Vision-Based Finger Tapping Test in Patients With Parkinson's Disease via Spatial-Temporal 3D Hand Pose Estimation

Zhilin Guo, Weiqi Zeng, Taidong Yu, Yan Xu, Yang Xiao, Xuebing Cao, and Zhiguo Cao, *Member, IEEE*

***Abstract*—Finger tapping test is crucial for diagnosing Parkinson's Disease (PD), but manual visual evaluations can result in score discrepancy due to clinicians' subjectivity. Moreover, applying wearable sensors requires making physical contact and may hinder PD patient's raw movement patterns. Accordingly, a novel computer-vision approach is proposed using depth camera and spatial-temporal 3D hand pose estimation to capture and evaluate PD patients' 3D hand movement. Within this approach, a temporal encoding module is leveraged to extend A2J's deep learning framework to counter the pose jittering problem, and a pose refinement process is utilized to alleviate dependency on massive data. Additionally, the first vision-based 3D PD hand dataset of 112 hand samples from 48 PD patients and 11 control subjects is constructed, fully annotated by qualified physicians under clinical settings. Testing on this real-world data, this new model achieves 81.2% classification accuracy, even surpassing that of individual clinicians in comparison, fully demonstrating this proposition's effectiveness. The demo video can be accessed at https://github.com/ZhilinGuo/ST-A2J.**

***Index Terms*—3D kinematic feature, depth camera, finger tapping test, MDS-UPDRS, Parkinson's disease, spatial-temporal 3D hand pose estimation, SVM.**

Manuscript received 18 August 2021; revised 6 March 2022 and 20 March 2022; accepted 21 March 2022. Date of publication 29 March 2022; date of current version 9 August 2022. This work was supported in part by the National Key R&D Program of China under Grant 2017YFC1310300, and in part by the National Natural Science Foundation of China under Grants 61502187, 81974200, 8187051653, and 81671108. *(Zhilin Guo, Weiqi Zeng, and Taidong Yu contributed equally to this work). (Corresponding authors: Zhiguo Cao; Xuebing Cao.)*

Zhilin Guo is with the Key Laboratory of Image Processing and Intelligent Control, Ministry of Education; School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan 430074, China, and also with the Fu Foundation School of Engineering and Applied Science, Columbia University, New York, NY 10027 USA (e-mail: zg2358@columbia.edu).

Taidong Yu, Yang Xiao, and Zhiguo Cao are with the Key Laboratory of Image Processing and Intelligent Control, Ministry of Education; School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan 430074, China (e-mail: taidongyu@hust.edu.cn; yang_xiao@hust.edu.cn; zgcao@hust.edu.cn).

Weiqi Zeng, Yan Xu, and Xuebing Cao are with the Department of Neurology, Union Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan 430022, China (e-mail: d201881482@hust.edu.cn; xuyanwxf@126.com; caoxuebing@126.com).

Digital Object Identifier 10.1109/JBHI.2022.3162386

I. Introduction

PARKINSON'S Disease (PD) is the second most common neurodegenerative disorder, affecting approximately 7 to 10 million people worldwide [1]. The main symptoms of PD are bradykinesia, rest tremor, rigidity, and postural instability [2]. In practice, physicians use a mixture of different examinations to evaluate these symptoms, including the finger tapping test, gait analysis, and speech impairment examination [3]. Among all PD symptoms, bradykinesia is the most essential according to the guidelines published by the Movement Disorder Society [4], and the finger tapping test (FTT) is one of the most widely used tests to assess bradykinesia [5]. During FTT, a PD patient is asked to tap the tip of the thumb and index finger in rapid succession with the largest possible amplitude for 10 times. Then, following Item 3.4 of *Movement Disorders Society Modified Unified Parkinson's Disease Rating Scale* (MDS-UPDRS) [3], a qualified clinician evaluates this hand movement by speed, amplitude, and rhythm to give a score ranging from 0 (normal) to 4 (severe). However, even though a trained clinician can supervise FTT and evaluate PD effectively, results may differ among different clinicians due to interobserver variability [6]. And more importantly, it is difficult for the naked eye to sequentially quantify metrics such as the amplitude and speed of a patient's finger movement. These metrics can be beneficial to clinicians in PD evaluations such as in early-stage PD detections or before-and-after-treatment comparisons [7]. Thus, there is need for machine-based, objective, and quantifiable evaluation for the UPDRS finger tapping test in PD evaluation [8].

To quantify FTT movement, some existing works proposed to use wearable sensors [9]. However, these sensors inevitably make physical contact with PD patient's hand which may impact the patient's movement due to sensors' self-weight and contact, and they require more complex calibration procedures [10]. Some vision-based approaches leveraged colored markers or gloves [11], [12] to assist the vision system, but their contact with patient's hands could still alter the fine motor movement. Other vision based approaches [13]–[15] avoided making any contact by employing RGB cameras and measured the relative finger motion orientation, or the 2D pixel-level distance between the thumb and the index fingertips. But since their calculations are in 2D, the physical maximum opening amplitude in 3D which is important for FTT evaluation cannot be accurately estimated.

To address the above problems, we propose to leverage 3D vision in FTT. Particularly, depth camera (Intel RealSense

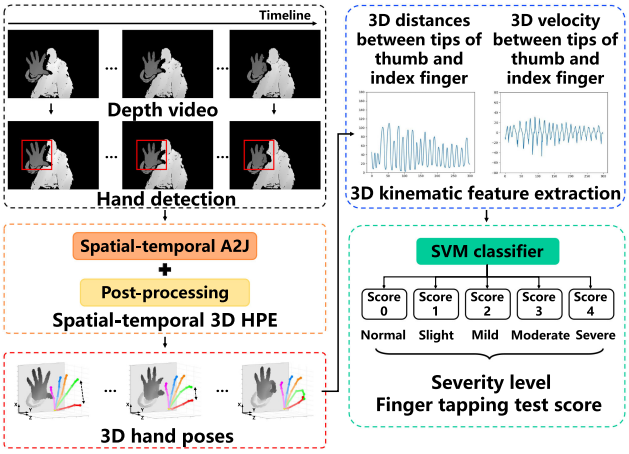


Fig. 1. The pipeline of the proposed vision-based finger tapping test approach in patients with Parkinson's Disease via spatial-temporal 3D Hand Pose Estimation (HPE).

SR300 [16]) is used to capture PD patient's spontaneous 3D hand movement in a contact-free manner. Then, spatial-temporal 3D hand pose estimation (HPE) is conducted on consecutive depth frames to predict 3D hand joint positions. Subsequently, PD patient's 3D hand kinematic features are extracted to execute supervised severity level classification with a Support Vector Machine (SVM). Advantages of this proposal include the utilization of depth camera that directly captures 3D spatial information, which is unavailable to methods using RGB cameras. Additionally, this novel spatial-temporal 3D HPE framework allows accurate estimations of 3D positions for all 21 keypoints of the hand throughout FTT, rather than finding intermediate metrics like the motion orientation of the index finger. As a result, our model can calculate the opening amplitude and frequency in 3D without using reference objects such as the patient's head, and these accurate results can be recorded for later comparisons after treatment. The main research idea of this proposition is shown in Fig. 1.

Many existing 3D HPE methods [17], [17]–[19] suffer from temporal jittering problem when estimating hand poses in videos [20], as they only extract information from a single image for each pose. This temporal jittering will severely disturb the estimation of the finger's movement amplitude, making motor impairment difficult to accurately detect. To alleviate this problem, we propose to extend A2J [17] previously developed by our team to Spatial-Temporal A2J (ST-A2J). Particularly, a temporal encoding module is constructed for ST-A2J to incorporate richer temporal contextual information while retaining the end-to-end running capacity under the deep learning framework. Additionally, we propose a pose refinement procedure that applies physical constraints to hand movement without the need for deep learning or large scale PD dataset in the process.

Next, to quantitatively evaluate and score the FTT, 123 features are extracted based on the relative 3D distance and speed of the thumb and index fingertips, including variance, energy, Fourier spectrum coefficients, etc. Then, embedded feature selection is applied to select 31 features as the final input of SVM to yield FTT scores.

To verify the effectiveness of this proposed method, we constructed a new PD dataset by collecting 112 FTT video clips from 48 PD patients and 11 control subjects using a depth camera at Wuhan Union Hospital. Physicians specialized in PD and Neurology independently evaluated each video clip to give a severity score. In the end, our proposition achieves 81.2% classification accuracy on this dataset using 5-fold cross-validation protocol [21], which even surpasses the accuracy of individual clinicians in comparison. Meanwhile, experiments of the new spatial-temporal 3D HPE model on HANDS 2017 dataset [22] also demonstrate the effectiveness of the ST-A2J proposition.

