Automatic Timed Up-and-Go Sub-Task Segmentation for Parkinson's Disease Patients Using Video-Based Activity Classification

Tianpeng Li, Jiansheng Chen, *Senior Member, IEEE*, Chunhua Hu, Yu Ma, Zhiyuan Wu, Weitao Wan, Yiqing Huang, Fuming Jia, Chen Gong, Sen Wan, and Luming Li, *Senior Member, IEEE*

***Abstract*—The timed up-and-go (TUG) test has been widely accepted as a standard assessment for measuring the basic functional mobility of patients with Parkinson's disease. Several basic mobility sub-tasks "Sit," "Sit-to-Stand," "Walk," "Turn," "Walk-Back," and "Sit-Back" are included in a TUG test. It has been shown that the time costs of these sub-tasks are useful clinical parameters for the assessment of Parkinson's disease. Several automatic methods have been proposed to segment and time these sub-tasks in a TUG test. However, these methods usually require either well-controlled environments for the TUG video recording or information from special devices, such as wearable inertial sensors, ambient sensors, or depth cameras. In this paper, an automatic TUG sub-task segmentation method using video-based activity classification is proposed and validated in a study with 24 Parkinson's disease patients. Videos used in this paper are recorded in semi-controlled environments with various backgrounds. The state-of-the-art deep learning-base 2-D human pose estimation technologies are used for feature extraction. A support vector machine and a long short-term memory network are then used for the activity classification and the subtask segmentation. Our method can be used to automatically acquire clinical parameters for the assessment of Parkinson's disease using TUG videos-only, leading to the possibility of remote monitoring of the patients' condition.**

***Index Terms*—Timed up-and-go, Parkinson's disease, human pose estimation, sub-task segmentation.**

I. Introduction

PARKINSON'S Disease (PD) is a mobility disorder that usually occurs in the elderly. The tests of basic functional mobility for PD patients have been explored by interdisciplinary researchers. The Timed Up-and-Go (TUG) is such a test that has been widely accepted in clinical practices for over 20 years [1]. The TUG test has been considered to be a succinct and efficient way to evaluate individuals' basic functional mobility mainly because it only includes the most critical fundamental activities in daily lives such as standing up, walking, turning around and sitting back. Fig. 1 illustrates a typical procedure of a 5-meter TUG test. Due to its effectiveness for detecting disorders in the gait and movement, the TUG test can be used to provide many useful PD assessment parameters such as the timing of the sub-tasks involved, the stride length and frequency [2]. Typically, the TUG test is executed under the supervision of clinicians, and the time durations of the sub-tasks are estimated using a stop watch through subjective observations. The patients are usually required to come to the hospital which may cause inconveniences considering their limited mobility. Furthermore, such a manual inspection based approach may cause heavy burden to the health care system in the aging societies. Therefore, automatic TUG analysis technologies, especially the subtask segmentation methods [3]–[13], have attracted a lot of research attention recently.

In previous studies, automatic TUG sub-task segmentation methods using only TUG videos as inputs [9]–[11] were obviously inferior to those using input signals captured from special devices such as wearable inertial sensors [7], [8], [12], [13], ambient sensors [3], [4] or depth cameras [5], [6] in terms of segmentation accuracies. Previous work showed that the accuracies of the video based automatic TUG sub-task segmentation methods were not high enough for clinic assessments even on TUG sequences recorded

Manuscript received November 22, 2017; revised April 25, 2018 and August 19, 2018; accepted September 26, 2018. Date of publication October 12, 2018; date of current version November 20, 2018. This work was supported in part by the National Natural Science Foundation of China under Grant 61673234 and Grant 81527901 and in part by the National Key Research and Development Program of China under Grant 2016YFC0105904 and Grant 2016YFC0105502. *(Corresponding author: Jiansheng Chen.)*

T. Li, J. Chen, Z. Wu, W. Wan, and Y. Huang are with the Department of Electronic Engineering, Tsinghua University, Beijing 100084, China (e-mail: ltp16@mails.tinghua.edu.cn; jschenthu@mail.tinghua.edu.cn; wu-zy18@mails.tinghua.edu.cn; wwt16@mails.tinghua.edu.cn; huang-yq17@mails.tinghua.edu.cn).

C. Hu, F. Jia, C. Gong, and S. Wan are with the School of Aerospace Engineering, Man-Machine-Environment Engineering Institute, Tsinghua University, Beijing 100084, China (e-mail: huchunhua@mail.tsinghua.edu.cn; jfmin@mail.tsinghua.edu.cn; gongc16@mails.tsinghua.edu.cn; wan-s13@mails.tsinghua.edu.cn).

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

L. Li is with the National Engineering Laboratory for Neuromodulation, Precision Medicine and Healthcare Research Center, Center of Epilepsy, School of Aerospace Engineering, Man-Machine-Environment Engineering Institute, Beijing Institute for Brain Disorders, Tsinghua-Berkeley Shenzhen Institute, Tsinghua University, Beijing 100084, China (e-mail: lilm@mail.tsinghua.edu.cn).

Digital Object Identifier 10.1109/TNSRE.2018.2875738

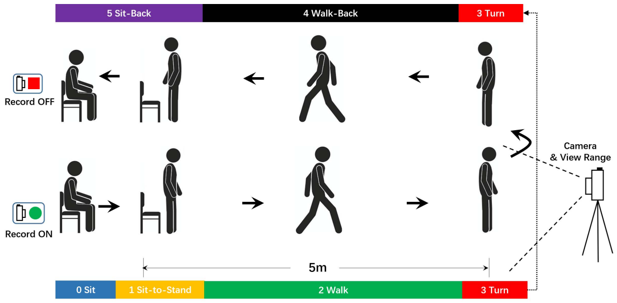


Fig. 1. The procedure and video recording protocol of the 5 meter TUG test. TUG sub-tasks are marked with different colors. The video camera is placed in front of the participant with the whole body covered in the view. The video recording starts with the participant sitting in a chair and being told to start; and stops when the participant finishes all sub-tasks and sits back in the chair.

in deliberately designed environments [9]–[11]. Nevertheless, in the past few years, the performance of the video based human activity analysis has been improved dramatically. For example, recent state-of-the-art human pose estimation technologies have demonstrated high reliability in detecting key points of the human body based on video information only [14], [15]. By combining these technologies with appropriate classification models, we propose an accurate and robust automatic TUG sub-task segmentation method which is purely based on video analysis. Since the only required input is a recorded TUG video sequence, our method practically allows the TUG test to be done remotely and analyzed automatically. This facilitates the building of telemedicine systems for low cost PD patient assessments and long term monitoring.

II. Related Work

Previous work on the automatic segmentation of TUG sub-tasks can be grouped into four categories according to the data capturing devices involved: inertial sensors, ambient sensors, video cameras and depth cameras [16].

Movement tracking devices like the inertial measurement units (IMUs) which contain accelerometers and gyroscopes can be attached to the human body to record activities. Signals generated by these sensors contain rich motion information for TUG analysis. Nguyen *et al.* [7] used a motion capturing suit attached with 17 inertial sensors, from which the generated data were first detrended to remove the sensor drift. A band pass filter with carefully designed cut-off frequencies was then used to reveal kinematic peaks that corresponded to different sub-tasks. Segmentation was finally accomplished by identifying the time stamps of the starting and ending positions of these detected peaks. This method was tested on 16 healthy elders with high detection sensitivity reported. Following this, Nguyen *et al.* [17] further optimized some parameters and achieved improved performances for 12 PD patients. Reinfelder *et al.* [8] used two inertial sensors attached on shoes to acquire 3D acceleration and orientation data, from which additional statistical features were computed. Four different classification models were adopted for distinguishing different sub-tasks. Salarian *et al.* [2] introduced a system called iTUG, in which seven inertial sensors together with a data-logger were mounted on the participants and each sub-task was detected using a unique subset of sensors. iTUG was tested on 12 PD patients and 12 healthy individuals. These inertial sensors based methods often achieve high experimental performances with accurate kinematic signals. However, installing and calibrating these sensors on PD patients are difficult without the help of experts. In addition, with too many additional devices, participants may feel uncomfortable, especially for those who are suffering from mobility disorders.

