Video Based Shuffling Step Detection for Parkinsonian Patients Using 3D Convolution

Xugang Cao, Youze Xue, *Graduate Student Member, IEEE*, Jiansheng Chen, *Senior Member, IEEE*, Xiaohe Chen, Yu Ma, Chunhua Hu, *Member, IEEE*, Huimin Ma, *Member, IEEE*, and Hongbing Ma

***Abstract*— Parkinson's Disease (PD) is a common neurodegenerative disease which impacts millions of people around the world. In clinical treatments, freezing of gait (FoG) is used as the typical symptom to assess PD patients' condition. Currently, the assessment of FoG is usually performed through live observation or video analysis by doctors. Considering the aging societies, such a manual inspection based approach may cause serious burdens on the healthcare systems. In this study, we propose a pure video-based method to automatically detect the shuffling step, which is the most indistinguishable type of FoG. Firstly, the RGB silhouettes which only contain legs and feet are fed into the feature extraction module to obtain multi-level features. 3D convolutions are used to aggregate both temporal and spatial information. Then the multi-level features are aggregated by the feature fusion. Skip connections are implemented to reserve information of high resolution and period-wise horizontal pyramid pooling is utilized to fuse both global context and local features. To validate the efficacy of our method, a dataset containing 268 normal gait samples and 362 shuffling step samples is built, on which our method achieves an average detection accuracy of 90.8%. Besides shuffling step detection, we demonstrate that our method can also assess the severity of walking abnormity. Our proposal facilitates a more frequent assessment of FoG with less manpower and lower cost, leading to more accurate monitoring of the patients' condition.**

***Index Terms*— Parkinsonian shuffling step, abnormal gait recognition, 3D convolution, severity assessment.**

Manuscript received April 21, 2020; revised September 9, 2020 and January 11, 2021; accepted January 12, 2021. Date of publication February 26, 2021; date of current version March 16, 2021. This work was supported in part by the National Natural Science Foundation of China under Grant 61673234 and Grant U20B2062, in part by the National Key Research and Development Program of China under Grant 2016YFC0105502, and in part by the Key Projects of Ministry of Science and Technology under Grant 2017YFC1001800. *(Xugang Cao and Youze Xue contributed equally to this work.) (Corresponding authors: Jiansheng Chen; Xiaohe Chen; Yu Ma; Chunhua Hu.)*

Xugang Cao is with the Department of Electronic and Information Engineering, Changchun University of Science and Technology, Changchun 130022, China (e-mail: 2018100407@mails.cust.edu.cn).

Youze Xue and Hongbing Ma are with the Department of Electronic Engineering, Tsinghua University, Beijing 100084, China (e-mail: xueyz19@mails.tsinghua.edu.cn; hbma@mail.tsinghua.edu.cn).

Jiansheng Chen is with the Department of Electronic Engineering, Tsinghua University, Beijing 100084, China, also with the Beijing National Research Center for Information Science and Technology, Tsinghua University, Beijing 100084, China, and also with the School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing 100083, China (e-mail: jschenthu@mail.tsinghua.edu.cn).

Xiaohe Chen is with the Suzhou Institute of Biomedical Engineering and Technology, Chinese Academy of Sciences, Suzhou 215163, China (e-mail: chenxh@sibet.ac.cn).

Yu Ma is with the Tsinghua University Yuquan Hospital, Beijing 100040, China (e-mail: mayu@mail.tsinghua.edu.cn).

Chunhua Hu is with the National Engineering Laboratory for Neuromodulation, School of Aerospace Engineering, Tsinghua University, Beijing 100084, China (e-mail: huchunhua@mail.tsinghua.edu.cn).

Huimin Ma is with the School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing 100083, China (e-mail: mhmpub@ustb.edu.cn).

Digital Object Identifier 10.1109/TNSRE.2021.3062416

I. Introduction

PARKINSON'S Disease (PD) is a progressive neurodegenerative disease of the central nervous system which usually occurs in the elder group. PD patients suffer from several kinds of movement disorders including static tremor, muscular rigidity, bradykinesia and freezing of gait. These movement disorders seriously affect the life quality of patients, and over 6.1 million individuals suffer from it worldwide [1]–[4]. Among these disorders, freezing of gait (FoG) is a common debilitating symptom that occurs mostly in the middle to later stage of PD. According to statistics, in the later stage, more than 60% PD patients suffer from FoG, and 70% PD patients' falls are related to FoG [5], [6]. Therefore, FoG is considered as a typical symptom to assess PD patients' condition. For example, after a Deep Brain Stimulation (DBS) surgery, FoG is often used to guide doctors to adjust the parameters of electrical stimulation. However, the assessment of FoG requires heavy labor of specialized doctors. This work prevents the patients from getting more frequent assessment during their rehabilitation. Commonly a PD patient takes only 2 to 4 assessments a year, which is not enough for continuous monitoring of the patient's condition changes. If the automatic method with high efficiency and low cost can be developed, more frequent assessments can be conducted. This could provide doctors with more detailed information when adjusting patients' treatments.

To describe FoG in detail, we refer to UPDRS standard of walking abnormity to assign FoG symptoms with different scores [7]. As is shown in Tab. I, FoG symptoms are divided into four different levels according to severity. Thompson and Marsden developed a similar division of FoG in [8], where mild cases of FoG are called shuffling step. In our study, we refer to FoG symptoms with score 1 and 2 as shuffling step. Patients scored more than 2 are not able to walk independently, thus these types of FoG are easy to recognize. In contrast,