Overall, the main contributions of this paper include:

- A novel research idea of conducting finger tapping test in PD patients using only 3D computer vision is established;
- An effective spatial-temporal 3D hand pose estimation approach is proposed;
- The first vision-based 3D hand dataset for PD patients with finger tapping test scores annotated by qualified clinicians according to MDS-UPDRS is constructed.

II. Related Work

A. Finger Tapping Test With Wearable Sensors

Wearable sensors have been used to capture PD patient's hand movement, including accelerometers [23], [24], gyroscopes [25], [26], magnetometers [10], [27], and Inertial Measurement Units (IMUs) [28]. Since these devices have to be attached to the patient's thumb or index finger, they may alter the patient's hand movement pattern, potentially weakening the accuracy and reliability of the test. Additionally, acceleration sensors often require the index finger to move parallel to the desktop and the collected data will be affected by gravitational artifacts [29]. Gyroscopes [26], [30] and accelerometers [23], [24] often omit the thumb's motion information. And IMUs employed in [28] only measure the angle between the index finger and the thumb, leaving out crucial information including the absolute maximum opening amplitude and speed of fingertips.

Moreover, wearable sensors have limited potential in clinical deployment since they require more complex calibration procedures than vision cameras [10], [31]. In contrast, vision-based methods which collect hand motion data in a contact-free fashion are cheaper, require less calibration, and can give a wider range of motion estimations.

B. Finger Tapping Test With Vision Sensors

Existing vision-based FTT approaches [11], [12], [14], [15], [32] can be categorized into 2D and 3D-based groups. They generally follow the idea of predicting hand joint positions first, and then extracting fingertip's motion features for PD severity level evaluation. The main drawback of 2D methods [14], [15], [32] is that only the 2D pixel-level hand joint positions can be acquired within the image plane. Consequently, the real-world 3D hand movement information cannot be fully revealed despite numerous normalization techniques suggested to alleviate its influence. Thus, a more suitable way to conduct FTT is by using 3D hand joint positions. However, existing 3D methods [14], [15], [32] still heavily rely on wearable devices (e.g., markers [15], [32] or gloves [14]) to facilitate 3D hand pose estimation, which

may still alter the patient's hand movement patterns. To address this, contact-free and effective 3D hand pose estimator for FTT is needed.

C. Finger Tapping Assessment With Smartphones

In parallel to the above clinical-oriented methods, self-administered PD evaluations using smartphones [32]–[36] are becoming increasingly popular. Among these methods, [35], [36] are two applications that utilize the finger tapping motion, which bear some resemblance to FTT from UPDRS Item 3.4. To be specific, [35] asks the participant to repeatedly tap the two sides of a rectangle on the touchscreen as fast as possible with an index finger, and correlates metrics such as correctness and movement distance with UPDRS scores to different PD patients from healthy controls. More recently, [36] requires the subject to place the smartphone on a horizontal surface and use two fingers from the same hand to touch two points on the touchscreen in an alternate fashion, and unsupervised clustering algorithm splits subjects into UPDRS severity groups.

Nonetheless, even though these methods leverage the finger tapping motion for more objective assessment, they only find indirect correlations between finger movement metrics collected by the touchscreen and UPDRS severity scores. And to our knowledge, no existing smartphone-related work has utilized the FTT test and rated the features as specified by UPDRS item 3.4. Hence, an FTT method that conducts the finger tapping and rates the finger's amplitude, velocity, and rhythm as instructed by UPDRS item 3.4 is called for.

D. 3D Hand Pose Estimation

According to input modality, existing 3D HPE approaches [17]–[19], [37], [38] can be categorized into depth map, 3D voxel, and point cloud based groups that rely on 2D CNN (Convolutional Neural Network) [39], 3D CNN [40], and point-set network [41] respectively. Among them, voxel-based approaches (e.g., V2V: Voxel-to-voxel Prediction Network [18]) suffer from large model size, which leads to training difficulty and poor running efficiency. Point cloud based methods (e.g., P2P, Point-to-point Regression Pointnet [42], and SO-HandNet, Self-Organizing Network for 3D Hand Pose Estimation [38]) are easily affected by point cloud sparsity as well as imaging noise. Also, the farthest-point-sampling process during feature learning is time consuming. On the other hand, depth map and 2D CNN approaches acquire a better trade-off between effectiveness and efficiency, where A2J (Anchor-to-Joint Regression Network) [17] previously proposed by our team is an outstanding example. By conducting local ensemble regression via preset dense anchors (i.e., local regressors), A2J achieves promising results with high generalization capability by employing richer global-local spatial context information. But for PD finger tapping test, A2J suffers from the jittering problem across consecutive frames since no temporal information is utilized.

To achieve satisfying FTT performance, 3D HPE should be spatially accurate and temporally consistent, (e.g., having consistent key-point velocity and acceleration across frames). But existing spatial-temporal 3D hand pose estimation approaches [20], [43] primarily focus on minimizing the traditional 3D mean distance error on one frame. The lack of consideration for temporal consistency often leads to the jittering of key-points on the temporal axis. Accordingly, we propose to extend A2J to spatial-temporal domain as Spatial-Temporal A2J (ST-A2J) to alleviate these problems. And since deep learning approaches are data-hungry whereas existing 3D PD hand data are limited, ST-A2J's result is refined without deep learning.

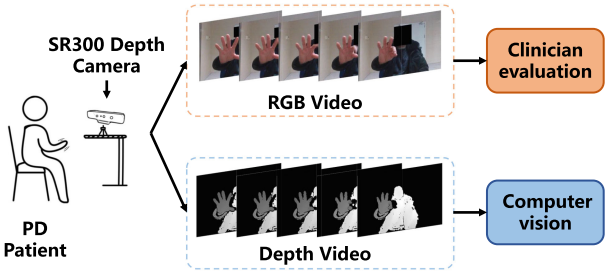


Fig. 2. Schematic diagram of data acquisition.

E. Kinematic Feature Extraction and Classification

Based on 2D or 3D hand pose estimation results, kinematic features of the hand (e.g., amplitude [14], [15], velocity [12], [15], and frequency [14], [15]) can be extracted to characterize finger tapping movement with feature-score correlation analysis [14]. And then supervised classifiers are used to generate the final rating scores [15]. Following the research paradigm, 3D hand kinematic features are extracted to acquire the severity rating score via SVM with max-margin mechanism, ensuring good generalization capability when the number of training samples is relatively small.

III. Clinical Finger Tapping Test Data Collection

A. System Overview

As shown in Fig. 2, Intel RealSense SR300 RGB-D camera [16] is used to simultaneously record RGB and depth videos of PD patient's FTT movement. SR300 is suitable for hand pose estimation due to its high precision [44], [45] compared to other depth cameras. The RGB videos are used for manual clinical evaluation which serves as severity annotation, and the depth videos are reserved as the input to our ST-A2J model. The resolution of the depth map is set to 640×480, and the frame rate is around 30 FPS.

During clinical test, SR300 camera is placed on a table and connected to our recording computer. Each patient is asked to sit in front of the camera, and certified physicians stand beside the table to give the patient instructions, ensuring that FTT procedures according to MDS-UPDRS [3] are correctly followed. Demonstrated in Fig. 3, our vision-based solution imposes no restriction on the patients' hand movement, as long as the hand appears in the camera's field of view.

B. Patients' Data and Annotation

Patients are recruited at the Department of Neurology, Wuhan Union Hospital, from July to October 2020. Participants are clinically diagnosed as idiopathic PD according to the 2015 Movement Disorder Society diagnostic criteria [4]. A total of

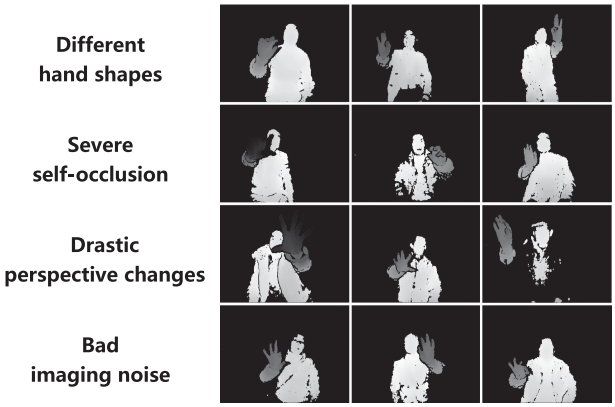


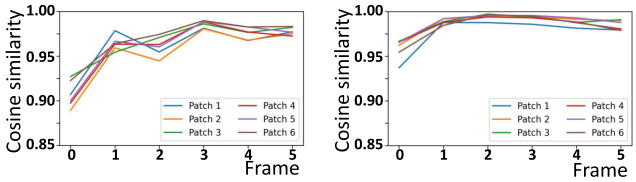
Fig. 3. Finger tapping sample frames from our dataset. Unconstrained hand movement presents great challenges.