To avoid the discomfort caused by wearing inertial sensors, Frenken *et al.* [4] alternatively used ambient sensors for capturing and achieved a comparable performance with inertial sensor based methods. A chair was equipped with an infared light barrier, four force sensors and a laser range scanner. The infared light barrier was installed under the armrests to detect whether the participant's back is in contact with the backrest. The force sensors were placed inside the chair legs to monitor the weight distribution. The laser range scanner was placed under the chair to measure the distance between the participant and the chair. The time duration of each sub-tasks as well as additional gait information were estimated by processing the sensor data with a microcontroller system mounted to the chair. Similar to the inertial sensor based methods, installing and calibrating the ambient sensors is also challenging.

A video camera is non-contact and more convenient to access than inertial and ambient sensors in daily life. Therefore, video based methods are more convenient than the sensor based methods considering that TUG videos can be captured nearly anywhere with simple environmental settings. However, the major difficulty of the video based methods comes from the human activity analysis. In previous video based methods, only part of the TUG sub-tasks could be stably detected and timed. Berrada *et al.* [9] used a fixed side-view camera to compute the time of standing up and several initial walking steps. Skrba *et al.* [10] only estimated the time of the walking stage. A common practice in these methods is to use image processing procedures to acquire silhouette of the participant and then segment different sub-tasks using some empirically selected geometrical features such as the height, width and aspect ratio. Although TUG videos were recorded in well-controlled environments with uniform backgrounds and fixed camera angles in these methods, the reported performances were still low.

In contrast to ordinary 2D video cameras, depth cameras capture extra 3D structures. In previous studies, the Microsoft Kinect system was most commonly used for depth capturing and analysis. Kitsunezaki *et al.* [6] used the Kinect's skeleton tracking mode to detect the starting and ending points of a TUG test based on which the total time of the TUG procedure was estimated with relatively high accuracy. Furthermore, Lohmann *et al.* [5] proposed the skeleton-TUG (sTUG) method for segmenting and timing all the sub-tasks in a TUG test also by using the Kinect's skeleton tracking mode. However, the sTUG was not fully automatic and the body key point tracking results provided by Kinect were further processed manually when segmenting the sub-tasks. Similar to the inertial and ambient sensors, depth cameras are relatively less common in daily lives and are comparatively

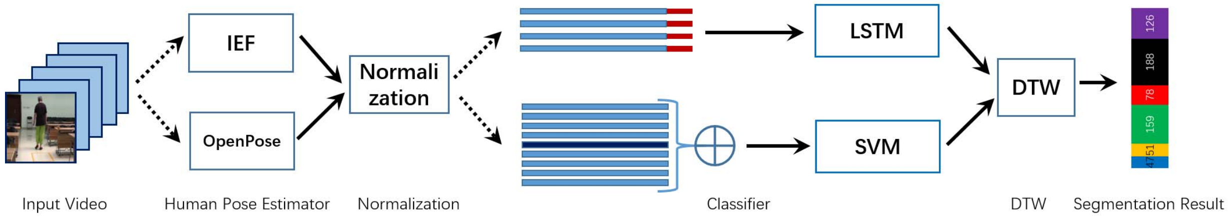


Fig. 2. The overall flowchart of the proposed method. The solid arrows illustrate the data flow where the the dash arrows indicate optional branches.

TABLE I
BASIC STATISTICS OF THE PARTICIPANTS

Parameter	Mean±Std	Range
Age (Years)	56.79 ± 9.48	$[37, 73]$
Weight (kg)	63.87 ± 10.37	$[49, 90]$
Height (cm)	164.83 ± 6.12	$[156, 178]$
BMI (kg/m^2)	23.51 ± 3.60	$[18.4, 31.3]$

TABLE II
MAJOR STATISTICS OF THE PD SEVERITY IN VIDEOS

Parameter	Mean±Std	Range	Max Score
UPDRS III Total	31.26 ± 17.48	$[1, 90]$	108
Gait	1.21 ± 0.91	$[0, 4]$	4
Postural Stability	1.96 ± 0.68	$[0, 4]$	4
Tremor	3.58 ± 5.14	$[0, 25]$	28
Bradykinesia	11.15 ± 6.86	$[0, 28]$	32
Rigidity	5.65 ± 4.22	$[0, 17]$	20

more difficult to set up and operate compared to ordinary 2D video cameras.

III. THE PROPOSED METHOD

Given a TUG video sequence captured in a semi-controlled environment by an ordinary 2D video camera, the proposed method automatically segments the video sequence into 6 sub-tasks, namely 'Sit,' 'Sit-to-Stand,' 'Walk,' 'Turn,' 'Walk-Back' and 'Sit-Back.' Fig. 2 shows the overall flowchart of the method. The input TUG video sequence is first processed by a human pose estimator to acquire coordinates of a set of key points of human body. Positions of these key points are treated as the raw feature and are spatially normalized for further analysis. Normalized features of consecutive frames are concatenated in order to take both spatial and temporal information into consideration. The concatenated features are input to a classifier to get frame-wise prediction of the sub-task classification. We use Dynamic Time Warping (DTW) algorithm to generate the sub-task segmentation. In this section, we will first introduce the data involved in this study and then explain the proposed method in detail.

A. Data Preparation

This study was approved by the IRB of Tsinghua University Yuquan hospital with the reference number YQ2015-08-12. A total of 24 PD patients who underwent a Deep Brain Stimulation (DBS) operation were involved in the data collection. The participants are different in age, weight and height as is shown in Table I. Also, the degree of motor disorder varies greatly among the participants. As is shown in Table II, the standard deviation of the UPDRS III (Unified Parkinson's Disease Rating Scale part III) score [18], which is the most commonly used mobility evaluation scale in the clinical study of PD, is as high as 17.48. This is also true for the five UPDRS III sub-scores: gait (item 28), postural stability (item 30), tremor (item 20, 21), bradykinesia (item 23, 24, 25, 26), rigidity (item 22). Aside from the mean and standard deviation, the dynamic ranges in the collected data and the possible highest values of some of the parameters are also listed in Table II.

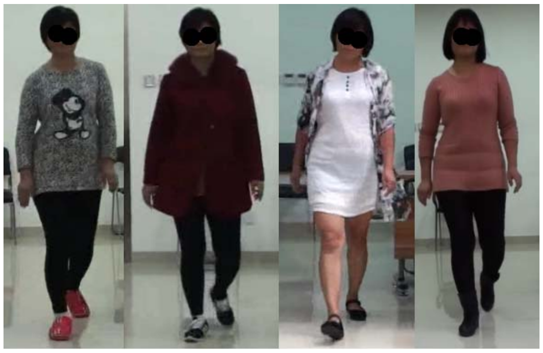


Fig. 3. Sample frames from different TUG videos of the same participant. The total times of TUG test are 46, 26, 13, 16 seconds respectively from left to right. To hide the participants' identities, eye areas are occluded in all the figures in this work.