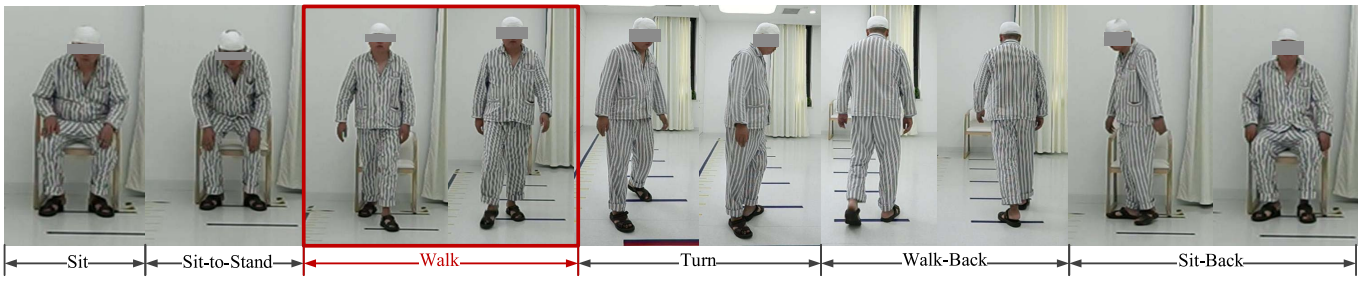

Fig. 1. The TUG test consists of six sub-tasks, including *Sit, Sit-to-Stand, Walk, Turn, Walk-Back*, and *Sit-Back*. We mainly focus on the sub-task *Walking* which is highlighted in red. To protect patients' privacy, the eye areas are covered in all figures in this paper.

TABLE I
UPDRS STANDARD FOR WALKING ABNORMITY

Score	Symptom Manifestation
0	Normal.
1	Slight abnormal. Possibily no movement of upper limbs or tend to shuffle.
2	Medium abnormal. Need for little or no help.
3	Severe abnormal. Incapable of walking without help.
4	Incapable of walking even with help.

patients with shuffling step are not that easy to be distinguished from normal people. They hit their feet to the ground toe to heel when walking, which needs careful observation of doctors to identify. Therefore, the shuffling step's detection is much more challenging than detection of FoG with scores 3 and 4.

Previous work has developed automatic methods for the detection of FoG as a whole. In [9], Hu *et al.* developed a dataset containing 45 subjects where most of them need help to walk. While the proposed method achieved great accuracy of FoG's detection on their dataset, experiments on our dataset in which subjects mainly suffer from shuffling step show that the method is not capable of effectively detecting shuffling steps. We focus on the detection of shuffling step and the assessment of shuffling step severity in this paper.

Many sensor-based methods have been proposed to detect FoG as the cost of sensors decreasing. Various gait motion parameters such as speed and orientation angles can be obtained by these sensors. However, clinical doctors still rely on timed up-and-go (TUG) tests heavily to diagnose and assess PD patients' condition in practice. Fig. 1 shows the six sub-tasks of the TUG test, including *Sit, Sit-to-Stand, Walk, Turn, Walk-Back*, and *Sit-Back*, which cover the most important activities in daily lives. Patients either perform TUG tests at hospitals observed by doctors or capture TUG videos at home and send them to doctors to get an assessment. In either way, TUG videos are captured and recorded,which provides the data foundation for automatic video-based assessment. Moreover, for patients receiving remote treatments, video-based methods are easier to be conducted since no specialized equipment is required. The only device required by TUG tests is a mobile phone with a camera. So, we follow the commonly used TUG test and develop an automatic method to detect and assess shuffling step based on TUG videos. Shuffling can be observed in the stage of *Walk*, *Turn* and *Walk-Back*. In the sub-task *Turn*, patients' legs and feet are severely occluded. And the *Turn* stage is often too short for careful observation. When patients are walking back, the toes would be occluded by legs, leading to severe loss of information about feet. Thus we only choose the sub-task *Walk* for analysis.

Video-based assessment of shuffling step is somewhat similar to the task of human gait recognition. The former detects abnormity in PD patients' gait, while the latter analyzes human gait to recognize human identity. Inspired by the commonly used framework in human gait recognition, we develop a two-stage pipeline. The first stage is feature extraction and the second stage is feature fusion. For preprocessing, a video clip is processed to produce the silhouettes of the subject. The RGB silhouettes are cropped to only contain legs and feet as the input of the feature extraction module. Different from general gait recognition in which explicit temporal relationship can be ignored [10], shuffling step is explicitly characterized by the hitting order of the toes and heels. This indicates that temporal information could be critical for shuffling step assessment. So for feature extraction, we utilize 3D convolutions to extract features of a frame sequence as a whole and produce multi-level features. Then the extracted features are fused by max operation through time and 2D convolutions are adopted to further extract spatial information. As the difference between patients' gait and normal people's gait is subtle, we argue that more information of high resolution is needed. Therefore skip connections from the shallow levels to the deep level are designed. At last, the fused features are further refined by period-wise horizontal pyramid pooling (PHPP) to combine both global context and local features before fed into the final classification layer. Beyond only detecting the existence of shuffling step, our method can also assess the severity of walking abnormity. This is useful, for example, in monitoring PD patients' condition changes during rehabilitation.

To validate the performance of our method, we collected 147 TUG videos from Tsinghua University Yuquan Hospital. Based on these videos, 362 positive samples with shuffling step and 268 negative samples with the normal step are sampled to formulate the shuffling step dataset. Besides, all the videos are given UPDRS scores according to their walking abnormity by clinical doctors from Yuquan Hospital. Our method achieves the shuffling step detection accuracy of 90.8% on the dataset, which is superior to state-of-the-art method in FoG detection.

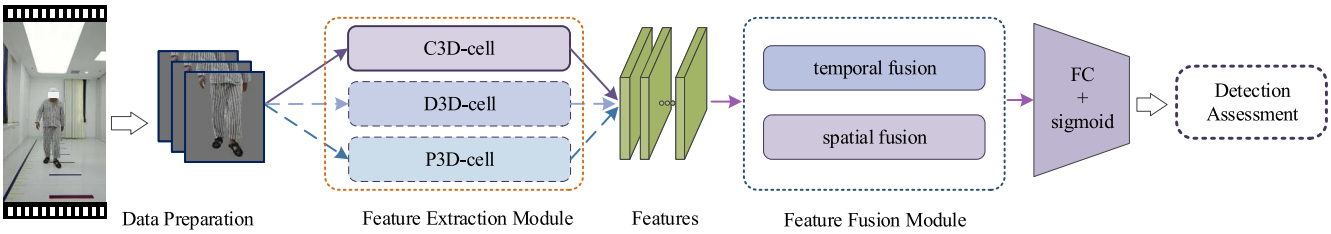

Fig. 2. The overall framework of our proposed method. The model consists of two parts, namely the feature extraction module and the feature fusion module. Finally, fully connected layers are used to produce the results of detection or assessment.