TABLE I
INFORMATION OF SIX CLINICIANS AND SCORING RESULTS OF 90 VIDEOS OF PD PATIENT FINGER TAPPING TEST

Clinician	Title	Experience (years)	Scores				
			0	1	2	3	4
No. 1	Junior	2	5	31	19	25	10
No. 2	Junior	3	17	21	23	22	7
No. 3	Junior	5	10	22	20	29	9
No. 4	Medium	10	0	12	29	40	9
No. 5	Senior	17	0	15	30	37	8
No. 6	Senior	21	1	22	16	43	8
Final	–	–	0	26	19	38	7

48 Parkinson's Disease patients (21 males and 27 females) are included. They average to be 64.1 years old (SD 8.7), with the youngest being 47 and the oldest 82. The left and right hands are recorded separately for each patient, and 44 clips of patients' left hands and 46 clips of right hands are included in our dataset. In addition, 22 video clips of FTT videos from 11 healthy subjects are recorded as the control group, which are also all rated by clinicians and result in severity ratings of 0's. In the end, this new finger tapping dataset contains a total of 112 videos from 48 PD patients and 11 healthy controls. The data collection is approved by the Medical Ethics Committee of Tongji Medical College, Huazhong University of Science and Technology. All participants provided informed consent prior to participating in the study. The scale of this new dataset is larger than previous works [11], [12], [32], [46] and it is the first contact-free 3D hand dataset for finger tapping test.

Observer subjectivity can lead to FTT score variance between clinicians due to different experience levels, fatigue, and being influenced by successive PD evaluations of consecutive patients. A total of six certified physicians participates in the evaluation of patients using our recorded RGB clips. To reduce the impact of subjectivity in our dataset, every RGB video is distributed to all six clinicians for independent evaluations. All clinicians are medical doctors and have completed *MDS-UPDRS and UDysRS Training Programs* [47], qualified to conduct PD evaluations. Each clinician evaluates all 112 video clips (90 from patients, 22 from healthy controls) and gives FTT scores according to Item 3.4 of MDS-UPDRS. Table I shows the titles of six clinicians and the corresponding rating scores for the 90 video clips of PD patients. Each video clip's final score for the dataset is determined by majority voting, as shown in the last row of the table. Divergence can be observed for most patient's scoring results given by different clinicians, which is what this research aims to reduce.



(a) A2J (b) ST-A2J

Fig. 4. Comparison of prediction results from A2J and ST-A2J using the collected PD dataset. 3D hand poses are projected onto the image plane for visualization.

IV. SPATIAL-TEMPORAL 3D HAND POSE ESTIMATION

This section details the motivation, network structure, pose refinement, and hand tracking of Spatial-Temporal A2J.

A. A2J and Its Temporal Jittering Challenge

Anchor-to-Joint Regression Network (A2J) [17] is the state-of-the-art 3D hand pose estimation method based on a single depth image. Its key idea is to extract local features by densely setting anchor points, and predict the offset vectors between each anchor point and each hand joint. Then, these offset vectors are aggregated through ensemble learning to ensure generalization, resulting in a good balance between local and global information.

Although A2J can achieve good prediction accuracy and high running efficiency on a single depth frame, problems arise when applying it to predict 3D hand poses of PD patients. As shown in Fig. 4, the positions of index fingertip estimated by A2J are visually unstable and inconsistent across 6 consecutive depth images, even though the frames appear similar. This un-smoothness across consecutive estimations is caused by the low number of informative anchor points (weight greater than 0.02), and the drastic change in positional distribution of these anchor points between frames.

To quantify this un-smoothness, features of six patches around the index fingertip are extracted in the last intermediate layer of the anchor proposal branch. By using $\mathbf{X}_p^t \in \mathbb{R}^{256}$ to represent the feature of the p-th local patch in the t-th depth image, where $t = [0, 1, 2, 3, 4, 5]$ and $p = [0, 1, 2, 3, 4, 5]$, the index finger region in Fig. 4 are divided into six image patches in order of row priority.

Then the Cosine Similarity (CS) is calculated between the feature of patch $\mathbf{X}_p^t$ and its temporal mean $\overline{\mathbf{X}}_p$:

$$CS(\mathbf{X}_p^t, \overline{\mathbf{X}}_p) = \frac{\mathbf{X}_p^t \cdot \overline{\mathbf{X}}_p}{\|\mathbf{X}_p^t\| \, \|\overline{\mathbf{X}}_p\|}, \tag{1}$$

where $\overline{\mathbf{X}}_p$ is defined as:

$$\overline{\mathbf{X}}_p = \frac{1}{T} \sum_{t=1}^{T} \mathbf{X}_p^t. \tag{2}$$

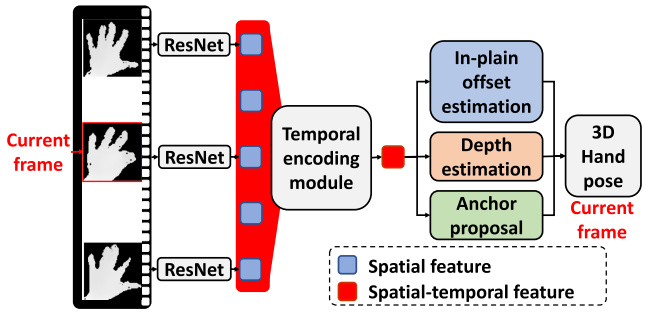


Fig. 5. The Cosine Similarity (CS) changes of the features that correspond to the 6 local patches in Fig. 5(a), between the 6 consecutive frames and the corresponding mean vector.

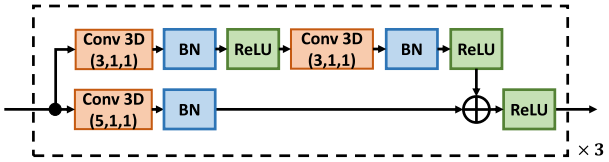


Fig. 6. The main technical pipeline of ST-A2J.

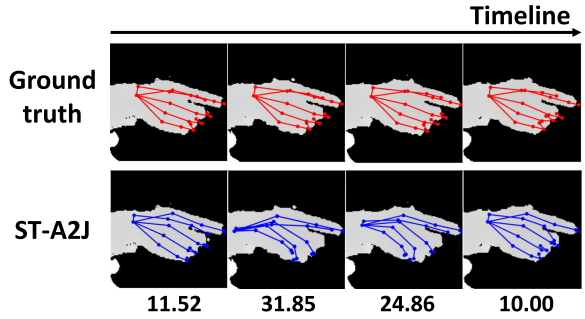


Fig. 7. Architecture of residual blocks in the temporal encoding module.

The CS change curve of the six local patches is shown in Fig. 5, where a closer-to-1 CS result represents a higher similarity of features. It can be observed from the CS change curve that the difference between the local image patches of the last four frames in Fig. 4 is not very obvious, whereas the CS between each feature $\mathbf{X}_p^t$ and the mean feature $\overline{\mathbf{X}}_p$ is relatively low in Fig. 5. This phenomenon reveals that the jittering of the hand joint is essentially caused by the inconsistency of features in the temporal dimension.

B. Spatial-Temporal A2J

As A2J's temporal jittering problem is caused by inconsistent spatial features across consecutive frames, the spatial features of these frames need to be temporally encoded to both retain the spatial information of the hand and achieve better consistency across frames. Hence, this work proposes the Spatial-Temporal A2J (ST-A2J), which extends A2J to the spatial-temporal domain through a temporal encoding module.

The main technical pipeline of our ST-A2J model is demonstrated in Fig. 6. Different from hand pose estimators using single frames [19], [37], [42], [50], the input of our ST-A2J is multiple consecutive depth hand images, where the middle frame is called the "current" frame. First, the pre-trained A2J backbone network is used to extract spatial features from each independent frame. Then, these features are stacked along the temporal dimension and served as the input of our temporal encoding module. Our temporal encoding module is composed of several 3D convolutional layers to capture the temporal context information of local spatial features. Compared to the spatial features extracted by A2J [17], the outputs of our temporal encoding module have the same feature dimension but are encoded with rich spatial-temporal context information of the hand when predicting the current frame. Then, the spatial-temporal features are sent to the in-plane offset estimation branch, depth offset estimation branch, and anchor proposal branch similar to A2J method [17]. Finally, the in-plane position and the depth value of each joint are obtained by the weighted sum of the in-plane offset and depth values of all anchor points.

The temporal encoding module is designed to eliminate the jittering of prediction results caused by inconsistent features. The convolution operation on the temporal dimension uses temporal context to smooth the features centered on the current frame and removes noise in the features, which is very helpful when the hand pose changes only slightly in consecutive frames. The temporal encoding module is composed of 3 residual blocks, which enables the ST-A2J network to extract temporal information in 13 consecutive depth images. Demonstrated in Fig. 7, each residual block is composed of several 3D convolutional, batch normalization (BN) and ReLU activation layers. Numbers in brackets represent the (depth, height, width) of the 3D convolution kernel.

