ACC	Acceptable ACC	Params (M)	FLOPs (G)	Memory (MB)	Inference Time (ms)
ST-GCN [29]	53.4	89.1	3.1	3.5	80.3	8.1
Motif-GCNs [54]	57.1	91.0	1.7	2.2	120.1	30.2
Two-Stream CNN [32]	60.6	96.5	9.2	0.1	2.5	2.2
2 s-AGCN [2]	61.9	93.3	6.9	7.9	188.2	49.7
Ours	**72.4**	**98.3**	**14.0**	**3.8**	**82.7**	**7.1**

Table 7
Comparison of other works in automatic MDS-UPDRS scoring of finger-tapping tests.

	Data source	Participants	Data Size	Classifier	ACC (%)
[17]	Sensor	25 PD 10 HC	–	SVM	68.5
[26]	Videos	60 PD	120 videos	SVM	90.0
Ours	**Videos**	**157 PD**	**744 videos**	**SVM**	**61.6**
	Videos	**157 PD**	**744 videos**	**Three-Stream CNN**	**72.4**

HC: health control.

sifier to realize 68.5% accuracy. Although wearable sensors are more sensitive and accurate than the vision-based pose estimation algorithm, our proposed method is superior for fine-grained action recognition. Hence, it is reasonable that the result (72.4% ACC) of our proposed vision-based DL scheme is slightly better than that of the sensor-based traditional feature engineering method [17].

In addition, Liu et al. [26] extracted 4 features by a hand pose estimation algorithm and achieved 90% automatic prediction accuracy with an SVM classifier. For comparison, we also applied the SVM classifier with feature engineering to our finger-tapping test dataset, according to [26] (implementation details in supplementary material). In this way, we obtained 61.6% accuracy, which is lower than the accuracy reported in [26]. However, Liu et al. primarily aimed to accelerate the pose estimation algorithm to make the overall system more convenient rather than improve the accuracy of pose estimation. Thus, the ideal dataset may be a reason for their considerably higher performance than the average level. In contrast, our method has the following competitive advantages: 1) To the best of our knowledge, we collected the world's largest clinical video dataset, making our study more suitable for meeting the requirement of clinical practice. 2) We provided more comprehensive evaluations to verify the reliability and robustness of our method, including the specific number of each class, confusion matrix, ROC curve, CDF, and so on. 3) The acceptable accuracy of 98.3% confirmed the usefulness of our method in clinical practice.

5.2. The effect of different input data length

The CNNs based approach requires the input with a fixed size, whereas each subject's time cost of 10 repetitions of finger-tapping movement is different. Thus, the determination of a proper data length is essential in this study. Too short length cannot cover sufficient information, and over-long of input frames inevitably introduce noise and redundant information. In around 6 s, most subjects can complete the finger-tapping task. Supplementary Table S1 compares the classification results with different input data length, including 90, 120, 150, 180, and 300 frames. Our proposed model achieved the best performance with the input length of 150 frames (5 s). It also can be seen that the performance is insensitive to the input size. Hence, we finally determined the 150 frames as the optimal length in this work.

5.3. Study limitations and future work

Moreover, lacking healthy controls is a limitation of our work. For finger-tapping tests in MDS-UPDRS, the score-0 class represents the movements of those who have no problem in the examination. Although 70 samples of PD patients were rated score-0, our study will be extended towards a more comprehensive and rigorous design, ideally with expanded data from age-matched health subjects in the future.

6. Conclusion

We proposed a vision-based fine-grained classification model for skeleton-based action recognition and realized an accurate automated MDS-UPDRS scoring for the finger-tapping test conducted in motor evaluation in PD. Our study can be summarized as follows: (1) A skeleton-based three-stream fine-grained classification network with Markov chain fusion was developed. (2) Fine-grained features were effectively captured with the modules of the spatial–temporal attention mechanism and symmetric bilinear pooling layers. (3) A novel mini-batch-based balanced algorithm with an inter-class sampling strategy was developed. Comprehensive experimental results on a large dataset demonstrated the accuracy and reliability of our method and confirmed its suitability for clinical practice. Furthermore, our approach does not need any additional wearable devices and provides a potential tool for remote assessment and mobile monitoring of PD patients in the future.

Declaration of Competing Interest

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

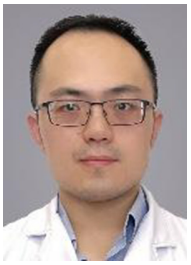
Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.neucom.2021.02.011.