Also, our method performs well in severity scoring of walking abnormity with the accuracy of 84.2%.

II. Related Work

In this section, previous methods related to the analysis of PD patients' movement disorders are introduced. Methods which only use RGB videos as input are referred to as *video-based methods*, while methods which require specialized equipment such as motion sensors or depth cameras are referred to as *specialized-equipment-based methods*.

1) Specialized-Equipment-Based Methods: Motion sensors have been widely used to capture motion information. Camps *et al.* [11] proposed an approach to recognize FoG where a waist-placed inertial measurement unit (IMU) was used to collect movement signals. They used an 8-layer 1D convolutional neural network to process the motion signals. Similarly, Mileti *et al.* [12] used wearable sensors on patients' lower limbs to collect movement signals of gait, then the data was analyzed to evaluate the condition of patients. Apart from motion sensors, depth cameras are also common tools to analyze human gait. Nguyen *et al.* [13] used a depth camera Kinect to obtain the 3D skeleton of the patients. The 3D skeleton containing abundant information about motion was then further utilized for detecting normal gait. Dranca *et al.* [14] also used Kinect. After obtaining the 3D skeletons by Kinect, they utilized the Bayesian networks to classify Parkinson abnormal gait into three kinds. With the help of specialized equipment, these methods showed great performance in detecting movement disorders. However, the usage of specialized equipment leads to inconvenience in practice. For example, the wearable sensors may interfere with the patients' movement, especially when patients suffer from severe movement disorders. Also, the calibration of these sensors is too difficult for patients to operate at home which prevents the large scale of application in a remote manner. As for depth cameras, though the cost of them is decreasing these years, very few families would buy them for daily use. Even if in hospitals, doctors need additional labor to establish them along with traditional TUG test. On the contrary, pure video-based methods need no additional work since TUG test videos are usually recorded. The videos can also be captured by commonly used mobile phones which means video-based methods can be easily and conveniently conducted in a remote manner.

2) Video-Based Methods: Deep learning has been proven to be powerful in video processing. Tang *et al.* [15] proposed a method to achieve accurate detection of toe-off events using a single camera. They used consecutive silhouettes difference maps (CSD-maps) to represent the gait pattern. They argued that the CSD-maps provided significant features for toe-off event detection. Hu *et al.* [9] proposed a vision-based method to recognize FoG. They first detected the keypoints of legs and feet and then employed the graph convolution neural networks (GCNN) [16] to obtained the features of the keypoints and combined the features extracting from C3D networks to classify the abnormal gait and the normal gait. Wolf *et al.* [17] proposed multi-view 3D Convolutional Neural Network (MV3DCNN) to capture spatial-temporal information from gait sequences. Optical flow image was utilized to enhance the performance when facing different clothings. To solve the problem that convolutional network couldn't deal with long image sequences, a gait sequence was cut into several short sequences as the input of the network. Thapar *et al.* [18] proposed a two-stage method to identify human gait from multiple views. A 3D convolutional neural network was designed to estimate the viewing angle and perform subject identification. Liu *et al.* [19] proposed a video-based method to quantify hand movement bradykinesia severity on PD patients. Human pose estimation method was used to get finger joints' locations and then an SVM classifier used them to generate score ratings. Generally, mild movement disorders like shuffling step have not received much attention yet. We hence propose a pure video-based method to automatically assess shuffling step.

III. The Proposed Method

Fig. 2 illustrates the overall framework of our method. Two major modules are designed in our method. Firstly, each frame of a sample is pre-processed into a RGB silhouette as the input of the feature extraction module. Mask R-CNN [20] is used to produce the bounding box of the subject and then NLGInet [21] is utilized to parse the human body from the bounded patch. Noted that shuffling step is a movement disorder which affects the behaviour of a patient's legs and feet most, the RGB silhouette is further cropped to only contain legs and feet. Next, the cropped RGB silhouettes of a sample are concatenated together and fed into the feature extraction module. The feature extraction module utilizes 3D convolutions to extract multi-level features. The feature volume of the i-th level is a 5-dimensional tensor, denoted as $V_i \in \mathbb{R}^{B*T*H*W*C}$, $i = 1, 2, 3$, where B, T, H, W, C refer to batch size, time span, height, width and channels respectively. Then the multi-level features are fused temporally and spatially. To aggregate information across time, the max