Fig. 5 shows ST-A2J's cosine similarity between feature $\mathbf{X}_p^t$ of the p-th local patch in the t-th depth image extracted and the mean vector $\overline{\mathbf{X}}_p$ (calcutated by (9)). Compared with A2J's Fig. 5(a), ST-A2J extracts features with higher cosine similarity, meaning that they are more consistent in the temporal dimension.

Additionally, as shown in Fig. 4, the distribution of informative anchor points obtained by the ST-A2J is also more similar across consecutive frames, and the prediction results of 3D hand pose are more accurate and consistent. This enables more accurate and meaningful analysis of hand movements in PD patients.

Although ST-A2J produces rich and descriptive spatial-temporal information, empirically it can still fail to perform on a few challenging cases. Fig. 8 shows the prediction results of ST-A2J on four consecutive frames with severe occlusion in HANDS 2017 dataset [22]. Particularly, the average 3D distance error (listed below each picture) is still relatively high as it exceeds 20 mm. Our intuition is that ST-A2J is still a deep learning based approach so it suffers from being data-hungry, and existing hand datasets for training cannot possibly cover all hand poses. Thus, we propose to refine ST-A2J's prediction result without deep learning from the temporal perspective.

C. 3D Hand Pose Refinement

3D hand movement is continuous in spatial-temporal dimensions and has strong physical constraints, which can be used to

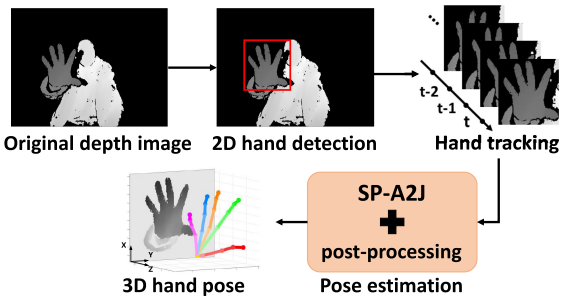


Fig. 8. ST-A2J's failure cases when severe self-occlusion happens, HANDS 2017 dataset. The numbers below each picture represent the average 3D distance error in millimeters.

effectively optimize the prediction results of ST-A2J and resolve the ambiguity caused by self-occlusion. These constraints can be summarized into two rules: 1) hand joints generally do not change drastically between consecutive frames; 2) bone lengths between hand joints remain consistent. In this work, these hand motion constraints are integrated by minimizing an objective function $E(X)$ to further refine ST-A2J's prediction. Specifically, $E(X)$ consists of temporal constraint $E_T(X)$, joint preference constraint $E_I(X)$ and bone length constraint $E_B(X)$, which can be formulated as:

$$E(X) = \lambda_1 E_T(X) + \lambda_2 E_I(X) + E_B(X). \tag{3}$$

Temporal constraint encourages smooth changes of 3D hand joint in adjacent frames and improves the anomaly detection results in individual frames:

$$E_T(X) = \frac{1}{T-1}\frac{1}{J}\sum_{t=2}^{T}\sum_{j=1}^{J}\|X_j^t - X_j^{t-1}\|, \tag{4}$$

where X is hand joints' 3D coordinates for optimization in all T frames. $X_j^t \in \mathbb{R}^3, j = (1, 2, \ldots, J), t = (1, 2, \ldots, T)$, is the 3D coordinates of the j-th joint point in the t-th frame. $J = 21$ represents hand joints number in this work.

Joint preference constraint ensures that the final result will stay close to the initial 3D hand pose $\widetilde{X}_j^t$, preventing it from being over-smoothed:

$$E_I(X) = \frac{1}{T}\frac{1}{J}\sum_{t}^{T}\sum_{j}^{J}\|X_j^t - \widetilde{X}_j^t\|, \tag{5}$$

where $\widetilde{X}_j^t$ is the prediction result of ST-A2J and we use it to initialize the optimization variable X.

Bone length constraint is composed of $E_{BC}(X)$ and $E_{BL}(X)$. It is defined as:

$$E_B(X) = \lambda_3 E_{BC}(X) + \lambda_4 E_{BL}(X), \tag{6}$$

where E_{BC} is to keep the length of the hand bone constant in a video, and $E_{BL}(X)$ is to restrict the length of each hand bone within a reasonable range. E_{BC} can be expressed as:

$$E_{BC}(X) = \frac{1}{T}\sum_{t}^{T}|B^t - \overline{B}|, \tag{7}$$

where $B^t \in \mathbb{R}^{24}$ is the bone length vector calculated using the 3D hand joints of the t-th frame. $\overline{B}$ presents the mean value of

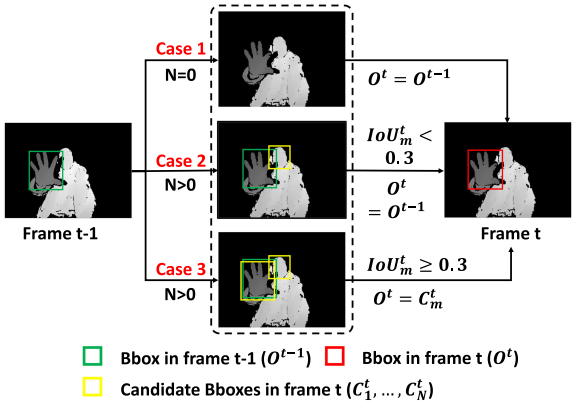


Fig. 9. Hand detection and tracking pipeline: from original depth image to 3D hand pose.

B^t over T frames. To constrain the bone length vector of each frame within a reasonable range, $E_{BL}(X)$ is defined as:

$$E_{BL}(X) = \frac{\lambda_5}{T}\sum_{t}^{T} max(\frac{B^t}{\|B^t\|} - B_{max}, 0) + max(B_{min} - \frac{B^t}{\|B^t\|}, 0), \tag{8}$$

where B_{max}, B_{min} are the upper and lower bounds of each normalized hand bone calculated in advance according to different hand shape on the large-scale 3D hand pose dataset HANDS 2017 [22]. The $\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5$ in the above formula are hyperparameters to balance the magnitude of each error term. Finally, the objective function $E(X)$ is optimized through the L-BFGS-B algorithm [51] in python to obtain the 3D coordinates of the hand joints in T frames.

D. Hand Detection and Tracking

To estimate 3D hand pose with ST-A2J in depth video, the local hand region needs to be extracted in advance as shown in Fig. 9 to eliminate background. Accordingly, YOLO V3 [52] is used to detect the hand in each depth frame due to its satisfying performance and efficiency towards object detection. Bounding box (Bbox) annotations of hand samples inferred by joints' positions from HANDS 2017 [22] dataset are used to train YOLO V3. To account for potential failure cases, hand tracking is also conducted to correct wrong detection results as demonstrated in Fig. 10. Particularly, The Bbox of frame $t-1$ is denoted by O^{t-1}, and N candidate Bboxes of frame t is denoted by $(C_1^t, \ldots, C_N^t)$. When $N = 0$ (case 1 in Fig. 10), since the plane position of the hand is close between two adjacent frames, O^{t-1} is directly adopted as the detection result of frame t. When $N > 0$, the Intersection over Union (IoU) between each candidate Bbox C_n^t and the detection result in $t-1$ frame O^{t-1} are calculated by the following formula:

$$IoU_n^t = \frac{C_n^t \cap O^{t-1}}{C_n^t \cup O^{t-1}}, \tag{9}$$

where $n \in [1, N]$; and $IoU_n^t \in [0, 1]$ measures the overlap between two Bboxes. Then the candidate Bbox C_m^t with the largest

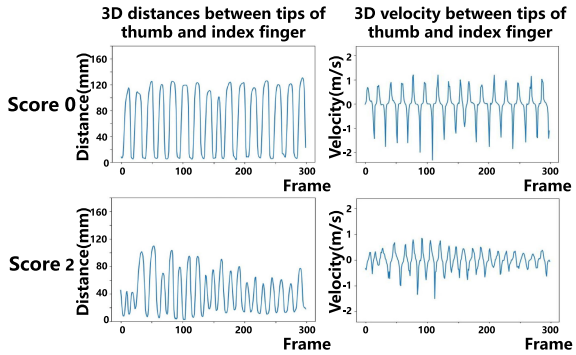


Fig. 10. Hand tracking for correcting false hand detection results by YOLO V3.

