Spatio-Temporal Attention and Magnification for Classification of Parkinson's Disease from Videos Collected via the Internet

Mohammad Rafayet Ali[1], Javier Hernandez[2],
E. Ray Dorsey[3], Ehsan Hoque[1], and Daniel McDuff[2]
[1] Computer Science, University of Rochester, New York, USA.
[2] Microsoft Research, Redmond, USA.
[3] Center for Health and Technology, University of Rochester Medical Center, New York, USA.

***Abstract*— We present an automated framework for detecting Parkinson's disease (PD) from videos collected through a scalable online platform. We analyzed 1380 videos of age-matched participants performing four standard motor tasks from the MDS-UPDRS. Our proposed framework leverages multiple deep neural networks to temporally and spatially segment the videos as well as magnify relevant motions. Frequency domain representations of the resulting data are then classified using supervised learning. Overall, the proposed framework achieves an accuracy of 82.5% when discriminating between those with PD and those without, and 61.8% when discriminating between those with PD with treatment, with PD without treatment, and those without PD. These results increased up to 91.8% and 73.5%, respectively, when combining the predictions of multiple models. To understand the contributions of each part of our framework we perform systematic ablation studies. We also compare between motion features based on pixel, phase-based and deep learning-based representations. This work demonstrates the possibility of identifying PD cues in challenging real-life settings with inexpensive webcams.**

***Index Terms*— Parkinson's, Deep learning, Online videos, Segmentation.**

I. INTRODUCTION

Parkinson's disease (PD) is a neurological disorder that affects over 10 million people worldwide [1]. PD is characterized by random, involuntary, and non-rhythmic movements of the body [2], [3], [4]. Many individuals only show subtle signs of tremor and do not get diagnosed with PD until the disease has progressed significantly. Even though there is no cure for PD, the use of certain drugs can help manage the symptoms. However, the doses of these drugs need to be carefully tuned based on the disease progression to minimize the side-effects [5]. This assessment process requires in-person visits with a neurologist who assesses the patient through physical examination. However, it is not uncommon that patients' symptoms are not clearly visible during the specific time of the visit. Furthermore, traveling back and forth to the clinic is often difficult for patients with movement disorders. These factors can lead to sparse and potentially inaccurate assessments. This inspired us to develop a novel approach that leverages machine learning to recognize PD symptoms using videos collected with their consent via a webcam from their own home. We believe video-based assessment can be a low-cost, convenient and scalable solution that provides access to care more frequently and remotely.

Unlike many other diseases, there is no objective diagnostic tool for PD. To get evaluated, patients need to perform tasks from the Movement Disorder Society - Unified Parkinson's Disease Rating Scale (MDS-UPDRS) [6] in front of a neurologist or an expert assessor. During the evaluation, the physician looks for different cues of PD such as how smoothly they are performing different tasks, how many breaks they need to have, how accurately the patient is performing the task etc. When considering motor tasks, the expert assessor looks for changes in motion, speed, tremor, breaks, rigidity, and other involuntary non-rhythmic motions which are very indicative of PD. Due to their relevance, this work focuses on the automated analysis of these type of motor tasks.

Computer vision analyses of gait [7] and analyses of wearable data [8] have already shown promise for the detection of PD during motor tasks. But there are still many challenges that need to be addressed [9]. In particular, Espay et al. [9] identify one of the biggest challenges as finding "objective biomarkers that improve the longitudinal tracking of impairments." We believe ubiquitous technologies such as low-cost cameras and computer vision algorithms offer opportunities to address this. In particular, we propose an automated pipeline for assessing PD symptoms from videos collected through an online tool from users homes. This approach comes with a unique set of challenges that have not been addressed by prior work. First, videos collected online often have poor resolution, making it difficult to recover subtle motions. Second, online recordings can be quite heterogeneous with large differences in illumination, background activities and body position of the subject [10]. Third, relevant PD cues may only be captured during a short period of time and/or by a small portion of the frame, leading to large amounts of unnecessary information. Finally, in a remote context it may be more difficult to guarantee that patients perform specific tasks, such as those in the MDS-UPDRS [6], while being in front of a camera and without the direct guidance of an expert. This paper addresses the aforementioned challenges by leveraging recent advancements in computer vision and motion magnification techniques [11],

This work was performed at Microsoft Research and the data for this research originated from the University of Rochester UDALL center, funded by the National Institute of Neurological Disorders and Stroke of the National Institutes of Health under Award Number P50NS108676.