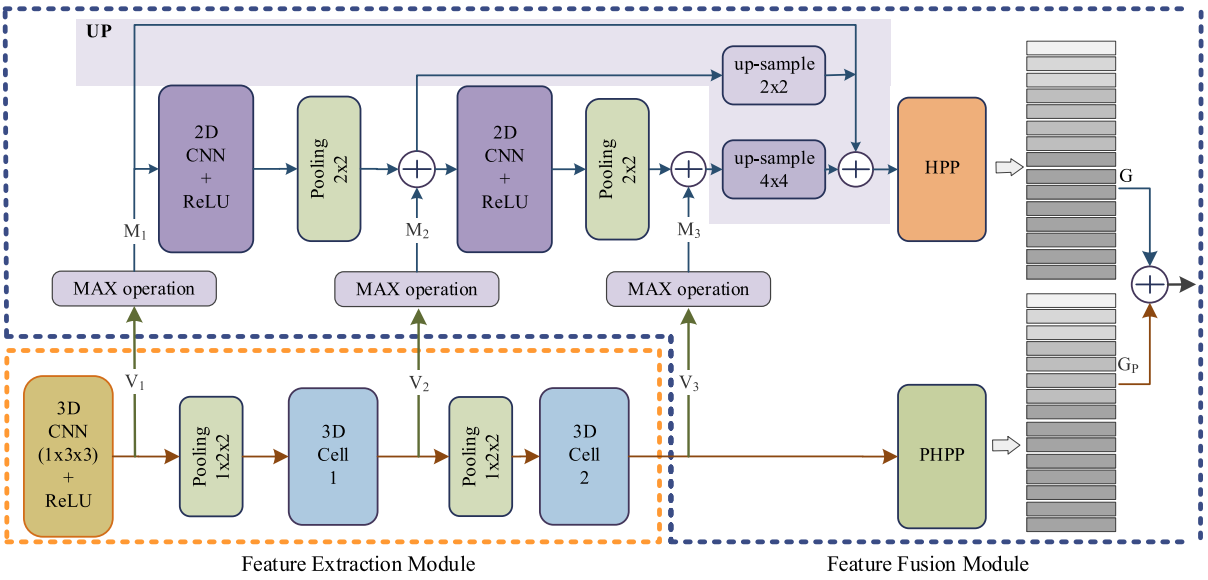

Fig. 3. The detailed architecture of our method. The feature extraction module is marked by the orange dashed line, and the feature fusion module is marked by the blue dashed line.

operation is utilized to extract most salient features in all frames. As for spatial fusion, multi-level features of different resolutions are combined and processed within multiple scales. The fused features are then flattened and fed into a classifier to produce the final detection result or the predicted severity level. Details of the structure of our proposed method are illustrated in Fig. 3.

A. The Feature Extraction Module

The feature extraction module is designed to extract multi-level temporal-spatial features. As is shown in Fig. 3, the feature extraction module consists of three levels. The first level extracts local features of each frame independently at original resolution. Noted that shuffling step is rather subtle compared to normal gaits, features of the first level are of the same resolution as input images to reserve information of high resolution. A 3D convolutional layer with 1*3*3 kernels is implemented to independently extract each frame's features in parallel. The second and third levels extract more global features and combine informative cues cross time. The features of each level are V_1, V_2 and V_3 respectively.

In human gait recognition, frames of a human gait sequence can be considered as independent [10]. However, we argue that shuffling step is characterized by the explicit temporal relationship which needs temporal aggregation. Specifically, when a patient with shuffling step is walking, his toes tend to hit the ground before his heel, while for normal people, the hitting order is just the opposite. If the order of frames in a sequence is disrupted, this important cue would be lost. Therefore, we utilize 3D convolutions to model the temporal relationship. To extract more global cues and reduce the cost of computation, pooling layers are used to downsample the feature volumes by 2 times before the second and the third levels.

In our method, we consider three types of convolutional cells. They are C3D-cell, D3D-cell, and P3D-cell.

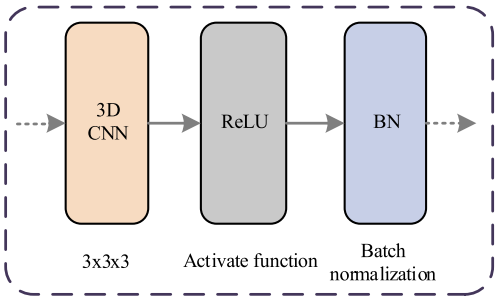

Fig. 4. The C3D-cell comes from C3D network [22].

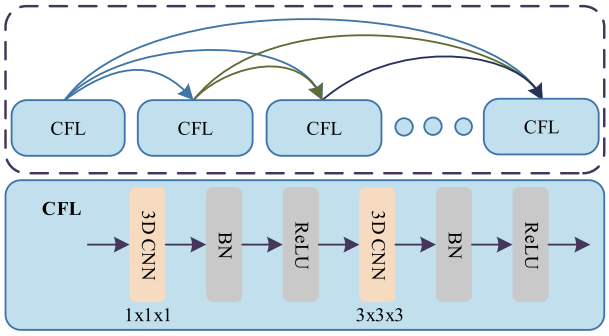

Fig. 5. D3D-cell contains multiple composite function layers (CFL), and the input of each CFL refers to the features of all former CFLs' outputs [23].

In [22], five 3D convolutional layers were cascaded to extract deep features. Following this design, we implement a basic cell consisting of a 3D convolutional layer, a ReLU layer and a batch normalization layer as the C3D-cell. As is shown in Fig. 4, the kernel size of the 3D convolutional layer is 3*3*3. D3D-cell is much more complicated [23]. Liu *et al.* proposed the D3D network for video-based person re-identification. A D3D-cell consists of six composite function layers (CFL), and these CFLs are densely connected in the form of densenet as shown in Fig. 5. In each CFL,

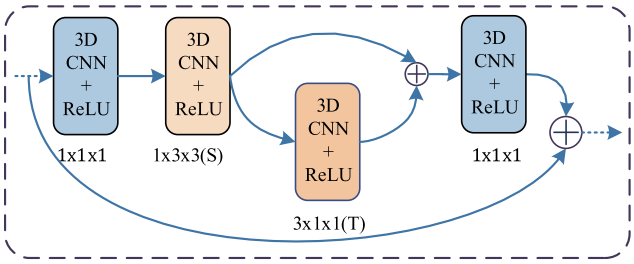

Fig. 6. P3D-cell derived from Pseudo-3D Residual Networks [24].

a 1*1*1 3D convolutional layer is used to adjust channels and a 3*3*3 convolutional layer is utilized to extract temporal-spatial features. Compared to C3D-cell, the D3D-cell has larger receptive field and requires more parameters. Inspired by the great success of ResNet [25] in numerous challenging image recognition tasks, Qiu *et al.* proposed the Pseudo-3D Residual Networks to extend residual networks to 3D convolutions [24]. The structure of the P3D-cell is illustrated in Fig. 6. The temporal processing and spatial processing are separated. D3D-cell and P3D-cell suffer from over-fitting much as the number of samples in our dataset are relatively small. The total number of layers in a D3D-cell is much more than C3D-cell. As for P3D-cell, the longest path in a cell is deeper than C3D-cell leading to higher complexity. As a result, we adopt C3D-cell as the 3D convolutional cell in our experiments. If the dataset is expanded in the future, it is possible that D3D-cell and P3D-cell can also exhibit good performance.