Following the clinic protocol of Yuquan hospital, TUG tests were carried out on each PD patient both before the operation and at every return visit after the operation. Each participant underwent 4-6 TUG tests during the whole treatment period, and the time interval between two TUG tests of the same participant is at least one month. For example in Fig. 3, the same participant underwent 4 TUG tests in different seasons with different clothes and hairstyles; and the TUG tests vary a lot in time length due to the changes in rehabilitation states. TUG videos were recorded while the participants were asked to perform the 5 meter TUG test on the hard floor with an armless chair in front of a camera as is shown in Fig. 1. The frame rate of all the video sequences were set to be 25 frames per second (fps), indicating that the time duration of each frame is 40 ms. 127 TUG video sequences were finally collected to form the dataset in this study. The aforementioned

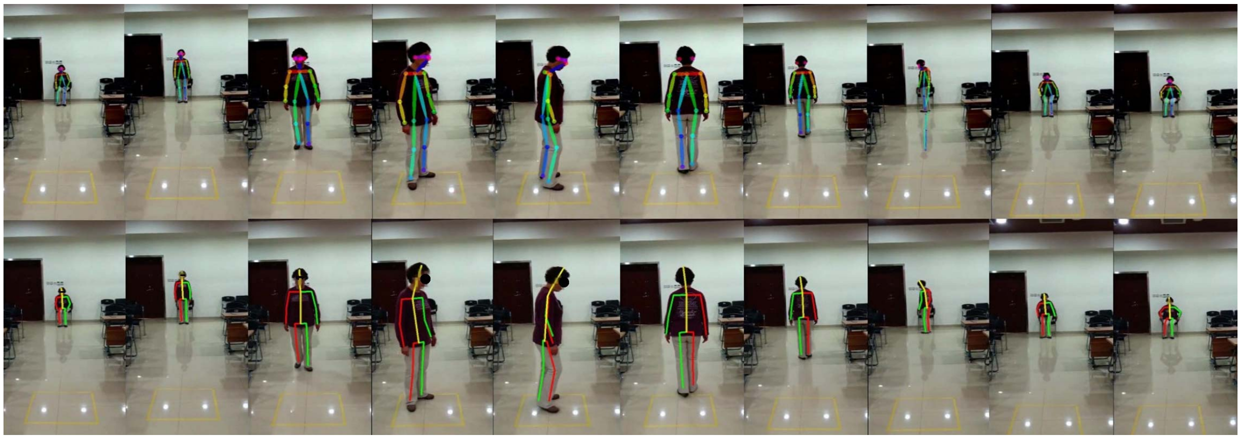


Fig. 4. Pose estimation examples in a TUG test video using OpenPose (First row) and IEF (Second row) respectively. 17(OpenPose) or 15(IEF) key points are estimated with high accuracy and stability. For simplicity, 9 common output key points of the two estimators are selected for further processing.

TABLE III
VISUAL GUIDELINES FOR THE SUB-TASK SEGMENTATION

Sub-Task	Label	Start Event
Sit-to-Stand	1	A. Lean forward, leave backrest B. Hip observably rise from chair
Walk	2	A. Stand straight, no more stand up movement B. First step forward
Turn	3	A. Arrive at turning point, stop walking B. Observable body rotation
Walk-Back	4	A. Finish 180 degree turn completely B. Observable backward velocity
Sit-Back	5	A. Arrive at chair, stop walking B. Observable body rotation

TABLE IV
SELECTED BODY KEY POINTS

Key Points	neck	R/L shoulder	R/L hip	R/L knee	R/L ankle
coordinates	$\vec{x}_1$	$\vec{x}_2, \vec{x}_3$	$\vec{x}_4, \vec{x}_5$	$\vec{x}_6, \vec{x}_7$	$\vec{x}_8, \vec{x}_9$

diversity in this dataset will help to promote the robustness and applicability of the proposed method.

To generate the segmentation ground truth, two experts were invited to segment the sub-tasks in all the 127 TUG videos independently. To ensure a common understanding, however, some guidelines for determining the starting points of the sub-tasks were presented to the experts. As is shown in Table III, a video frame was considered as the starting point of the corresponding sub-task when any one of the starting events occurred. For each sub-task in a TUG video, the two starting points (time) marked by the two experts were averaged as the ground truth. The intra-rater reliability between the labeling results of the two experts was high with the $ICC = 0.99$ (intra-class correlation coefficient). Following this, one of the labels shown in the second column of Table III was assigned to each frame to denote which sub-task the frame belongs to. Label 0 denotes the 'Sit' sub-task at the beginning of the TUG test, which is omitted in Table III.

B. Human Pose Estimation

It is possible to perform human activity classification directly based on images or video frames [19], [20]. However, such methods usually require a large quantity of training data which cannot be easily acquired in this study. The TUG test video only contains simple human activities which can possibly be distinguished by using low dimensional features such as the human pose. To obtain the human pose information, a pose estimator can be used to detect participants' body key point coordinates frame by frame. Iterative Error Feedback (IEF) [15] and OpenPose [14] are the two deep learning based human pose estimators explored in our study. Fig. 4 shows some human pose estimation results using IEF and Openpose respectively. To enhance the visibility, the body key points are connected by straight lines in Fig. 4. The IEF model detects 15 human key points and the OpenPose model detects 17 human key points. According to the start events listed in Table III, we selected $K = 9$ common body key points listed in Table IV for representing the human poses. Given a TUG test video sequence consisting of N frames, a vector $X^i = (\vec{x}^i_1, \vec{x}^i_2, \ldots, \vec{x}^i_K)$, $(i = 1, 2, \cdots, N)$ is used to denote the coordinates of the K detected body key points in the i_{th} frame. The subscripts of the coordinates of different body key points are shown in Table IV.

1) Iterative Error Feedback: The IEF model requires an input RGB image with the human body centered in the image [15]. However, this cannot be ensured by assuming that the TUG test video is recorded in a semi-controlled environment. Therefore, before applying the IEF, a human detector is used to locate the position of the human body. We chose the Faster R-CNN [21] detector to detect the body area which was then cropped out as the input to the IEF model. The basic idea of the IEF method is to iteratively correct the positions of the estimated body key points. In each iteration, a Convolution Neural Network (CNN) operates on an augmented input space created by concatenating the input RGB image with a visual representation of the detected key point positions in the last iteration to predict a correction that brings the key points closer to the ground truth. The correction is then applied to the detected point positions to generate new point positions for the current iteration. This procedure is repeated iteratively and each iteration is expected to produce a more accurate estimation of the body key point positions than the previous iteration. For the first iteration, the body key points are set to fixed positions. In the following iterations, the visual representation of detected key points is generated using 2D-Gaussian functions centered at each key point with fixed standard deviations.

2) OpenPose: OpenPose is a library for real-time multi-person key point detection [14]. It takes a RGB image as input and produces the 2D positions of body key points for each person in the image. In OpenPose, a deep CNN is used to jointly predict the confidence maps of body part locations and a set of 2D vector fields of part affinities, which encode the degree of association between different body parts. Similar to the IEF model, the CNN in OpenPose works in several stages. In the first stage it produces a rough prediction of the confidence maps and the affinity fields from the input RGB image. For the following stages the CNN takes the prediction of the last stage together with the input RGB image to produce refined predictions. Finally, the confidence maps and the affinity fields are parsed by a greedy inference process to generate the body key points positions for each person in the image. OpenPose can take raw video frames as input and doesn't need additional detectors for locating the human body. Compared to IEF, OpenPose is more advantageous in terms of the key point detection accuracy. The PCKs (Probability of Correct Keypoint) on the MPII dataset [22] are 89.3% and 81.0% for OpenPose and IEF respectively.