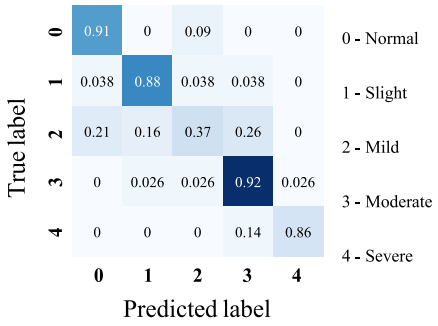


Fig. 11. Comparison of 3D distance and velocity between individual samples with score 0 and score 2.

IoU_m^t is derived. Next, the final detection result of the frame t is determined by:

$$O^t = \begin{cases} C_m^t, & \text{for} \quad IoU_m^t > 0.3, \\ O^{t-1}, & \text{otherwise.} \end{cases} \tag{10}$$

If $IoU_m^t \leq 0.3$ (case 2 in Fig. 10), the overlap between O^{t-1} and C_m^t is relatively small, and C_m^t tends to be the false detection. Thus, O^{t-1} is still regarded as the hand detection result of frame t. And if $IoU_m^t > 0.3$ (case 3 in Fig. 10), C_m^t is selected as the detection result of frame t.

V. Finger Tapping Assessment

Using accurate 3D hand poses produced by ST-A2J, the assessment score is calculated in three steps: *feature extraction*, *feature selection*, and *classification*.

A. Feature Extraction

To quantitatively analyze FTT movement, the distance between tips of the thumb and index finger and the corresponding velocity are calculated as the basic signals similar to some earlier works [14], [15], [32]. However, our method improves from previous approaches by extracting a more robust 3D distance with real physical meaning without using smoothing filters or normalization operations to process the distance and velocity signals, successfully preserving more motion details. After employing the proposed spatial-temporal 3D HPE method, it is observed that the amplitude of tapping with score 2 tends to decrease and the overall tapping velocity is relatively slow, whereas the sample with score 0 has a larger tapping amplitude, better consistency, and faster movement as demonstrated in Fig. 11.

According to MDS-UPDRS [3], the final score of finger tapping test depends on a variety of kinematics characteristics, such as amplitude, velocity and rhythm. To transform these kinematic characteristics into a quantitative representations, tsfresh [53] is used to calculate a variety of time series features based on 3D distance and velocity signals. Tsfresh [53] is a fast and standard machine learning library for time series feature extraction and has been widely used. Table II details the names, parameters, and corresponding dimensions of each feature we calculated using this library. In the end, 124-dimensional features are extracted for each video clip to describe the finger tapping movement.

TABLE II
Kinematic Feature Extraction for Finger Tapping Test, the Specific Calculation Formula Refers to [53]

Signal	Feature name	Parameter	Dimension
Distance	variance	None	1
	mean_abs_change	None	1
	autocorrelation	lag(1,2,· · · ,9)	9
	fft_coefficient	abs[0,0.1,· · · ,10)	100
	linear_trend	intercept, stderr	2
Velocity	variance	None	1
	abs_energy	None	1
	quantile	[0.1,· · · ,0.4,0.6,· · · ,0.9]	8
	linear_trend	stderr	1

B. Feature Selection

To eliminate redundant features and improve the accuracy of the classification algorithm, feature-selection is implemented based on the 124-dimensional features extracted. To verify the effectiveness of the filtered features, statistical analysis is performed on each dimension feature. Specifically, Spearman correlation coefficient between each dimensional feature and the assessment score are calculated while taking into consideration the sequentiality between the 0 to 4 scores, similar to other methods [14], [46].

C. Classification

To generate the final test score from 0 to 4 while taking into account the relatively small sample size, we experimented with several supervised classification algorithms in machine learning. These algorithms include k-nearest neighbors (KNN), random forests (RF), eXtreme Gradient Boosting (XGBoost), support vector machine with linear kernel (SVM-L) and Radial Basis Function kernel (SVM-R). We choose these algorithms for their ability to correspond to medical evaluation criteria and their visualization potentials, as well as considerations for data size and computation limitations in real-world applications. To quantitatively evaluate the performance of different classifiers, a five-fold cross-validation method [21] is adopted to obtain the final classification accuracy.

TABLE III
CLASSIFICATION ACCURACY COMPARISON OF SIX PARTICIPATED CLINICIANS AND OUR PROPOSED METHOD ON OUR DATA

Clinician	No.1	No.2	No.3	No.4	No.5	No.6	Ours
Title	Junior	Junior	Junior	Medium	Senior	Senior	-
Experience	2	3	5	10	17	21	-
Accuracy	0.589	0.571	0.687	0.687	0.696	0.786	**0.812**

TABLE IV
COMPARISON OF CLASSIFICATION ACCURACY ACHIEVED BY DIFFERENT FEATURE SELECTION METHODS AND DIFFERENT CLASSIFIERS

Selection Method	KNN	RF	XGBoost	SVM-L	SVM-R
None	0.661	0.713	0.731	0.686	0.714
Filter	0.750	0.722	0.739	0.732	0.714
SFS	0.731	0.758	0.703	0.750	0.794
RFE	0.642	0.705	0.695	0.668	0.705
Embeded	0.768	0.740	0.749	0.768	**0.812**

The best results are in bold.

VI. IMPLEMENTATION DETAILS

• ST-A2J is implemented using PyTorch with Adam optimizer, with the same network parameters as the original A2J for fair comparison. The learning rate is set to 0.00035 with a weight decay of 0.0001, and ST-A2J is trained for 27 epochs as the learning rate is set to decay by 0.2 every 7 epochs;

• Annotated data from HANDS 2017 dataset [22] is used to train ST-A2J. The hand image is cropped by Bbox and resized to 176×176, and the mean of all hand point clouds is used as the center point. The mean and variance of the point cloud after subtracting the depth value of the center point are used to normalize hand images;

• To optimize $E(X)$, λ_5 is set to 100, and $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ are determined by ablation experiments. L-BFGS-B algorithm [51] is employed to optimize the 3D hand joints X in $T = 100$ frames and the maximum number of iterations 100,000, resulting in $T \times J \times 3$ optimized parameters;

• Each of the FTT video clips has 300 frames, lasting about 10 seconds. 4 feature selection methods are implemented with Scikit-learn [54], and the optimal parameters of each classifier are determined by the GridSearchCV function.

VII. EXPERIMENTS

A. Effectiveness of Different Scoring Methods

The accuracy of each individual clinician's PD evaluation of 112 video clips is calculated and compared to that of the proposed ST-A2J. As shown in Table III, embedded feature selection and SVM-R classifier achieve the highest classification accuracy after using ST-A2J. It is observed:

• The proposed ST-A2J pipeline achieves a classification accuracy of 81.2%, higher than that of any individual clinician. ST-A2J's accuracy is 2.6% higher than a senior clinician with 21 years of clinical experience, and 22.3% higher than a junior clinician with 2 years. This result validates ST-A2J's high accuracy and consistency over the given PD samples;

• The variance in clinician's accuracy validates that, understandably, clinicians with much fewer years of experience may have higher inconsistency. This reveals the need for automatic and consistent FTT evaluations, and demonstrates the importance of the proposed ST-A2J and vision-based FTT.

B. Analysis of Different Classification Methods

As presented in Table IV, ST-A2J's final classification accuracy is greatly influenced by choosing different feature selection methods (see Sec. V-B) and classifiers (see Sec. V-C). It is observed that:

• Without feature selection, the classification accuracy is relatively low using the original 124-dimensional features, with the highest accuracy obtained with XGBoost classifier [55] at 73.1%. This verifies the necessity of feature selection;

• The combination of embedded feature selection and SVM classifier with the RBF kernel (SVM-R) achieves the highest classification accuracy of 81.2%, thanks to the high-dimensional feature mapping capability of the RBF kernel function and the max-margin mechanism of the SVM.

Fig. 12 shows the confusion matrix corresponding to the method with 81.2% accuracy. One can observe that class 0 and 3 have high classification accuracy (over 90%), followed by class 1 and 4 (above 85%). However, more mis-classifications occur within class 2 with an accuracy of only 37%. Its cause is the highly similar finger tapping movement of adjacent severity levels, a main challenge of accurately scoring FTT.

Additionally, we also employed leave-one-out strategy to check for the variance of correlation. For each of the 112 samples, we train the classifier on all other 111 samples and test whether the classifier can correctly predict the unseen sample. Our classifier achieves 76.79% overall classification accuracy, with a variance of 0.8444.

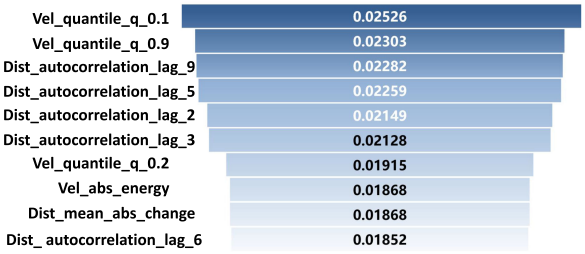


Fig. 12. Confusion matrix corresponding to the proposed model of 81.2% accuracy. Classes are severity levels from 0 (normal) to 4 (severe) according to MDS-UPDRS.