B. The Feature Fusion Module

The extracted multi-level spatial-temporal features are aggregated by the feature fusion module, as is shown in Fig. 3. The features are fused in both temporal and spatial dimensions.

1) Temporal Fusion: All vision cues related to shuffling step need to be aggregated through time. Shuffling step is a kind of movement disorder which is hard to distinguish from normal gaits and usually the typical characteristics of shuffling step only appear among a few consecutive frames. Also, in different stages of a gait cycle, shuffling step symptom could appear at different locations. For example, at the hitting moments of feet, pixels around the toes may be critical. While in the process of moving legs, the rigidity of legs may present discriminative features. Therefore an effective mechanism to identify and aggregate informative features during different stages is required. To process features across time at a fine-grained level, the pooling layers in the feature extraction module only down-sample the feature volumes spatially while the time span of feature volumes remains unchanged. After the process of 3D convolutions, a max operation across time is utilized to reserve the most descriptive features of shuffling step's pattern at each pixel, leading to three levels' feature maps $M_i \in \mathbb{R}^{B*H*W*C}, i = 1, 2, 3$.

2) Spatial Fusion: As is shown in Fig. 3, M_1, M_2 and M_3 are aggregated together with 2D convolutions and pooling layers. Further, we argue that high resolution features are valuable for detecting shuffling step, since shuffling step is characterized by certain local structures. For example, the length of the movement of toes on an image is very small while the raising abnormity of toes is one of the most important features of shuffling step. Thus, we develop a block called UP to make a better combination of shallow high-resolution features with deep features. In the UP block, skip connections from shallow layers to deep layers are designed and features from different layers are summed up. To match the original resolutions, upsampling layers are used. Then the summed features are spatially fused by a successful pooling mechanism called horizontal pyramid pooling (HPP) [26]. HPP divides a feature map horizontally into several strips as is shown in Fig. 7. According to the height of the strips, these strips contain information of different scales. In our experiments, the feature map is divided into 1,2,4 and 8 strips respectively leading to 15 strips with different scales. Each strip is pooled spatially by global average pooling(GAP) and global max pooling(GMP), and the GAP result and GMP result are summed up. After the above mentioned operation, each strip is represented by a C dimensional vector and all the 15 vectors are concatenated into a new feature map. Then independent fully-connected layers are implemented to transform these C dimensional vectors into C' dimensions, denoted as $G \in \mathbb{R}^{B*N*C'}$, where $N = 15$ in our experiments.

Moreover, we notice that shuffling step has characteristics at different time granularity. For example, the hitting order of toes and heels can be determined by a few frames near the exact hitting moment, while detecting abnormity of moving legs requires global analysis of a whole gait sequence. Therefore, we propose to conduct spatial fusion within multiple time spans. We propose a new pooling mechanism called period-wise horizontal pyramid pooling (PHPP) to conduct spatial fusion directly on V_3 before max operation. Fig. 7 illustrates the difference between PHPP and HPP. To be concise, the diagrams only display operation for each channel. Empirically, we divide the whole time span into 1,2 and 3 periods respectively, leading to a total of 6 periods. For each period, horizontal pyramid pooling is done and the resultants are concatenated across channel dimension. To adjust channels of the resulting feature maps, independent fully-connected layers are used to produce $G_p \in \mathbb{R}^{B*N*C'}$. Then G_p is added with G and the resultant is flattened as input of the final classification layer.

IV. Experimental Results

A. Data Preparation

Approved by Tsinghua University Yuquan Hospital, we have collected a dataset of totally 18 PD patients and 42 normal people. Each patient took several TUG tests before and after Deep Brain Stimulation (DBS) operation. The time interval between two TUG tests is at least one month so that the collected TUG videos are of rich diversity. As for normal people, identical TUG tests are conducted in Tsinghua University Yuquan Hospital to reduce the environmental biases in data collection. In total, 147 TUG videos are collected. Fig. 8 illustrates several TUG video fragments in our study.

All TUG videos are of the frame rate of 25 frames per second (FPS), namely the interval between consecutive frames