C. Sub-Task Segmentation

Based on the detected body key point coordinates, a classifier can be used to predict the activity label l_i for each frame to denote which sub-task the frame may belong to. Two different machine learning models, namely the SVM and the LSTM were adopted as the classifier in our work. The SVM model is a direct implementation of the structure risk minimization inductive principle [23]. It is suitable for solving small sample problems by providing theoretically optimal generalization capability. An important property of SVM is that the model parameters can be solved through convex optimization so that any local optimum is guaranteed to be a global optimum. The LSTM model is a recursive artificial neural network that aims at dealing with the long-short term dependence for sequential inputs [24]. The LSTM parameters are usually solved through back propagation using algorithms such as the stochastic gradient descent (SGD) or Adam [25]. The global optimum of the model cannot be achieved due to the non-convex nature [24]. However, compared to the SVM model, the descriptive capability of the LSTM model is more powerful due to its deeper architecture. Different inputs were designed for these two kinds of classifiers.

1) Frame-Wise Classification by SVM: To achieve a high segmentation accuracy, it is crucial to correctly classifying frames near the boundary of two adjacent sub-tasks. Essentially, this is very difficult when only using the single frame information for classification. For example, a frame A near the end of the sub-task 'Sit-to-Stand' can be visually very similar to a frame B at the beginning of the sub-task 'Walk' since both frames contain a person standing still on the ground. This kind of ambiguity can be solved by considering the temporal information provided by neighboring frames. Intuitively, frame A is more likely to be immediately after frames containing a bending body while frame B is more likely to be immediately before frames containing a rising foot. Based on this consideration, we adopted a sliding window based approach for generating the input to the classifier. For the i_{th} frame, the detected key point coordinates of a total of 9 frames $Y^i = \{X^{i-4r}, X^{i-3r}, \dots, X^{i+3r}, X^{i+4r}\}$ were used to form the classifying feature, where r is a adaptive factor decided by the total number of frames N in order to level out the variance of movement speed among different participants. We empirically set $r = N/(20 \times FPS)$, where FPS is the frame rate of the TUG video. As we have mentioned before, $FPS = 25$ for all the TUG videos in our dataset. The window size 9 was experimentally determined according to the sub-task segmentation performance.

It can be observed from Fig. 4 that the sizes of the participant in different frames vary a lot when the distance between the camera and the participant changes over time. Also, the positions of the participants may be different across different videos taken in semi-controlled environments. These variations in the original values of the key point coordinates may affect the classification performance. To deal with this problem, we normalized the key point coordinates spatially before feeding them into the classifier. Specifically, for the i_{th} frame, the key point coordinates in its classifying feature Y^i are first centered by the coordinate of the neck in the i_{th} frame, namely $\vec{x}_1^i$; and are then scaled by the Euclidean distance between the neck and the right hip in the i_{th} frame, namely $|\vec{x}_1^i - \vec{x}_4^i|_2$. Eq. (1) and (2) show the normalized classifying feature $\bar{Y}^i$ for the i_{th} frame.

$$\bar{Y}^i = \{\bar{X}^{i-4r}, \bar{X}^{i-3r}, \dots, \bar{X}^{i+3r}, \bar{X}^{i+4r}\} \tag{1}$$

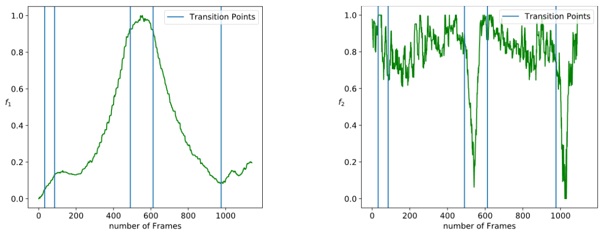
$$\bar{X}^{i+j*r} = \{\frac{(\vec{x}_t^{i+j*r} - \vec{x}_1^i)}{|\vec{x}_1^i - \vec{x}_4^i|_2}\}_{t=1,2,\dots,K}, \quad j = -4, -3, \dots, 3, 4 \tag{2}$$

The normalized feature $\bar{Y}^i$ contains a total of 81 coordinates of key points on the body. It is used as the input to a linear SVM based classifier. It should be noticed that the trained classifier may be inclined to predict the class which occurs more frequently in the training samples. In our study, the number of frames in different sub-tasks are seriously unbalanced. In order to level out the classification bias, we used a class weight inversely proportional to the occurrence frequency of each sub-task during SVM training [26]. The occurrence frequency of each sub-task and the corresponding class weight are shown in Table V.

2) Frame-Wise Classification by LSTM: Unlike common human activity videos, the sequence of actions and the trajectory of the human movement are highly similar among

TABLE V
OCCURRENCE FREQUENCIES AND CLASS WEIGHTS OF SUB-TASKS

Sub-task	Occurrence Frequency	Class Weight
Sit	0.045	22.14
Sit-to-Stand	0.110	9.13
Walk	0.337	2.97
Turn	0.184	5.44
Walk-Back	0.302	3.31
Sit-Back	0.222	4.49



(a) (b)

Fig. 5. Sample global temporal features, f_1 in (a) and f_2 in (b). Green curves show the features and blue lines indicate the ground truth transition points between adjacent sub-tasks.

different TUG video sequences. As such, the long term temporal information in the TUG video can be utilized to achieve more accurate activity classification results. Considering the capability of LSTM for analyzing the long term dependency in the input data, besides the normalized body key point coordinates $\bar{X}^i$ defined in Eq. (2), we introduced two extra global temporal features f_1 and f_2 to the LSTM input.

In the TUG test, participants are asked to try to walk along a straight line. Let $\vec{x}^i_{mankle} = (\vec{x}^i_8 + \vec{x}^i_9)/2$ be the middle point of two ankles in the i_{th} frame. Although there might be some deviation from the straight line trajectory caused by the mobility of the participants or by camera movements, the Euclidean distance between $\vec{x}^i_{mankle}$ and $\vec{x}^1_{mankle}$ is still highly correlated to the spatial distance between the participant and the chair where the TUG test starts. To eliminate the differences in the resolutions and video shooting distances among different TUG videos, we normalized $|\vec{x}^i_{mankle} - \vec{x}^1_{mankle}|$ to the range of $[0, 1]$ as Eq. (3). In a TUG video sequence, a small value of f^i_1 means that the participant is near the chair while a large value of f^i_1 indicates that the participant is close to the turning area. An example of f_1 extracted from a TUG video is shown in Fig. 5(a). It should be noted that f_1 may degenerate to a constant under some special video shooting angles. For example, $f_1 = 0$ when the optical axis of the video camera is collinear with the trajectory of the feet. However, this seldom happens in practice.

$$f^i_1 = \frac{|\vec{x}^i_{mankle} - \vec{x}^1_{mankle}|}{\max\limits_{j=1,2,\ldots,N}\{|\vec{x}^j_{mankle} - \vec{x}^1_{mankle}|\}} \tag{3}$$

Secondly, the orientation of the participant's body can be used to effectively distinguish the sub-tasks 'Turn' and 'Sit-Back' from other sub-tasks. However, accurately calculating the body orientation may require 3D reconstruction of the participant which can be highly complicated. For sub-task segmentation, a coarse estimation should be good enough. Noticing that the width of the hip changes with the body rotation in the TUG video, we used the distance between the left and right hip points as a simplified representation for the rotation angle of the participant's body. Similarly, we normalized this distance as is shown in Eq. (4). Again, the body length $|\vec{x}^i_1 - \vec{x}^i_4|$ is used for the normalization in order to level out the differences caused by the variations of the body sizes in the videos. Roughly speaking, f_2 has a cosine relation with the rotation angle of the participant's body. $f_2 \rightarrow 1$ indicates that the participant is facing or back facing the camera; and $f_2 \rightarrow 0$ indicates that the participant is side facing the camera. An example of f_2 extracted from a TUG video is shown in Fig. 5(b).