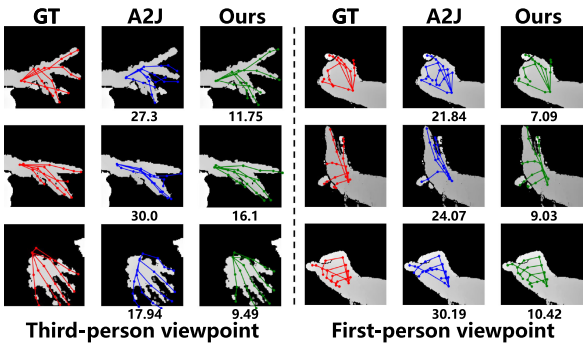


Fig. 13. Top ten features and their corresponding importance after embedded feature selection.

TABLE V
COMPARISON OF CLASSIFICATION ACCURACY WITH DIFFERENT 3D HAND POSE ESTIMATORS AND DIFFERENT DIMENSIONS OF HAND JOINT POSITIONS

Joint type	A2J	ST-A2J	Refined ST-A2J
2D	0.687	0.741	0.749
3D	0.714	0.767	**0.812**

TABLE VI
PERFORMANCE COMPARISON OF OUR ST-A2J MODEL AND SEVERAL OTHER HAND POSE ESTIMATORS

Method	Dist Err	Vel Err	Acc Err	Bone Err	KPA
Hand Pointnet [57]	20.653	11.129	18.596	110.125	4.9%
P2P [43]	16.841	10.937	18.500	111.535	10.4%
Global Reg [56]	9.921	3.253	5.074	66.420	37.7%
DeepPrior++ [59]	9.528	3.698	5.878	62.743	39.1%
ST-DeepPrior++	8.898	2.265	3.270	59.235	49.1%
A2J [17]	9.132	3.206	5.042	64.972	55.1%
ST-A2J	8.405	2.176	3.130	61.009	70.8%
Refined A2J	8.901	2.253	3.243	44.888	57.8%
Refined ST-A2J	**8.266**	**1.800**	**2.412**	**43.958**	**76.9%**

C. EFFECTIVENESS OF HAND POSE ESTIMATION IN 3D

The higher effectiveness of 3D poses over 2D coordinates is highlighted by converting the 3D joint position estimates from 3 state-of-the-art methods (A2J, ST-A2J, and refined ST-A2J) into 2D image coordinates to extract the corresponding 2D kinematic features, and demonstrating their classification results in Table V, after applying embedded feature selection and SVM classifier. It is observed that:

- Motion features extracted using 3D hand joints achieve better classification accuracy than those from 2D hand joints, indicating 3D hand poses' higher robustness and 3D kinematic features' stronger capacity to distinguish different PD severity levels, verifying the importance of 3D hand pose;
- ST-A2J method and the refinement step are both helpful to improve the classification accuracy. It shows that the spatial-temporal information of hand movement not only helps to improve the prediction accuracy of 3D hand joints (verified in Sec. VII-E) but also contributes to the classification of finger tapping test of PD patients.

D. DIRECT COMPARISONS OF 3D HAND POSE ESTIMATORS

Model performances are compared between the original A2J model [17], the proposed ST-A2J model, and the Refined A2J as well as Refined ST-A2J model which follow optimization approach proposed in Sec. IV-C. Additionally, state-of-the-art regression-based hand estimator DeepPrior++ which uses a single depth frame as input is re-implemented and compared, along with its extension by us to ST-DeepPrior++ that takes consecutive depth frames like ST-A2J. A Global Regression based method [56] and two Point Cloud based methods [42], [57] are also compared. Global Regression is applied to encode depth image with pre-trained ResNet-50 [48], and then 3D joint positions are regressed through one fully-connected layers. On the other hand, point cloud methods are used to retrain Hand Pointnet [57] and P2P [42] on HANDS 2017 dataset [22]. Experimentation results are reported in Table VI.

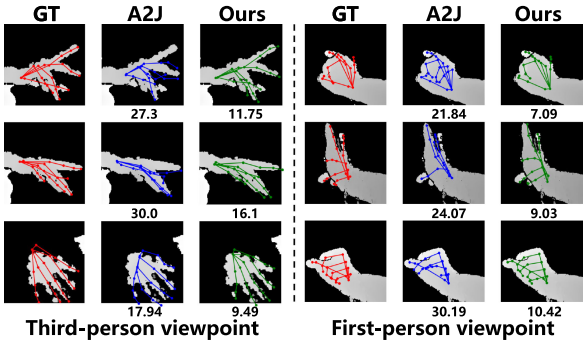


Fig. 14. Visualized hand pose comparisons between Ground truth (GT), A2J, and the proposed ST-A2J on HANDS 2017 dataset [22]. Numbers are average 3D distance error in millimeters.

To conduct a comprehensive and quantitative evaluation of the 3D hand poses, we propose to use Velocity Error (**Vel-Err**), Acceleration Error (**Acc-Err**), Bone Consistency Error (**Bone-err**), Average 3D Distance Error (**Dist-Err**), and Keypoint Pixel Accuracy (**KPA**). The Vel-Err and Acc-Err measure the difference between the predicted and ground-truth 3D velocity and acceleration for each hand point, calculated using the finite-difference between individual frames while ignoring the sampling interval Δt. Bone-err calculation is described in (7). Dist-Err measures the 3D Euclidean distance between the predicted and the ground truth 3D joint positions in millimeters (mm). And KPA measures the percentage of keypoint pixel coordinates that are correctly estimated in the 3D depth-image space for all 21 hand keypoints. These error and accuracy terms need to be considered jointly when estimating the 3D hand pose of PD patients, because lower Dist-Err and higher KPA indicate better overall correctness in 3D HPE, whereas the other three errors represent whether the temporal movement is smooth and natural.

All quantitative experiments are conducted using the HANDS 2017 dataset [22]. HANDS 2017 dataset contains 99 segments from BigHand2.2M [44] and First-Person Hand Action [45], including a variety of hand poses in the first and third-person viewpoints. Each segment has 2700-3300 consecutive depth frames with 3D coordinates hand joints provided as the ground truth, collected using magnetic sensors attached to fingertips and wrists. Our training set contains 262,092 continuous depth images from 1-89 segments in the HANDS 2017 dataset and the test set contains 30,728 images from the 90-99 segments.

It can be observed from visualization results in Fig. 14 that our ST-A2J obtains more accurate prediction results than A2J even under severe self-occlusion. And from the experiment results shown in Table VI, it can be summarized that:

- Compared with methods using only a single frame of spatial information, the proposed ST-A2J achieves the lowest errors and highest accuracy in all five evaluation criteria. Compared to the best results obtained by A2J method, ST-A2J achieves 7.9%, 32.1%, 37.9%, 6.0%, and 28.4% improvement on five evaluation metrics respectively, verifying the effectiveness of the proposed temporal encoding module;
- Both taking multiple frames (13 frames) as input, ST-A2J and ST-DeepPrior++'s strong performance increases over the original A2J and DeepPrior++ models on HPE

demonstrate the power of our proposed Spatial-Temporal pipeline. And ST-A2J's advantage in prediction performance over ST-DeepPrior++ shows the necessity of using A2J backbone.

- Based on the prediction results of both A2J and ST-A2J, the 3D hand pose refinement approach can further optimize the 3D position of the hand joint and achieve lower prediction errors. Refined ST-A2J achieves the best prediction performance among all models compared, demonstrating the effectiveness of the proposed hand pose refinement procedure;

VIII. Discussion

This paper proposes a vision-based automatic PD finger tapping test using spatial-temporal 3D hand pose estimation. The novelty of this work is the utilization of depth camera and consecutive depth frames for PD evaluation, and the spatial-temporal HPE model that accurately estimates the hand keypoints in 3D to alleviate the jittering of hand poses in the temporal axis. The proposed FTT evaluation pipeline closely follows UPDRS item 3.4 Finger Tapping [3], collecting FTT movement features and evaluating scores all according to UPDRS instructions. The proposed method captures rich spatial information compared to RGB cameras or wearable sensors and allows clinicians to accurately model the hand and evaluate the finger tapping test, giving quantifiable hand movement readings that can be recorded for later comparisons.

A daily clinical FTT evaluation after employing the proposed ST-A2J method could proceed as follows: a patient will sit in front of a depth camera and read instructions (or watch a demonstration video) on how to perform FTT. Next, the patient will perform FTT, *"tapping the index finger on the thumb 10 times as quickly AND as big as possible"* as instructed by UPDRS item 3.4 [3], while being recorded by the depth camera of the proposed St-A2J system without requiring clinician's supervision. Subsequently, ST-A2J estimates the patient's robust 3D hand pose from the depth video, and extracts movement features such as amplitude, velocity, and rhythm as specified in UPDRS item 3.4. And as a result, the SVM classifier gives the patient's FTT score from 0 to 4 directly corresponding to UPDRS, while the quantified movement features can be recorded for any future usage.