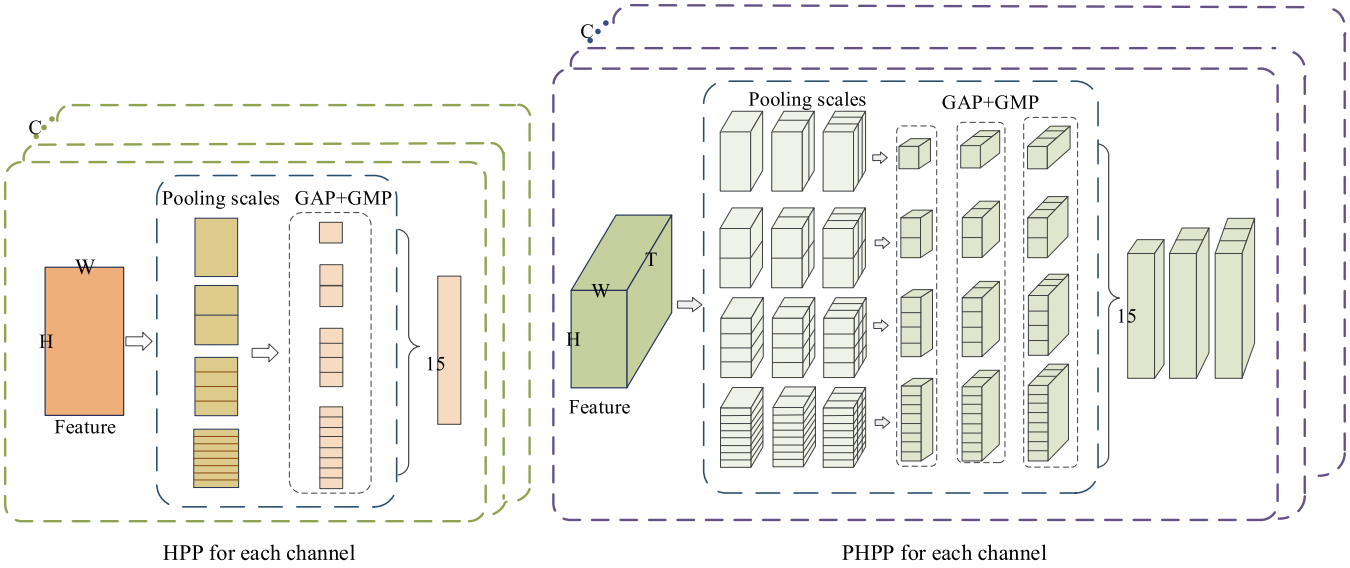

Fig. 7. Illustration of horizontal pyramid pooling (HPP) and period-wise horizontal pyramid pooling (PHPP). To be concise, the diagrams only display operation for one channel. The resultants for all the channels are concatenated together.

Fig. 8. Sample frames from Yuquan Hospital. The first row displays frames collected from Parkinson patient with shuffling step and the second row refers to the normal gait.

is 40 ms. According to our observation on the collected data, normal individuals or PD patients with shuffling step spend under 0.8s to complete a cycle of gait, which is defined as the time interval between two consecutive hitting moments on the ground of the same foot. Therefore we collect 25 consecutive frames which contain a complete cycle of gait as a sample. In general, for one TUG video three to five samples are produced. Moreover, all the 147 videos are given UPDRS scores by professionals to describe the severity of shuffling step following the standard shown in Tab. I. Tab. II shows the scoring of our samples. It is noticeable that our samples are all mild cases scored 0,1 and 2. As is discussed in Section I, for patients with scores more than 2 who couldn't walk independently without help, FOG can actually be detected

TABLE II
UPDRS SCORES OF THE SAMPLES

Patients	Total	*score*-0	*score*-1	*score*-2
18	362	0	119	243
Healthy	Total	*score*-0	*score*-1	*score*-2
42	268	268	0	0

in a straightforward way. On the contrary, it is much more meaningful and more difficult to distinguish a mild case from normal people. Samples with zero scores are denoted as negative and the others are regarded as positive, leading to a total of 362 positive samples and 268 negative samples.

Since our work specifically focuses on the sub-task *Walk* in TUG tests, we use the method proposed by Li *et al.* [27] to achieve automatic sub-task segmentation. After *Walk* fragments are separated out, Mask R-CNN [20] is used to detect human body area and produce bounding boxes around the body centers. Based on these bounding boxes, the human body area is cropped out and resized into 128*64 for each frame of a sample. Moreover, NLGInet [21] is used to perform human parsing to eliminate the interference of background. Noticed that shuffling step is more related to patients' legs and feet, the input images are further cropped to only contain the lower part of the body. Empirically, we directly crop the lowest quarter of the 128*64 images, leading to inputs of 32*64. Based on the above pre-processing, we produce a total of 6 types of input as is shown in Fig. 9. The first three types contain the full body of the subject, and the size of them is 128*64. Type-I is the original RGB version of the full image. Type-II is the silhouette version. Type-III is produced by eliminating background of Type-I, and is called the RGB silhouettes. Type-IV to VI only contain the lower part of the body. They are directly obtained by cropping the lowest quarter of the full size version. Experiments show that Type-VI performs the best compared to others, thus the following experimental results are all produced with Type-VI inputs.

Fig. 9. Illustration of six types of inputs. The first row corresponds to Type-I,II,III, and the second row displays Type-IV,V,VI.

B. Detection of Shuffling Step

A fully-connected binary classification layer is implemented following the proposed feature fusion module to detect the existence of shuffling step. As is mentioned above, samples with zero scores are considered as negative samples and the other samples are regarded as positive. To make the experimental results more stable and more reliable, the complete dataset is randomly divided into three parts to perform three-fold cross validation. Samples of the same subject are restricted to be in exactly one fold, so that the training set and the validation set do not have samples from the same participant.

Three metrics are used to assess the performance of our method and several previous state-of-the-arts. The average classification accuracy of three folds is referred to as *acc*. Besides, *prec* is used to assess the precision of the predicted positive samples and *rec* is used to measure the detection ratio of shuffling step as is defined in (1) and (2). Similar to *acc*, the average *prec* and *rec* are calculated on three folds.

$$prec = \frac{true\ positive}{true\ positive + false\ positive} \tag{1}$$

$$rec = \frac{true\ positive}{true\ positive + false\ negative} \tag{2}$$

To validate the effectiveness of our method, we reproduce several methods on our dataset. GaitSet [10] is a successful method for human gait recognition which is designed to recognize human identity from a gait sequence. D3D is also a classical architecture designed for person re-identification problem initially [23]. The final multi-class classification layers of them are replaced with binary fully-connected layers to produce the detection results of shuffling step. C3D is proposed as a universal 3D descriptor for video analysis. We follow the architecture designed in [22] which is initially designed for video-based action classification. Five C3D cells and two fully-connected layers are concatenated to produce a binary classification results. P3D [24] is another effective architecture for video-based action classification, and the final fully-connected layer is modified to produce binary output. JGP-GCNN [9] is recently proposed to detect FoG symptom from videos which is the most relevant method with ours. All the above mentioned methods are fed with inputs of the same formats as ours.