References

[1] C.M. Tanner, S.M. Goldman, Epidemiology of Parkinson's Disease, Neurol. Clin. 14 (1996) 317–335.
[2] L. Shi, Y. Zhang, J. Cheng, 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.
[3] P.S. Group, Levodopa and the progression of Parkinson's disease, N. Engl. J. Med. 351 (2004) 2498–2508.
[4] A.L. Benabid, Deep brain stimulation for Parkinson's disease, Curr. Opin. Neurobiol. 13 (2003) 696–706.
[5] R.A. Hauser, F. Deckers, P. Lehert, Parkinson's disease home diary: Further validation and implications for clinical trials, Mov. Disord. 19 (2004) 1409–1413.
[6] N. Zhu, D. Zhang, W. Wang, X. Li, B. Yang, J. Song, X. Zhao, B. Huang, W. Shi, R. Lu, A novel coronavirus from patients with pneumonia in China, 2019, N. Engl. J. Med. (2020).
[7] C.G. Goetz, B.C. Tilley, S.R. Shaftman, G.T. Stebbins, S. Fahn, P. Martinez-Martin, W. Poewe, C. Sampaio, M.B. Stern, R. Dodel, Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): Scale presentation and clinimetric testing results, Movement Disorders 23 (2008) 2129–2170.
[8] M. Yokoe, R. Okuno, T. Hamasaki, Y. Kurachi, K. Akazawa, S. Sakoda, Opening velocity, a novel parameter, for finger tapping test in patients with Parkinson's disease, Parkinsonism Related Disorders 15 (2009) 440–444.
[9] C.G. Goetz, G.T. Stebbins, Assuring interrater reliability for the UPDRS motor section: utility of the UPDRS teaching tape, Mov. Disord. 19 (2004) 1453–1456.
[10] P. Arias, V. Robles-García, N. Espinosa, Y. Corral, J. Cudeiro, Validity of the finger tapping test in Parkinson's disease, elderly and young healthy subjects: Is there a role for central fatigue?, Clin Neurophysiol. 123 (2012) 2034–2041.
[11] A.L. Taylor Tavares, G.S. Jefferis, M. Koop, B.C. Hill, T. Hastie, G. Heit, H.M. Bronte-Stewart, Quantitative measurements of alternating finger tapping in Parkinson's disease correlate with UPDRS motor disability and reveal the improvement in fine motor control from medication and deep brain stimulation, Movement Disorders 20 (2005) 1286–1298.
[12] J.S. Wefel, B.D. Hoyt, P.J. Massman, Neuropsychological functioning in depressed versus nondepressed participants with Alzheimer's disease, Clin. Neuropsychol. 13 (1999) 249–257.
[13] J.N. Leijnse, N.H. Campbell-Kyureghyan, D. Spektor, P.M. Quesada, Assessment of individual finger muscle activity in the extensor digitorum communis by surface EMG, J. Neurophysiol. 100 (2008) 3225–3235.
[14] E. Strauss, E.M. Sherman, O. Spreen, A compendium of neuropsychological tests: Administration, norms, and commentary, American Chemical Society, 2006.
[15] I. Shimoyama, T. Ninchoji, K. Uemura, The finger-tapping test: A quantitative analysis, Arch. Neurol. 47 (1990) 681–684.
[16] I. Teshuva, I. Hillel, E. Gazit, N. Giladi, A. Mirelman, J.M. Hausdorff, Using wearables to assess bradykinesia and rigidity in patients with Parkinson's disease: A focused, narrative review of the literature, J. Neural Transm. 126 (2019) 699–710.
[17] O. Martinez-Manzanera, E. Roosma, M. Beudel, R. Borgemeester, T. van Laar, N. M. Maurits, A method for automatic and objective scoring of bradykinesia using orientation sensors and classification algorithms, IEEE Trans. Biomed. Eng. 63 (2015) 1016–1024.
[18] F. Parisi, G. Ferrari, M. Giuberti, L. Contin, V. Cimolin, C. Azzaro, G. Albani, A. Mauro, Inertial BSN-based characterization and automatic UPDRS evaluation of the gait task of Parkinsonians, IEEE Trans. Affective Comput. 7 (2016) 258–271.
[19] M. Giuberti, G. Ferrari, L. Contin, V. Cimolin, C. Azzaro, G. Albani, A. Mauro, Assigning UPDRS scores in the leg agility task of Parkinsonians: Can it be done through BSN-based kinematic variables?, IEEE Internet Things J. 2 (2015) 41–51.
[20] P. Schwab, W. Karlen, PhoneMD: Learning to diagnose Parkinson's disease from smartphone data, Proc. AAAI Conf. Art. Intell. (2019) 1118–1125.
[21] S.-T. Tang, C.-H. Tai, C.-Y. Yang, J.-H. Lin, Feasibility of smartphone-based gait assessment for parkinson's Disease, J. Med. Biol. Eng. 40 (2020) 582–591.
[22] J.C. Pérez-Ibarra, A.A. Siqueira, H.I. Krebs, Identification of gait events in healthy and parkinson's disease subjects using inertial sensors: A supervised learning approach, IEEE Sens. J. 20 (2020) 14984–14993.
[23] M.D. Hssayeni, J. Jimenez-Shahed, B. Ghoraani, Hybrid feature extraction for detection of degree of motor fluctuation severity in parkinson's disease patients, Entropy 21 (2019) 137.
[24] M.H. Li, T.A. Mestre, S.H. Fox, B. Taati, Vision-based assessment of parkinsonism and levodopa-induced dyskinesia with pose estimation, J. NeuroEng. Rehabil. 15 (2018) 97.
[25] M.H. Li, T.A. Mestre, S.H. Fox, B. Taati, Automated assessment of levodopa-induced dyskinesia: Evaluating the responsiveness of video-based features, Parkinsonism Related Disorders 53 (2018) 42–45.
[26] Y. Liu, J. Chen, C. Hu, Y. Ma, D. Ge, S. Miao, Y. Xue, L. Li, Vision-based method for automatic quantification of parkinsonian bradykinesia, IEEE Trans. Neural Syst. Rehabil. Eng. 27 (2019) 1952–1961.