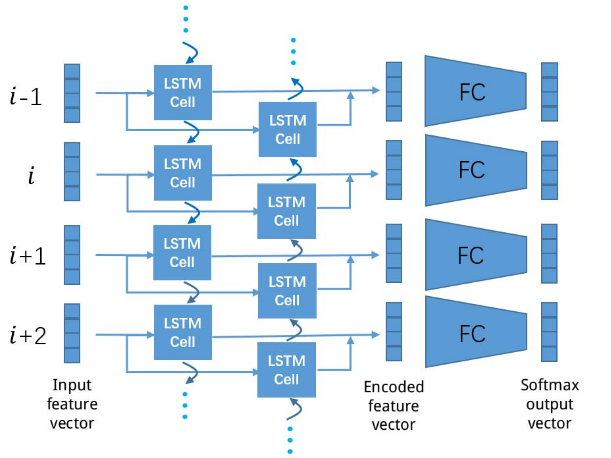


Fig. 6. The LSTM architecture in which arrows indicate the flow of data.

$$f^i_2 = \frac{\frac{|\vec{x}^i_4-\vec{x}^i_5|}{|\vec{x}^i_1-\vec{x}^i_4|} - \min\limits_{j=1,2,\ldots,N}\{\frac{|\vec{x}^j_4-\vec{x}^j_5|}{|\vec{x}^j_1-\vec{x}^j_4|}\}}{\max\limits_{j=1,2,\ldots,N}\{\frac{|\vec{x}^j_4-\vec{x}^j_5|}{|\vec{x}^j_1-\vec{x}^j_4|}\} - \min\limits_{j=1,2,\ldots,N}\{\frac{|\vec{x}^j_4-\vec{x}^j_5|}{|\vec{x}^j_1-\vec{x}^j_4|}\}} \tag{4}$$

Together with the normalized body key point coordinates $\bar{X}^i$, the two global temporal features f_1 and f_2 are used as the input of an LSTM model of which the architecture is shown in Fig. 6. In order to combine the temporal information both in the positive and the negative time direction, we adopted a bi-directional LSTM architecture for encoding the input features [27]. The encoded features are then sent to a fully-connected layer with the softmax function for predicting the frame labels. A softmax cross-entropy loss function L was adopted as defined in Eq. (5), in which $\vec{y}_i$ is the one-hot label vector of the i_{th} frame; and $\vec{p}_i$ is the corresponding softmax output of the LSTM network. Both $\vec{y}_i$ and $\vec{p}_i$ are 6-dimensional vectors in which each dimension corresponds to a sub-task.

$$L = -\sum_{i=1}^{N} \vec{y}_i \cdot \log \vec{p}_i \tag{5}$$

Considering the similarity between the frames of adjacent sub-tasks described above, the misclassification between adjacent sub-tasks is somewhat tolerable. Comparatively speaking,

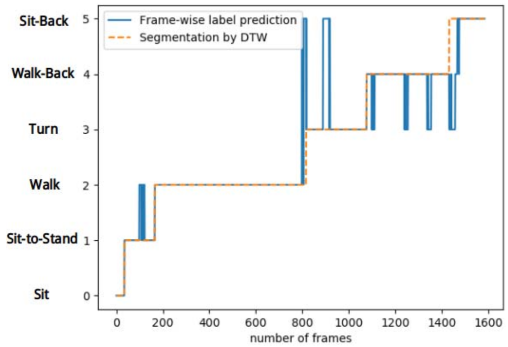


Fig. 7. A sample sub-task segmentation result using DTW.

other types of misclassification should be avoided as much as possible. Therefore, we introduced a punishment factor λ_i to punish misclassification among frames of non-adjacent sub-tasks. If $|\arg\max \vec{y}_i - \arg\max \vec{p}_i| \geq 2$, which means our model classifies a frame F_i into a wrong sub-task not adjacent to the ground truth sub-task, we set $\lambda_i = 10$ to punish this situation, and $\lambda_i = 1$ otherwise. Besides the misclassification punishment factor, the class weight same as the SVM based method was also used. The product of the two factors was used to weigh the softmax cross-entropy loss as is shown in Eq. (6), in which the class weight w_i corresponds to the ground truth sub-task of the i_{th} frame. The class weight value of each sub-task is shown in Table V. The weighted softmax cross-entropy loss brought about more stable convergence in the training and higher accuracy in the test than the one without weighting.

$$L_w = -\sum_{i=1}^{N} \lambda_i w_i \vec{y}_i \cdot \log \vec{p}_i \tag{6}$$

During the training procedure, a stochastic gradient descent algorithm with the learning rate equals 0.0001 was applied to minimize the loss function. To avoid overfitting, a dropout algorithm with probability 0.5 was used on the LSTM cell [28], and the weight decay was used in the fully connected layer with the decay rate equals 0.003.

3) Sub-Task Segmentation by DTW: After the classifier output a prediction label l_i for each frame, the transition points between sub-tasks still has to be determined based on the frame-wise predictions. Considering the strict time order of TUG sub-tasks, we performed the dynamic time warping (DTW) [29] algorithm to find the most appropriate sub-task segmentation. The main idea of the DTW algorithm is to use a dynamic programming procedure to find a best path that is non-decreasing on l_i along the time axis under a pre-defined criterion. The algorithm can correct most of the frame classification errors and determine the optimal transition points based on the frame-wise prediction, as is illustrated in Fig. 7. Formally, given a frame-wise prediction $P = \{l_1, l_2, \ldots, l_N\}$, the final classification result $P' = \{l_1', l_2', \ldots, l_N'\}$ will be computed using DTW. Due to the strict time order of TUG sub-tasks, P' should be non-decreasing as is described in Eq. (7). At the same time, P' should be as close to P as possible in terms of the cost function defined in Eq. (8). The minimization of $C(P, P')$ can be efficiently solved using dynamic programming. Finally, a segmentation well matching the frame-wise classification result, while simultaneously fitting in the time order of TUG sub-tasks can be directly generated from P'.

$$l_i' \in \{l_{i-1}', \quad l_{i-1}' + 1\}, \; l_1' = 0, \; l_N' = 5 \tag{7}$$

$$C(P, P') = \sum_{i=1}^{N} h(l_i, l_i'), \quad h(l_i, l_i') = \begin{cases} 0 & l_i = l_i' \\ 1 & l_i \neq l_i' \end{cases} \tag{8}$$

IV. Experimental Results

We tested all four possible algorithm combinations in our method, namely IEF + SVM (I + S), IEF + LSTM (I + L), OpenPose + SVM (O + S) and OpenPose + LSTM (O + L) on the TUG video dataset described in Section III-A. For quantitative evaluations, we deployed the 5-fold cross validation procedure. The 127 TUG video sequences were evenly divided into 5 folds according to the identities of participants, so that each fold contained 24-26 videos from 4-6 participants and videos of one participant only belongs to one fold. In each cross validation round, one of the 5 folds was used for testing while the others were used for training the classification models. Average results of all the 5 rounds are reported for the performance metrics. To evaluate the sub-task segmentation performance, we mainly studied two categories of metrics: the frame classification accuracy and the sub-task timing accuracy.