TABLE III
AVERAGE RESULTS OF 3-FOLD CROSS-VALIDATION

Methods	*acc(%)*	*prec(%)*	*rec(%)*
GaitSet [10]	84.9	87.1	85.2
JGR-GCNN [9]	79.1	87.5	76.5
C3D [22]	85	87.9	82.7
D3D [23]	87.5	89.4	85.1
P3D [24]	84.1	86.3	82.0
Ours	**90.8**	**92.1**	**90.8**

TABLE IV
TRUTH TABLE OF THE SAMPLES

	Total	Positive	Negative
Positive	362	333	29
Negative	268	29	239

Tab. III shows the quantitative comparison between our method and several state-of-the-arts. Our method achieves an accuracy of 90.8%, outperforming others with a large margin. Our method also achieves both higher precision and recall which demonstrates that our method not only enhances the overall accuracy but also maintains a good balance between sensitivity and specificity. And Tab. IV shows the truth table of our method for the detection of positive and negative samples, where positive samples represent shuffling step and negative samples represent normal gait.

To further demonstrate the effectiveness of our proposed method, ablation study is conducted and Tab. V presents the results. For comparison, We replace the 3D convolutional cells with parallel 2D convolutions, denoted as "Ours-2D". It is shown that 3D convolutions contribute a lot in both precision and recall, in line with our intuition that temporal relationship is critical for shuffling step's detection. Moreover, the combination of shallow high resolution features with deep features is also effective especially for the detection recall. Tab. V shows that the high resolution features contribute significantly on the detection recall with more than 5 percent at the cost of a slight sacrifice on detection precision. It is reasonable that the high resolution features are capable of detecting the subtle difference between mild cases of shuffling step and normal gaits, which may be the reason of the great improvement on recall. Further, the proposed period-wise horizontal pyramid pooling (PHPP) helps to achieve a better balance between sensitivity and specificity and enhance the overall accuracy of the detection. The period-wise temporal fusion makes it possible to aggregate information at different time scales, enabling global consideration of both long-term features and short-term characteristics. More interestingly, accuracy on three folds further reveals the ability of PHPP to stabilize the performance of the model, which implies better generalization on different data composition.

TABLE V
ABLATION STUDY

Methods	*acc(%)*	*prec(%)*	*rec(%)*	*acc-fold0(%)*	*acc-fold1(%)*	*acc-fold2(%)*	*params*	*FLOPs*	*time(ms)*
Ours-2D w/o UP & PHPP	84.9	87.1	85.2	86.5	84.1	84.1	1.4M	2.8M	3.6
Ours w/o UP & PHPP	88.1	89	88.2	90.1	84.1	90.1	1.5M	2.9M	3.7
Ours w/o PHPP	89.7	88.2	**93.6**	**91.7**	**90.3**	87.1	1.6M	3.2M	5.1
Ours	**90.8**	**92.1**	90.8	90.1	**90.3**	**91.8**	1.9M	3.8M	7.5

TABLE VI
COMPARISON OF INPUT FORMATS

Formats	*acc(%)*	*prec(%)*	*rec(%)*
Type-I	84.8	84.6	90.2
Type-II	81.1	83.9	80.4
Type-III	82.9	83.7	87.4
Type-IV	84.9	85.0	90.1
Type-V	83.9	83.5	87.3
Type-VI	**90.8**	**92.1**	**90.8**

TABLE VII
IMPACT OF ADDITIONAL DATA FROM CASIA

Data	*acc(%)*	*prec(%)*	*rec(%)*
Original Dataset	90.8	**92.1**	90.8
Extended Dataset	**91.3**	89.7	**92.0**

Tab. V also lists the number of parameters of our proposed method. It is notable that the 3D convolutional cells and the skip connections from high resolution features to deep features do not increase the number of parameters greatly. And also the PHPP block only increases a small portion of parameters which proves that the enhancement achieved by our method does not merely result from higher complexity. As for *FLOPs* and speed, only several milliseconds are needed to analyze a 25-frames sequence on a Geforce RTX 2080Ti GPU. Though the UP and PHPP structures consume a little bit more time, the proposed method is still fast enough for real time applications. And it is very promising to apply our method on mobile platforms in real time in the future.

C. Impact of Input Formats

Fig. 9 displays the six types of inputs. The impacts of them are evaluated using our proposed method on the detection accuracy. Tab. VI shows the quantitative results. It is observable that the cropped versions consistently outperform their counterpart full size version. This may be explained by the intuition that shuffling step is much more related with abnormity of legs and feet than the upper part of body, so that the cropped versions would guide the network to focus directly on the most important parts. Moreover, the silhouette versions lead to unsatisfactory results. This implies that the RGB variance of the image, which represents illumination, textures and body structures, contains rich motion cues of the subject. Since Type-VI performs the best on the dataset so that all the other experiments are carried out with this kind of inputs. Noticed that compared with results shown in Tab. III, the input formats are even more influential than the network architecture. This proves that input format is also a critical factor of detection accuracy which needs careful consideration.

Fig. 10. Five frames come from CASIA data [28]. We select 96 samples from CASIA data to balance our dataset.

D. Experiments With Additional Data

As the scale of our dataset is relatively small and the positive samples and negative samples are not very balanced, we refer to the public dataset for the extension. We refer to CASIA gait dataset to increment the number of negative samples [28]. CASIA gait dataset contains three sub-sets, namely Dataset A, Dataset B and Dataset C. Among them, Dataset B provides multi-view walking videos for 124 individuals. We select the front-view videos of 24 subjects to generate 96 negative samples. Fig. 10 illustrates several samples generated from CASIA data. The additional 96 samples are randomly divided into three folds and added into our dataset. Tab. VII shows the performance of our method on the extended dataset. The extension of the dataset boosts the average accuracy to 91.3%. This implies that with more data available, our method would perform even better on the detection accuracy.

E. Severity Assessment of Walking Abnormity

Compared with detection of shuffling step, a more challenging and more practical analysis on PD patients is the severity assessment of walking abnormity. In our dataset, each video is assigned with a score as its label according to UPDRS standard shown in Tab. I. We design a three-class classification task based on these TUG videos. The dataset is divided into three folds and cross-validation is performed. Noticed that one TUG video could contain more than one sample. If a video only contains one sample, then the prediction score of this sample is assigned to the video as the final prediction. When a video contains more than one sample, we calculate the most number of prediction as to the video's prediction score. In this way, our method achieves an average scoring accuracy of 84.2%. As we do not find similar research assessing the

severity of shuffling step based on RGB videos, we do not conduct experiments for comparison. Our method is helpful for assessing PD patients' condition changes during their recovery. The accuracy at present is largely restricted by lack of labelled data. In the future, it is promising to achieve more accurate and more continuous assessment of PD patients with the expansion of dataset.