[27] R. Guo, X. Shao, C. Zhang, X. Qian, Sparse adaptive graph convolutional network for leg agility assessment in Parkinson's disease, IEEE Trans. Neural Syst. Rehabil. Eng. (2020).
[28] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, Y. Sheikh, OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields, arXiv preprint arXiv:1812.08008, (2018).
[29] S. Yan, Y. Xiong, D. Lin, Spatial temporal graph convolutional networks for skeleton-based action recognition, Thirty-Second AAAI Conference on Artificial Intelligence2018).
[30] G. Hu, B. Cui, S. Yu, Skeleton-based action recognition with synchronous local and non-local spatio-temporal learning and frequency attention, 2019 IEEE International Conference on Multimedia and Expo (ICME), (IEEE2019), pp. 1216-1221.
[31] Y. Du, Y. Fu, L. Wang, Skeleton based action recognition with convolutional neural network, 2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR), (IEEE2015), pp. 579-583.
[32] C. Li, Q. Zhong, D. Xie, S. Pu, Skeleton-based action recognition with convolutional neural networks, 2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), (IEEE2017), pp. 597-600.
[33] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, The kinetics human action video dataset, arXiv preprint arXiv:1705.06950, (2017).
[34] A. Shahroudy, J. Liu, T.-T. Ng, G. Wang, Ntu rgb+ d: A large scale dataset for 3d human activity analysis, Proceedings of the IEEE conference on computer vision and pattern recognition2016), pp. 1010-1019.
[35] S. Poria, E. Cambria, R. Bajpai, A. Hussain, A review of affective computing: From unimodal analysis to multimodal fusion, Inform. Fusion 37 (2017) 98–125.
[36] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, A.Y. Ng, Multimodal deep learning, ICML2011).
[37] J. Yu, Y. Rui, B. Chen, Exploiting click constraints and multi-view features for image re-ranking, IEEE Trans. Multimedia 16 (2013) 159–168.
[38] J. Yu, Y. Rui, D. Tao, Click prediction for web image reranking using multimodal sparse coding, IEEE Trans. Image Process. 23 (2014) 2019–2032.
[39] X. Zhang, L. He, K. Chen, Y. Luo, J. Zhou, F. Wang, Multi-view graph convolutional network and its applications on neuroimage analysis for parkinson's disease, AMIA Annual Sympos. Proc., (American Medical Informatics Association (2018) 1147.
[40] Q.W. Oung, M. Hariharan, H.L. Lee, S.N. Basah, M. Sarillee, C.H. Lee, Wearable multimodal sensors for evaluation of patients with Parkinson disease, 2015 IEEE International Conference on Control System, Computing and Engineering (ICCSCE), (IEEE2015), pp. 269-274.
[41] L. Nie, L. Zhang, Y. Yang, M. Wang, R. Hong, T.-S. Chua, Beyond doctors: Future health prediction from multimedia and multimodal observations, in: Proceedings of the 23rd ACM international conference on Multimedia, 2015, pp. 591–600.
[42] J.C. Vasquez-Correa, J.R. Orozco-Arroyave, R. Arora, E. Nöth, N. Dehak, H. Christensen, F. Rudzicz, T. Bocklet, M. Cernak, H. Chinaei, Multi-view representation learning via GCCA for multimodal analysis of Parkinson's disease, 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), (IEEE2017), pp. 2966-2970.
[43] T.-Y. Liu, Easyensemble and feature selection for imbalance data sets, 2009 International Joint Conference on Bioinformatics, Systems Biology and Intelligent Computing, (IEEE2009), pp. 517-520.
[44] X.-Y. Liu, J. Wu, Z.-H. Zhou, Exploratory undersampling for class-imbalance learning, IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 39 (2008) 539–550.
[45] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, SMOTE: Synthetic minority over-sampling technique, J. Art. Intell. Res. 16 (2002) 321–357.
[46] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, Proceedings of the IEEE international conference on computer vision2017), pp. 2980-2988.
[47] F. Yang, Y. Wu, S. Sakti, S. Nakamura, Make Skeleton-based Action Recognition Model Smaller, Faster and Better, Proceedings of the ACM Multimedia Asia, 2019), pp. 1-6.
[48] M. Zolfaghari, G.L. Oliveira, N. Sedaghat, T. Brox, Chained multi-stream networks exploiting pose, motion, and appearance for action classification and detection, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2904–2913.
[49] X. Wei, Y. Zhang, Y. Gong, J. Zhang, N. Zheng, Grassmann pooling as compact homogeneous bilinear pooling for fine-grained visual classification, Proceedings of the European Conference on Computer Vision (ECCV)2018), pp. 355-370.
[50] A. Savitzky, M.J. Golay, Smoothing and differentiation of data by simplified least squares procedures, Anal. Chem. 36 (1964) 1627–1639.
[51] F. Parisi, G. Ferrari, M. Giuberti, L. Contin, V. Cimolin, C. Azzaro, G. Albani, A. Mauro, 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. Inf. 19 (2015) 1777–1793.
[52] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556, (2014).
[53] H. Zhang, I. Goodfellow, D. Metaxas, A. Odena, Self-attention generative adversarial networks, arXiv preprint arXiv:1805.08318, (2018).
[54] Y.-H. Wen, L. Gao, H. Fu, F.-L. Zhang, S. Xia, Graph CNNs with motif and variable temporal block for skeleton-based action recognition, Proc. AAAI Conf. Art. Intell. (2019) 8989–8996.