On frame classification accuracy, we calculated the $precision$ ($prec$), $recall$ (rec) and $F1$ score for each sub-task; and the total classification accuracy (acc) for the video sequence. The $acc.$ is defined as the percentage of the frames correctly classified in a video sequence; and the other three metrics are defined in Eq. (9), in which TP,FP,FN stands for the number of true positive, false positive and false negative frames for a given sub-task respectively. It should be noticed that the metric $sensitivity$ used in some previous work [8] is equivalent to the acc in our work.

$$\begin{aligned} prec &= TP/(TP + FP), \quad rec = TP/(TP + FN) \\ F1 &= 2 \times (prec \times rec)/(prec + rec) \end{aligned} \tag{9}$$

In the TUG test, knowing the time duration of each sub-task is critical for the evaluation or rehabilitation purposes. We therefore introduced the $Time\ Error$ (TE) for evaluating the sub-task timing accuracy. For a sub-task, TE is defined as the absolute difference between its time duration estimated by our method and that by experts. For a TUG video, TE is defined as the sum of the TEs of all the sub-tasks. As the lengths of TUG videos varies, we further introduce the $Relative\ Time\ Error$ (RTE) which is defined as the overall TE of a video divided by its length.

Fig. 8 shows sample sub-task segmentation results. Videos 1 and 2 are of the similar length (∼25 seconds), indicating that the participants finished the test quite fluently. Video 3 is the longest (∼267 seconds) sequence in our dataset, indicating that the participant was probably suffering from a severe dyskinesia. Some of the footsteps are too subtle to be recognized, leading to the poor segmentation performance. It can

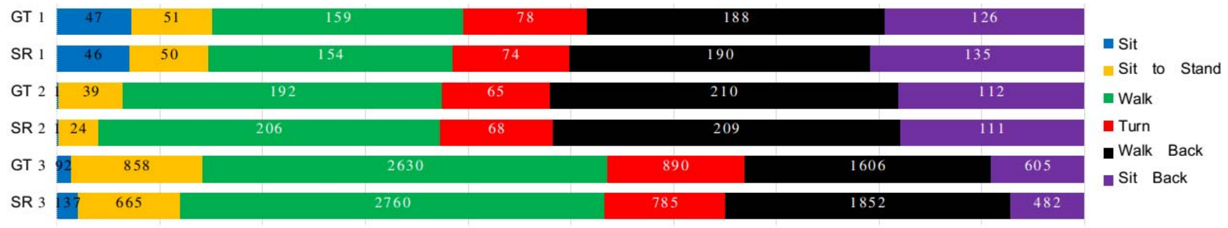


Fig. 8. Sample sub-task segmentation results (using O + L) of three selected TUG videos. GT stands for *G*round *T*ruth labeled by experts; and SR stands for *S*egmentation *R*esult acquired using the proposed method. Different colors are used to represent different sub-tasks; and the number of frames for each sub-task is also shown. For better visualization, the results are normalized to the same length although the number of frames for different video sequences are different. For the 'Sit-to-Stand' sub-task in video 1, $TP = 49$, $FP = 1$, $FN = 2$, so that $prec = 49/(49+1) = 98.0\%$, $rec = 49/(49+2) = 96.1\%$, $F1 = 97.0\%$. For the three videos, the *acc*s are 95.3%, 96.9%, 93.2%; the *TEs* are 0.88s, 1.36s, 33.68s; and the *RTEs* are 3.4%, 5.5%, 12.6%.

TABLE VI
CLASSIFICATION RESULTS FOR DIFFERENT ALGORITHM COMBINATIONS

	I+S			I+L			O+S			O+L		
Sub-Task	$prec\%$	$rec\%$	$F1\%$	$prec\%$	$rec\%$	$F1\%$	$prec\%$	$rec\%$	$F1\%$	$prec\%$	$rec\%$	$F1\%$
Sit	70.8	93.6	80.6	76.4	90.2	82.7	78.4	91.7	84.5	**81.6**	**89.0**	**85.1**
Sit-to-Stand	81.1	78.1	79.6	77.6	85.4	81.3	82.7	81.0	81.8	**80.8**	**86.3**	**83.5**
Walk	96.5	92.3	94.3	97.4	90.6	93.9	**96.5**	**93.9**	**95.2**	96.7	93.5	95.1
Turn	89.5	86.3	87.9	83.3	95.4	89.0	**87.8**	**90.3**	**89.0**	84.4	92.4	88.2
Walk-Back	95.4	94.0	94.7	**97.4**	**94.9**	**96.1**	95.5	93.8	94.7	96.6	94.4	95.5
Sit-Back	95.1	98.2	96.6	96.6	97.9	97.2	95.3	98.1	96.7	**97.5**	**97.0**	**97.3**
Average $acc(\%)$		91.9			92.7			92.8			**93.1**	

TABLE VII
TIME ERRORs FOR DIFFERENT ALGORITHM COMBINATIONS

Sub-Task	I+S(sec.)	I+L(sec.)	O+S(sec.)	O+L(sec.)
Sit	0.45	0.49	0.47	**0.40**
Sit-to-Stand	1.30	1.17	1.09	**1.02**
Walk	1.17	1.19	**1.03**	1.07
Turn	0.96	0.68	0.77	**0.58**
Walk-Back	0.92	0.67	0.84	**0.66**
Sit-Back	0.48	0.35	0.40	**0.32**
Total	5.18	4.56	4.59	**4.05**

TABLE VIII
PERFORMANCE COMPARISON AMONG DIFFERENT METHODS

Methods	$sensitivity$ or acc (%)	TE (sec.)	RTE (%)
[5]	-	1.32	11.2
[8]	81.80	-	-
[7]	-	0.55	-
[17]	-	**0.52**	-
[4]	-	2.54	16.4
I+S (ours)	91.9	5.17	13.5
I+L (ours)	92.7	4.56	11.9
O+S (ours)	92.8	4.59	12.0
O+L (ours)	**93.1**	4.05	**10.6**

be observed that the two categories of metrics are related but not completely consistent. Numerical classification accuracies of the 5 fold test are shown in Table VI. In each row, results of the algorithm combination corresponding to the highest $F1$ or acc are highlighted. The classification accuracies for the sub-tasks 'Walk,' 'Walk-Back,' and 'Sit-Back' are apparently higher than that of the other three sub-tasks. Sub-tasks 'Sit' and 'Sit-to-Stand' are relatively difficult to segment because the duration of these two sub-tasks are often too short to provide sufficient training samples. As for the sub-task 'Turn,' the intrinsic ambiguity in its beginning and ending points leads to the low segmentation accuracy. Generally speaking, O + L out-performs the other three algorithm combinations in most cases. Similar conclusions can be drawn from the Time Errors shown in Table VII.

We also compared the performance of our method to that of some previous methods, in which either $sensitivity$ or the overall TE was commonly reported as the evaluation metric. Considering that the average time duration of the TUG test videos in our dataset is much longer than that in previous works, we also list RTE in Table VIII for fair comparison. In terms of the $sensitivity$ (acc) or the RTE, our method even out-performs some of the previous methods in which ambient/inertial sensors [4], [8] or depth cameras [5] were used. The TEs in [7] and [17] are much smaller than ours. However, their methods used 17 inertial sensors, leading to inconveniences in the practical implementation. Also, since the average time duration of the TUG tests was not presented in [7] and [17] their $RTEs$ cannot be estimated.