V. Conclusion

In this paper, we propose a video-based automatic method for shuffling step detection and severity assessment. 3D convolutions are adopted to aggregate informative temporal cues. In the feature fusion module, multi-level features are fused both temporally and spatially. Extensive experiments demonstrate the effectiveness of our proposed method. We also explore the possibility of automatically scoring walking abnormity. With the development of large-scale dataset, it is promising to achieve remote and automatic assessment of PD patients' condition more accurately in the future.

References

[1] E. R. Dorsey *et al.*, "Global, regional, and national burden of Parkinson's disease, 1990–2016: A systematic analysis for the global burden of disease study 2016," *Lancet Neurol.*, vol. 17, no. 11, pp. 939–953, Nov. 2018.

[2] B. R. Bloem, J. M. Hausdorff, J. E. Visser, and N. Giladi, "Falls and freezing of gait in Parkinson's disease: A review of two interconnected, episodic phenomena," *Movement Disorders*, vol. 19, no. 8, pp. 871–884, 2004.

[3] L. M. de Lau and M. M. B. Breteler, "Epidemiology of Parkinson's disease," *Lancet Neurol.*, vol. 5, no. 6, pp. 525–535, 2006.

[4] S. Sveinbjornsdottir, "The clinical symptoms of Parkinson's disease," *J. Neurochem.*, vol. 139, pp. 318–324, Oct. 2016.

[5] T. Virmani, C. B. Moskowitz, J.-P. Vonsattel, and S. Fahn, "Clinicopathological characteristics of freezing of gait in autopsy-confirmed Parkinson's disease," *Movement Disorders*, vol. 30, no. 14, pp. 1874–1884, Dec. 2015.

[6] O. Moore, C. Peretz, and N. Giladi, "Freezing of gait affects quality of life of peoples with Parkinson's disease beyond its relationships with mobility and gait," *Movement Disorders*, vol. 22, no. 15, pp. 2192–2195, 2007.

[7] S. Fahn and R. L. Elton, "Unified Parkinson's disease rating scale," in *Recent Developments in Parkinson's Disease*. Florham Park, NJ, USA: MacMillan Health Care Information, 1987, pp. 153–163 and 293–305.

[8] J. D. Schaafsma, Y. Balash, T. Gurevich, A. L. Bartels, J. M. Hausdorff, and N. Giladi, "Characterization of freezing of gait subtypes and the response of each to levodopa in Parkinson's disease," *Eur. J. Neurol.*, vol. 10, no. 4, pp. 391–398, Jul. 2003.

[9] 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.

[10] H. Chao, Y. He, J. Zhang, and J. Feng, "Gaitset: Regarding gait as a set for cross-view gait recognition," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 8126–8133.

[11] J. Camps *et al.*, "Deep learning for detecting freezing of gait episodes in Parkinson's disease based on accelerometers," in *Proc. Int. Work-Conf. Artif. Neural Netw. (IWANN)*, Jun. 2017, pp. 344–355.

[12] I. Mileti *et al.*, "Gait partitioning methods in Parkinson's disease patients with motor fluctuations: A comparative analysis," in *Proc. IEEE Int. Symp. Med. Meas. Appl. (MeMeA)*, May 2017, pp. 402–407.

[13] T.-N. Nguyen, H.-H. Huynh, and J. Meunier, "Skeleton-based abnormal gait detection," *Sensors*, vol. 16, no. 11, p. 1792, Oct. 2016.

[14] L. Dranca *et al.*, "Using kinect to classify Parkinson's disease stages related to severity of gait impairment," *BMC Bioinf.*, vol. 19, no. 1, p. 471, Dec. 2018.

[15] Y. Tang, Z. Li, H. Tian, J. Ding, and B. Lin, "Detecting toe-off events utilizing a vision-based method," *Entropy*, vol. 21, no. 4, p. 329, Mar. 2019.

[16] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," 2016, *arXiv:1609.02907*. [Online]. Available: https://arxiv.org/abs/1609.02907

[17] T. Wolf, M. Babaee, and G. Rigoll, "Multi-view gait recognition using 3D convolutional neural networks," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2016, pp. 4165–4169.

[18] D. Thapar, A. Nigam, D. Aggarwal, and P. Agarwal, "VGR-Net: A view invariant gait recognition network," in *Proc. IEEE 4th Int. Conf. Identity, Secur., Behav. Anal. (ISBA)*, Jan. 2018, pp. 1–8.

[19] 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.

[20] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2961–2969.

[21] T. Li, W. Wan, Y. Huang, J. Chen, C. Hu, and Y. Ma, "Improving human parsing by extracting global information using the non-local operation," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2019, pp. 2961–2965.

[22] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4489–4497.

[23] J. Liu, Z.-J. Zha, X. Chen, Z. Wang, and Y. Zhang, "Dense 3D-convolutional neural network for person re-identification in videos," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 15, no. 1, pp. 1–19, Feb. 2019.

[24] Z. Qiu, T. Yao, and T. Mei, "Learning spatio-temporal representation with pseudo-3D residual networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5533–5541.

[25] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[26] Y. Fu *et al.*, "Horizontal pyramid matching for person re-identification," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, vol. 33, Jul. 2019, pp. 8295–8302,

[27] T. Li *et al.*, "Automatic timed up-and-go sub-task segmentation for Parkinson's disease patients using video-based activity classification," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 26, no. 11, pp. 2189–2199, Nov. 2018.

[28] S. Yu, D. Tan, and T. Tan, "A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition," in *Proc. 18th Int. Conf. Pattern Recognit. (ICPR)*, vol. 4, Aug. 2006, pp. 441–444.