Xiaohua Qian received his Ph.D. in EE from Jilin University in 2012. He is currently an Associate Professor at the Shanghai Jiao Tong University (SJTU) School of BME. Before joining SJTU, Dr. Qian worked at The University of Texas Health Science Center at Houston School of Biomedical Informatics as an Assistant Professor. Prior to that, Dr. Qian was a Research Fellow at Wake Forest University's School of Medicine. During his doctoral program, he studied Medical Physics at Duke University Medical Center from 2010 to 2012. His primary research interests include machine learning, medical imaging analysis, and computer-aided diagnosis/surgery.



Chencheng Zhang completed his Ph.D. at Shanghai Jiao Tong University School of Medicine in 2019. He is currently also a junior principal investigator in Shanghai Research Center for Brain Science and Brain-Inspired Intelligence. His research interests include neuromodulation and brain-machine interface.

Xiangxin Shao received his Ph.D. degree in circuits and systems from Jilin University in 2010. He had visited Portland State University as a visiting scholar in 2017. He works as dean of the school of Electrical and electronic engineering at the Changchun University of Technology. His current research interests include signal processing and image processing.

Hao Li is currently pursuing his master's degree with the School of BME, Shanghai Jiao Tong University, China, under the supervision of Assoc. Prof. X. Qian. Prior to this, He earned his B.S. degree from the School of Computer Science, Beijing University of Post and Telecommunication, China. His research interests include skeleton-based action recognition and medical image analysis.