Investigating the improvements in the TUG sub-task time durations is a common practice for evaluating the effect of the PD treatment such as the DBS. Suppose the time

TABLE IX
ESTIMATED SUB-TASK TIME DURATIONS AND *trrs* FOR A PD PATIENT TREATED BY DBS

Sub-Task (i)	t_i by experts (sec.)	t'_i by experts (sec.)	t_i by O+L (sec.)	t'_i by O+L (sec.)	trr_i by experts	trr_i by O+L
Walk (2)	7.24	6.60	6.76	6.08	9%	10%
Turn (3)	3.84	1.88	3.84	1.76	51%	54%
Walk-Back (4)	7.88	6.48	8.68	6.20	18%	29%
Sit-Back (5)	7.80	3.04	7.28	3.28	61%	55%

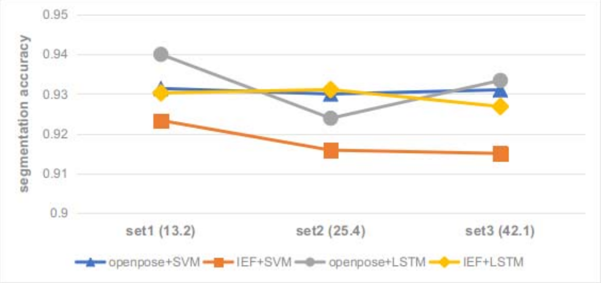


Fig. 9. Segmentation performances for different UPDRS III scores.

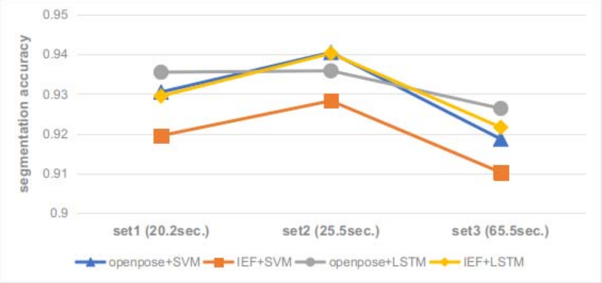


Fig. 10. Segmentation performances for different TUG time durations.

durations for a PD patient to finish a TUG sub-task i are t_i and t'_i respectively before and after the DBS, we define the time reduction rate as $trr_i = (t_i - t'_i)/t_i$ to quantitate the improvements. To further validate the effectiveness of our proposal method to detect these improvements, we compared the time durations and $trrs$ estimated by experts and our method (O + L) for eight PD patients before and after the DBS surgery. We omitted the sub-tasks 'Sit' and 'Sit-to-Stand' because their time reductions are usually much less obvious comparing to the other sub-tasks. Table IX shows the comparison results for one PD patient, for whom the most obvious time reductions come from sub-tasks 'Turn' and 'Sit-Back' according to the human labeled sub-task time durations, indicating that the DBS treatment had substantially improved the patient's ability in completing the turning activity. The same conclusion can also be drawn from the estimated results using our method, implying that the segmentation accuracy is high enough at least for this specific clinic evaluation task. We computed the correlation coefficients between the $trrs$ estimated by experts and by our method as per sub-tasks for all the eight PD patients, and the results are 0.99 ('Walk'), 0.93 ('Turn'), 0.98 ('Walk-back') and 0.98 ('Sit-back'). This means that statistically our method complies well with human experts in the sub-task duration estimation.

Intuitively, automatic TUG analysis becomes more difficult for patients with more severe motor impairments due to ambiguous movements caused by trembling or freezing. We therefore further explored the robustness of our method towards different levels of PD severity. We sorted the videos according to the participants' UPDRS III scores and then divided them evenly into three subsets in order. The average UPDRS III scores for the three subsets are 13.2, 25.4 and 42.1 respectively. The average accs for the three subsets shown in Fig. 9 indicate that the segmentation performance of our method is quite stable to the changes in the UPDRS III score. Considering that most items in the UPDRS III score are not decided by the TUG test, we reordered the videos according to the TUG time duration and performed the experiment again. The average TUG time durations in the three subsets are 20.2$sec.$, 25.5$sec.$ and 65.5$sec.$ It can be observed from Fig. 10 that the proposed method performs the best on TUG videos of moderate lengths. The segmentation performance drops slightly for very short or very long videos. This can be understood considering the difficulty in frame classification for too fast or too slow motions. Overall, the experimental results indicate that our method is robust to the variation of patients' PD severity.

TABLE X
HARDWARE REQUIREMENTS OF DIFFERENT METHODS

Method	Hardware Requirements
[5]	2 Microsoft Kinects.
[7], [17]	A motion capture suit with 17 inertial sensors attached.
[8]	2 inertial sensors attached on shoes.
[4]	2 light batteries, 4 force sensors, and 1 laser range sensors.
Ours	**1 video camera.**

V. DISCUSSIONS ON PROS AND CONS

One of the major advantages of our method is that it is convenient to deploy. The proposed method is totally based on visual information and the only device required is an ordinary video camera which can be easily found, for example, in almost every modern mobile phone. Table X compares the hardware requirements of different methods. The TUG test video can be recorded by the patients' family at home following the video recording protocol shown in Fig. 1, and can be analyzed remotely and automatically using our method.

Comparing to that in previous work, the dataset in our study contains real TUG videos from more participants, as is shown in Table XI. This has enabled us to incorporate machine learning approaches, so that comprehensive activity patterns

TABLE XI
PARTICIPANTS OF DIFFERENT METHODS

Methods	Participants
[5]	5 elderly and 4 young healthy people.
[7]	16 elderly healthy people.
[17]	12 PD patients (mean score 0.08 of UPDRS gait item)
[8]	5 PD patients (mean UPDRS score < 17)
[4]	5 healthy people.
Ours	24 PD patients (mean UPDRS III score > 31)

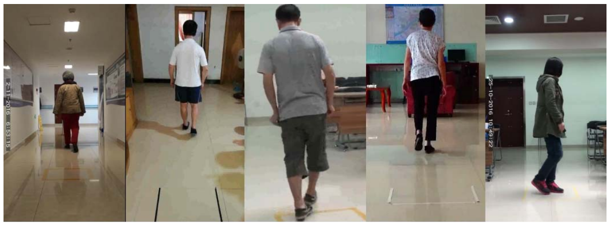


Fig. 11. Background variations in our dataset. Previous work using 2D vision based approaches usually used fixed background like a giant white curtain [10]. Our method is invariant to the background variation largely due to the high robustness of the state-of-the-art human pose estimators.

in TUG tests can be automatically learned to facilitate more accurate sub-task segmentations. The degree of motor disorder among the our participants are more diverse than that in previous work. As is shown in Table XI, in many previous work, the participants were healthy people or PD patients with relatively slight motor disorders, while some of our participants suffered from very severe motor disability. As such, the time which patients took to finish the TUG tests vary from around ten seconds to several minutes in our dataset. In addition, the average time duration is 38.2 sec., which is much longer than that in previous studies (for example [4, Sec. 15.6] and [5, Sec. 13.0]). In addition, different from previous video based methods in which videos were captured in a fixed background, there are various backgrounds existing in our dataset as is shown in Fig. 11. Videos in our dataset covers a large range of situations that may happen in real life TUG tests rather than in laboratory environments. This enables the proposed method to generalize well in practice.