The advantage of the proposed ST-A2J method over wearable sensors [9], including accelerometers [59], [60], and magnetometer [34], [61], [62], is that ST-A2J models the spatial-temporal movement of all hand keypoints while making no contact. This allows the model to extract and evaluate key features of FTT corresponding to UPDRS or any other evaluation criteria, without the possibility that fine motor movement from patient is altered by wearable equipment. Clinically, rather than carefully calibrating and operating expensive wearable sensors, the proposed system only requires a commercial depth camera and a computer to generate accurate hand poses to evaluate the finger tapping test.

As mentioned earlier, several self-administered methods exist that utilize smartphones to perform PD evaluations [32]–[36]. These methods usually collect and analyze data of the user's interaction with the smartphone, including the movement of finger on the screen, memory test, or speech analysis, and find correlations between the interaction patterns and PD severity levels. A disadvantage of the proposed work to these methods is that our ST-A2J pipeline still requires a dedicated depth camera. However, whereas many of the self-administered test show promising correlation between smartphone interaction pattern and PD, the clinical importance of our work is that it directly follows the UPDRS testing procedure for FTT. As a result, our model gives quantifiable metrics such as opening amplitude and speed closely corresponding to UPDRS, and these results can be further analyzed or stored for later comparisons.

Within these methods that use smartphone applications, [35] and [36] both attempt to record the finger tapping movement to assess PD, similar to the proposed method in this work. In [35], the correctness, movement distance, and "inter-tap dwelling time" of the index fingertip on the touchscreen is recorded, and the correlation between these metrics and UPDRS scores are calculated to determine PD severity. And in [36], unsupervised learning is leveraged to cluster samples according to UPDRS severity, using the positions and frequencies of the taps on the touchscreen as well as information from the smartphone's accelerometer. Whereas these methods use the smartphone to indirectly infer the finger tapping movement, the proposed ST-A2J method directly calculates the 3D hand pose during FTT, giving straightforward pose estimations for any desired features. As a result, these smartphone-based methods can only find correlations with UPDRS severity, or cluster subjects according to their movement patterns and UPDRS scores. In comparison, the proposed ST-A2J can directly calculate features including amplitude, velocity, and rhythm which are explicitly specified in UPDRS Item 3.4 for the clinical evaluation of PD.

Other automatic FTT evaluation methods are usually designed only for FTT task. For example, [13] only detects the relative orientation and position of the index finger and the thumb. In contrast, since ST-A2J is a robust 3D hand pose estimation model that calculates keypoint coordinates of the full hand, it can be generalized to many other UPDRS items [3] and tasks. For example, after some appropriate modification and additional model training, this method can also extend to Item 3.5 Hand Movement, Item 3.6 Pronation/Supination, and Item 3.15 Postural tremor of hands from UPDRS, or even full body pose estimations.

A practical limitation of the proposed work is that its collected dataset has imbalanced labels, as less severe cases are more frequent and vice versa. To be exact, of our 90 video clips from PD patients, 26, 19, 38, and 7 samples are rated from 1 (slight) to 4 (severe), respectively. When trained on an imbalanced dataset like this, a machine learning model has the potential to bias towards giving lower severity scores due to the higher frequency of these training samples. And it should be highlighted that on the contrary, physicians are trained to rate the full range of scores on items such as the MDS-UPDRS, including the most severe ones. Many reasons could lead to this label imbalance, and in our clinical observation one of the causes is that the most severe PD patients have impaired motor functions and may have difficulties commuting to the hospital. Nonetheless, to account for this label imbalance and validate the proposed method, five-fold cross-validation is performed before reporting model accuracy in Table III. Additionally, leave-one-out strategy, an extreme case of k-fold cross-validation, is further conducted and results in 76.79% mean accuracy.

IX. Conclusion

In conclusion, a novel 3D vision-based finger tapping test method for PD patients in a contact-free manner is proposed in this paper according to MDS-UPDRS. Without any wearable devices or markers, this method achieves a test accuracy of 81.2% on a newly established 3D PD hand dataset of 59 subjects collected under practical clinical condition, verifying the feasibility of the proposition. To achieve this goal, an effective spatial-temporal 3D hand pose estimation approach is proposed, and its effectiveness is demonstrated using the large-scale HANDS 2017 dataset. Currently, only single-center study is conducted. In the future, multi-center studies including FTT evaluation using the proposed method without in-person instructions or supervision is planned to further verify the applicability of our proposition. And further extension of the model to other PD evaluation items, as well as correlating with bradykinesia and tremor information to enhance the evaluation and management of PD, is also scheduled.

Acknowledgment

The authors would like to thank clinicians Xiaoman Yang, Chi Cheng, Xiaomei Yang, and Weijian Peng from the Department of Neurology, Union Hospital, Tongji Medical College, Huazhong University of Science and Technology for their professional suggestion and assistance in this work.Thanks to Boshen Zhang and Wenzheng Zeng from Key Laboratory of Image Processing and Intelligent Control, Ministry of Education; School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, for their effort in the construction of our dataset.

Thanks to Mingyang Zhang and Changlong Jiang from Key Laboratory of Image Processing and Intelligent Control, Ministry of Education; School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, for their timely assistance in our additional experiments.

References

[1] Parkinson Association of the Carolinas, "Statistics on Parkinson's disease," Accessed: Feb. 22, 2021. [Online]. Available: https://jnnp.bmj.com/content/79/4/368

[2] J. Jankovic, "Parkinson's disease: Clinical features and diagnosis," *J. Neurol.*, vol. 79, pp. 368–376, 2008.

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

[4] R. B. Postuma *et al.*, "MDS clinical diagnostic criteria for Parkinson's disease," *Movement Disord.*, vol. 30, no. 12, pp. 1591–1601, 2015.

[5] H. Ling, L. A. Massey, A. J. Lees, P. Brown, and B. L. Day, "Hypokinesia without decrement distinguishes progressive supranuclear palsy from Parkinson's disease," *Brain*, vol. 135, no. 4, pp. 1141–1153, Apr. 2012.

[6] C. G. Goetz and G. T. Stebbins, "Assuring interrater reliability for the UPDRS motor section: Utility of the UPDRS teaching tape," *Movement Disord.*, vol. 19, no. 12, pp. 1453–1456, Dec. 2004.

[7] L. Henderson *et al.*, "Scales for rating motor impairment in Parkinson's disease: Studies of reliability and convergent validity," *J. Neurol. Neurosurgery Psychiatry*, vol. 54, no. 1, pp. 18–24, 1991.

[8] J. A. Stamford, P. N. Schmidt, and K. E. Friedl, "What engineering technology could do for quality of life in Parkinson's disease: A review of current needs and opportunities," *IEEE J. Biomed. Health Informat.*, vol. 19, no. 6, pp. 1862–1872, Nov. 2015.

[9] C. F. Pasluosta, H. Gassner, J. Winkler, J. Klucken, and B. M. Eskofier, "An emerging era in the management of Parkinson's disease: Wearable technologies and the Internet of Things," *IEEE J. Biomed. Health Informat.*, vol. 19, no. 6, pp. 1873–1881, Nov. 2015.

[10] K. Shima, T. Tsuji, A. Kandori, M. Yokoe, and S. Sakoda, "Quantitative evaluation of human finger tapping movements through magnetic measurements," *IEEE/ASME Trans. Mechatronics*, vol. 24, no. 1, pp. 186–196, Feb. 2019.

[11] D. Buongiorno *et al.*, "A low-cost vision system based on the analysis of motor features for recognition and severity rating of Parkinson's disease," *BMC Med. Informat. Decis. Mak.*, vol. 19, no. 9, pp. 1–13, 2019.

[12] R. Krupicka *et al.*, "Bradykan: A motion capture system for objectification of hand motor tests in Parkinson disease," in *Proc. E-Health Bioeng. Conf.*, 2017, pp. 446–449.

[13] T. Khan, D. Nyholm, J. Westin, and M. Dougherty, "A computer vision framework for finger-tapping evaluation in Parkinson's disease," *Artif. Intell. Med.*, vol. 60, no. 1, pp. 27–40, Jan. 2014.

[14] S. Williams, Z. Zhao, A. Hafeez, D. C. Wong, and J. E. Alty, "The discerning eye of computer vision: Can it measure Parkinson's finger tap bradykinesia?," *J. Neurological Sci.*, vol. 416, 2020, Art. no. 117003.

[15] 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.