Nevertheless, there are still some fail cases. In some TUG videos, the participants could not finish the TUG test fluently. Typically, some participants stopped several times during the whole test, leading to the increase in the frame-wise classification errors. For short stops, the DTW algorithm is still able to correct most of the incorrect labels and generate a reasonable segmentation result. However for long stops, DTW also fails, leading to very poor sub-task segmentation results. Fig. 12 shows a typical fail case caused by two very long stops in the TUG video. One stop is during the 'Sit-to-Stand' sub-task and the classifier assigns the label 'Sit' to many frames during the stop, so that the estimated 'Sit' sub-task is significantly longer than the ground truth. Another stop is at the beginning of the 'Sit-Back' sub-task and the classifier mistake a considerable portion of the stop as in the 'Walk-Back' sub-task. The two stops are too longer so that the misclassified labels cannot be corrected by DTW. The segmentation accuracy of this video is only $acc = 47.2\%$, which is much lower than the average. The time durations of 'Sit,' 'Sit-to-Stand,' 'Walk-back' and 'Sit-back' all deviate significantly from the ground truth.

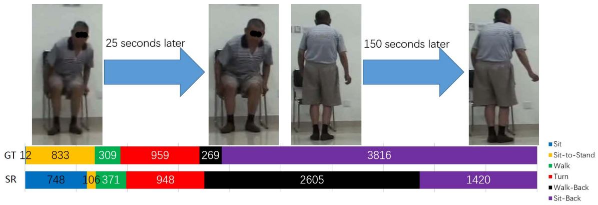


Fig. 12. A fail case caused by the long time stops due to gait freezing.

VI. CONCLUSIONS

We propose a fully-automated TUG sub-task segmentation method for TUG videos. The use of deep learning based human pose estimation technology makes the proposed method much easier to be applied in practice than previous work. Our proposal can possibly help doctors to monitoring PD patients' rehabilitation states remotely and automatically on a regular basis. In the future, we will focus on video based analysis of other standardized motor function tests included in the UPDRS III; and symptom detection of tremor and gait freezing for PD patients. We believe that this work can well serve as a foundation towards a fully-automated UPDRS III evaluation.

VII. ACKNOWLEDGMENT

The authors would like to thank Chen Yao, Wai Kin Daniel Ko, Ling Yang, Xi Fang and Yanan Sui for their contributions.

REFERENCES

[1] D. Podsiadlo and S. Richardson, "The timed 'Up & Go': A test of basic functional mobility for frail elderly persons," *J. Amer. Geriatrics Soc.*, vol. 39, no. 2, pp. 142–148, 1991.

[2] A. Salarian, F. B. Horak, C. Zampieri, P. Carlson-Kuhta, J. G. Nutt, and K. Aminian, "iTUG, a sensitive and reliable measure of mobility," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 18, no. 3, pp. 303–310, Jun. 2010.

[3] A. Ziegl, R. Modre-Osprian, A. Sánchez, M. Falgenhauer, P. Kastner, and G. Schreier, "Timed Up-and-Go device for unsupervised functional assessment of elderly patients," *Stud. Health Technol. Inform.*, vol. 236, no. 1, p. 298, 2017.

[4] T. Frenken, B. Vester, M. Brell, and A. Hein, "aTUG: Fully-automated Timed Up and Go assessment using ambient sensor technologies," in *Proc. Int. Conf. Pervas. Comput. Technol. Healthcare*, 2011, pp. 55–62.

[5] O. Lohmann, T. Luhmann, and A. Hein, "Skeleton Timed Up and Go," in *Proc. IEEE Int. Conf. Bioinf. Biomed.*, Oct. 2012, pp. 1–5.

[6] N. Kitsunezaki, E. Adachi, T. Masuda, and J.-I. Mizusawa, "Kinect applications for the physical rehabilitation," in *Proc. IEEE Int. Symp. Med. Meas. Appl.*, May 2013, pp. 294–299.

[7] H. P. Nguyen *et al.*, "Auto detection and segmentation of physical activities during a Timed-Up-and-Go (TUG) task in healthy older adults using multiple inertial sensors," *J. Neuroeng. Rehabil.*, vol. 12, no. 1, p. 36, Jan. 2015.

[8] S. Reinfelder, R. Hauer, J. Barth, J. Klucken, and B. M. Eskofier, "Timed Up-and-Go phase segmentation in Parkinson's disease patients using unobtrusive inertial sensors," in *Proc. Conf. IEEE Eng. Med. Biol. Soc.*, Aug. 2015, pp. 5171–5174.

[9] D. Berrada, M. Romero, G. Abowd, M. Blount, and J. Davis, "Automatic administration of the Get Up and Go test," in *Proc. Int. Conf. Mobile Syst., Appl., Services*, 2007, pp. 73–75.

[10] Z. Skrba *et al.*, "Objective real-time assessment of walking and turning in elderly adults," in *Proc. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Sep. 2009, pp. 807–810.

[11] F. Wang, M. Skubic, C. Abbott, and J. M. Keller, "Quantitative analysis of 180 degree turns for fall risk assessment using video sensors," in *Proc. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Aug./Sep. 2011, pp. 7606–7609.

[12] A. Weiss, A. Mirelman, A. S. Buchman, D. A. Bennett, and J. M. Hausdorff, "Using a body-fixed sensor to identify subclinical gait difficulties in older adults with IADL disability: Maximizing the output of the Timed Up and Go," *PLoS ONE*, vol. 8, no. 7, p. e68885, 2013.

[13] M. R. Adame *et al.*, "TUG test instrumentation for Parkinson's disease patients using inertial sensors and dynamic time warping," *Biomed. Eng.*, vol. 57, no. SI-1 Track-E, pp. 1071–1074, 2012.

[14] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2016, pp. 7291–7299.

[15] J. Carreira, P. Agrawal, K. Fragkiadaki, and J. Malik, "Human pose estimation with iterative error feedback," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4733–4742.

[16] G. Sprint, D. J. Cook, and D. L. Weeks, "Toward automating clinical assessments: A survey of the Timed Up and Go," *IEEE Rev. Biomed. Eng.*, vol. 8, pp. 64–77, 2015.

[17] H. Nguyen, K. Lebel, P. Boissy, S. Bogard, E. Goubault, and C. Duval, "Auto detection and segmentation of daily living activities during a Timed Up and Go task in people with Parkinson's disease using multiple inertial sensors," *J. Neuroeng. Rehabil.*, vol. 14, no. 1, p. 26, 2017.

[18] C. G. Goetz, "The unified Parkinson's disease rating scale (UPDRS): Status and recommendations," *Movement Disorders*, vol. 18, no. 7, pp. 738–750, 2003.

[19] H. Wang, D. Oneata, J. Verbeek, and C. Schmid, "A robust and efficient video representation for action recognition," *Int. J. Comput. Vis.*, vol. 119, no. 3, pp. 219–238, 2016.

[20] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 1933–1941.

[21] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[22] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, "2D human pose estimation: New benchmark and state of the art analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 3686–3693.

[23] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York, NY, USA: Springer, 2006.

[24] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016. [Online]. Available: http://www.deeplearningbook.org

[25] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, 2015, pp. 1–15.

[26] H. He and Y. Ma, Eds., *Imbalanced Learning: Foundations, Algorithms, and Applications*. Hoboken, NJ, USA: Wiley, 2012.

[27] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Trans. Signal Process.*, vol. 45, no. 11, pp. 2673–2681, Nov. 1997.

[28] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.

[29] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-26, no. 1, pp. 43–49, Feb. 1978.