[16] *Intel RealSense Camera SR300*. Accessed: Feb. 22, 2021. [Online]. Available: https://ark.intel.com/content/www/us/en/ark/products/92329/intel-realsense-camera-sr300.html

[17] F. Xiong *et al.*, "A2J: Anchor-to-joint regression network for 3D articulated pose estimation from a single depth image," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 793–802.

[18] G. Moon, J. Yong Chang, and K. Mu Lee, "V2V-PoseNet: Voxel-to-voxel prediction network for accurate 3D hand and human pose estimation from a single depth map," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 5079–5088.

[19] L. Ge, H. Liang, J. Yuan, and D. Thalmann, "Robust 3D hand pose estimation from single depth images using multi-view CNNs," *IEEE Trans. Image Process.*, vol. 27, no. 9, pp. 4422–4436, Sep. 2018.

[20] Y. Cai *et al.*, "Exploiting spatial-temporal relationships for 3 d pose estimation via graph convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 2272–2281.

[21] R. Kohavi *et al.*, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Proc. Int. Joint Conf. Artif. Intell.*, vol. 14, 1995, pp. 1137–1145.

[22] S. Yuan, Q. Ye, G. Garcia-Hernando, and T. K. Kim, "The 2017 hands in the million challenge on 3D hand pose estimation," 2017, *arXiv:1707.02237*.

[23] J. Stamatakis *et al.*, "Finger tapping clinimetric score prediction in Parkinson's disease using low-cost accelerometers," *Comput. Intell. Neurosci.*, vol. 2013, pp. 1–13, 2013, doi: 10.1155/2013/717853.

[24] P. Kang, J. Li, B. Fan, S. Jiang, and P. B. Shull, "Wrist-worn hand gesture recognition while walking via transfer learning," *IEEE J. Biomed. Health Informat.*, vol. 26, no. 3, pp. 952–961, Mar. 2022.

[25] A. Salarian *et al.*, "Quantification of tremor and bradykinesia in Parkinson's disease using a novel ambulatory monitoring system," *IEEE Trans. Biomed. Eng.*, vol. 54, no. 2, pp. 313–322, Feb. 2007.

[26] J. W. Kim *et al.*, "Quantification of bradykinesia during clinical finger taps using a gyrosensor in patients with Parkinson's disease," *Med. Biol. Eng. Comput.*, vol. 49, no. 3, pp. 365–371, 2011.

[27] Y. Sano *et al.*, "Reliability of finger tapping test used in diagnosis of movement disorders," in *Proc. Int. Conf. Bioinf. Biomed. Eng.*, 2011, pp. 1–4.

[28] M. Djurić-Jovičić, N. Jovičić, A. Roby-Brami, M. Popović, V. Kostić, and A. Djordjević, "Quantification of finger-tapping angle based on wearable sensors," *Sensors*, vol. 17, no. 2, p. 203, Jan. 2017.

[29] R. J. Elble, "Gravitational artifact in accelerometric measurements of tremor," *Clin. Neuriophysiol.*, vol. 116, no. 7, pp. 1638–1643, 2005.

[30] M. Djurić-Jovičić *et al.*, "Implementation of continuous wavelet transformation in repetitive finger tapping analysis for patients with PD," in *Proc. Telecommun. Forum Telfor*, 2014, pp. 541–544.

[31] J. Li *et al.*, "Three-dimensional pattern features in finger tapping test for patients with Parkinson's disease," in *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2020, pp. 3676–3679.

[32] A. Lauraitis, R. Maskeliūnas, R. Damaševičius, and T. Krilavičius, "A mobile application for smart computer-aided self-administered testing of cognition, speech, and motor impairment," *Sensors (Basel)*, vol. 20 no. 11, Jun. 2020, Art. no. 3236.

[33] A. Zhan *et al.*, "Using smartphones and machine learning to quantify Parkinson disease severity: The mobile Parkinson disease score," *JAMA Neurol.*, vol. 75, no. 7, pp. 876–880, 2018.
[34] L. Omberg *et al.*, "Remote smartphone monitoring of Parkinson's disease and individual response to therapy," *Nature Biotechnol.*, Aug. 2021.
[35] C. Y. Lee, S. J. Kang, S.-K. Hong, H.-I. Ma, U. Lee, and Y. J. Kim, "A validation study of a smartphone-based finger tapping application for quantitative assessment of bradykinesia in Parkinson's disease," *PLoS One*, vol. 11, no. 7, Jul. 2016, Art. no. e0158852.
[36] D. Surangsrirat, Panyawut Sri-iesaranusorn, A. Chaiyaroj, P. vateekul, and R. Bhidayasiri, "Parkinson's disease severity clustering based on tapping activity on mobile device," *Sci. Rep.*, vol. 12, no. 1, Feb. 2022, Art. no. 3142.
[37] X. Chen, G. Wang, H. Guo, and C. Zhang, "Pose guided structured region ensemble network for cascaded hand pose estimation," *Neurocomputing*, vol. 395, pp. 138–149, 2020.
[38] Y. Chen, Z. Tu, L. Ge, D. Zhang, R. Chen, and J. Yuan, "So-HandNet: Self-organizing network for 3D hand pose estimation with semi-supervised learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 6960–6969.
[39] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.
[40] S. Ji, W. Xu, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 221–231, Jan. 2013.
[41] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "PointNet: Deep hierarchical feature learning on point sets in a metric space," 2017, *arXiv:1706.02413*.
[42] L. Ge, Z. Ren, and J. Yuan, "Point-to-point regression PointNet for 3D hand pose estimation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 475–491.
[43] Y. Wu *et al.*, "Context-aware deep spatiotemporal network for hand pose estimation from depth images," *IEEE Trans. Cybern.*, vol. 50, no. 2, pp. 787–797, Feb. 2020.
[44] S. Yuan, Q. Ye, B. Stenger, S. Jain, and T. K. Kim, "Bighand2.2M benchmark: Hand pose dataset and state of the art analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4866–4874.
[45] G. Garcia-Hernando, S. Yuan, S. Baek, and T. K. Kim, "First-person hand action benchmark with RGB-D videos and 3D hand pose annotations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 409–419.
[46] J. H. Shin *et al.*, "Objective measurement of limb bradykinesia using a marker-less tracking algorithm with 2D-video in PD patients," *Parkinsonism Related Disord.*, vol. 81, pp. 129–135, 2020.
[47] *MDS-UPDRS and UDYSRS training programs*. [Online]. Available: https://www.movementdisorders.org/MDS/Education/Rating-Scales/Training-Programs.htm
[48] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
[49] J. Deng *et al.*, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.
[50] L. Ge, H. Liang, J. Yuan, and D. Thalmann, "3D convolutional neural networks for efficient and robust hand pose estimation from single depth images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1991–2000.
[51] C. Zhu, R. H. Byrd, P. Lu, and J. Nocedal, "Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization," *ACM Trans. Math. Softw.*, vol. 23, no. 4, pp. 550–560, 1997.
[52] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*.
[53] M. Christ, N. Braun, J. Neuffer, and A. W. Kempa-Liehr, "Time series feature extraction on basis of scalable hypothesis tests (tsfresh-A Python package)," *Neurocomputing*, vol. 307, pp. 72–77, 2018.
[54] F. Pedregosa *et al.*, "Scikit-Learn: Machine learning in python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.
[55] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2016, pp. 785–794.
[56] A. Toshev and C. Szegedy, "DeepPose: Human pose estimation via deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 1653–1660.
[57] L. Ge, Y. Cai, J. Weng, and J. Yuan, "Hand PointNet: 3D hand pose estimation using point sets," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 8417–8426.
[58] M. Oberweger and V. Lepetit, "DeepPrior : Improving fast and accurate 3D hand pose estimation," in *Proc. IEEE Int. Conf. Comput. Vis. Workshop*, 2017, pp. 585–594.
[59] I. Shimoyama, T. Ninchoji, and K. Uemura, "The finger-tapping test. A quantitative analysis," *Arch. Neurol.*, vol. 47, no. 6, pp. 681–684, Jun. 1990.
[60] R. Okuno, M. Yokoe, K. Fukawa, S. Sakoda, and K. Akazawa, "Measurement system of finger-tapping contact force for quantitative diagnosis of Parkinson's disease," in *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2007, pp. 1354–1357.
[61] A. Kandori *et al.*, "Quantitative magnetic detection of finger movements in patients with Parkinson's disease," *Neurosci. Res.*, vol. 49, no. 2, pp. 253–260, Jun. 2004.
[62] K. Shima, T. Tsuji, A. Kandori, M. Yokoe, and S. Sakoda, "Measurement and evaluation of finger tapping movements using log-linearized Gaussian mixture networks," *Sensors (Basel)*, vol. 9, no. 3, pp. 2187–2201, 2009.