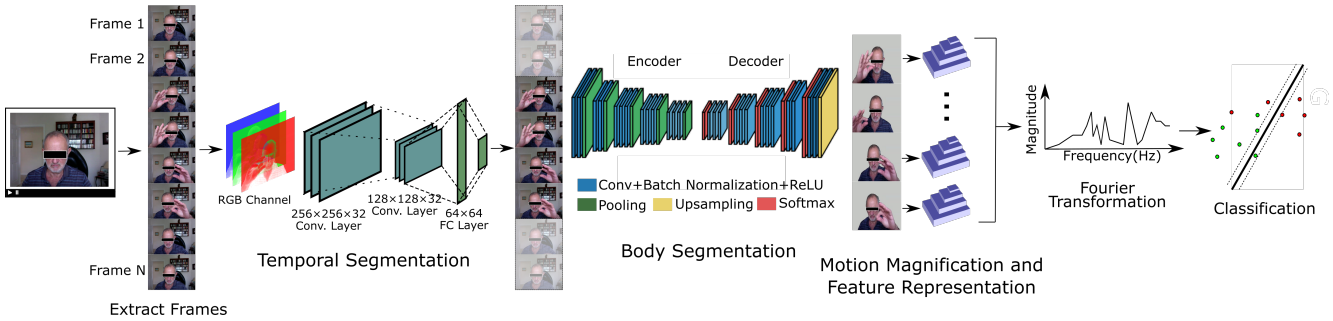


Fig. 1: Overview of our framework: 1) frames are extracted from a video, 2) a CNN-based classifier identifies the frames in which participants are performing each task, 3) a body segmentation algorithm removes the background, 4) motion magnification is applied and magnified pyramids are generated as the feature representation, 5) frequency components are generated by applying Fourier transformation, and 6) an SVM classifier is used to recognize PD cues.

[12], [13]. In particular, we analyze 1380 videos from 345 age-matched (between 55 and 75 of age) individuals performing 4 motor tasks involving hands from MDS-UPDRS. We propose a processing pipeline and evaluate it in a 2-class classification problem (PD = 206, and non-PD N = 139) as well as 3-class classification problem (PD with medication N = 87, PD with medication N = 119, and non-PD N = 139).

Our contributions can be summarized as follows. First, we propose a novel pipeline for detecting PD-related movement disorder from videos. This framework includes components to help find and filter relevant motion information. Second, we evaluate our framework on a 2-class classification problem (non-PD vs PD) and a more challenging and clinically relevant 3-class classification problem (non-PD, PD on medication, PD without medication). Finally, we perform ablation tests to systematically compare different segmentation and feature representation approaches. The remainder of the paper reviews relevant work on the automated detection of PD, the database we use as well as the proposed framework, and provides a thorough analysis and discussion of the results.

II. RELATED WORK

A. Automated PD Detection Tools.

The application of computers in medicine has led to the development of novel approaches for detecting PD using audio signals [14], [15], [16], [17], [18], wearable sensor data [19], [20], and video data [21], [22], [23], [24]. Arora et al. [25] used a smartphone application to assess voice, posture, gait, response time, and finger tapping. In a study with 20 participants, they were able to discriminate between PD and non-PD (with sensitivity = 96.2% and specificity = 96.9%). Sahyoun et al. [26] presented a smartwatch-based application, PakNosis, which aimed to measure PD symptoms remotely using motion tests and qualitative questionnaire. Tzallas et al. [19] built PERFORM, a smart-watch based algorithm which processes the sensor signals to help professionals monitor the severity of PD-related symptoms such as freezing of gait. Lonini et al. [20] studied the value of using wearable sensors at different body locations and used convolutional neural networks to recognize characteristics of PD during regular daily activities. Although smartphones and wearable sensors are common, they often require office visits for measurements [27], [23] and sensors and smartphone are inaccessible to individuals. In contrast, video recording devices are ubiquitous and non-invasive. Additionally, video-conferencing based virtual office visits are getting popular in telemedicine [28]. Uhríková et al. [29] proposed a method to detect motion disorder in video data using Fourier transform and frequency analysis. Orphanidou et al. [30] used accelerometer data from eight participants with PD to detect the freezing of gait events. They applied seven different machine learning algorithms and were able to predict the freezing of gait with over 90% accuracy. Bandini et al. [31] analyzed facial expressions of 17 healthy and 17 PD participants to detect PD-related facial hypomimia. They found that the disgust and anger facial expressions were the most impaired in participants with PD. Meigal et al. [32] proposed a motion video-based PD symptom tracking method. Their experiment shows that video cameras provide reliable capture quality for PD patient motion video tracking.

B. Feature Representations.

Researchers have explored a wide variety of methods to measure and magnify motions in video. Optical flow is the estimation of "apparent velocities of movement of brightness patterns in an image" [33]. As with many tasks in computer vision, deep learning-based architectures currently provide the best results on flow estimation problems [34]. Computational methods can not only be used to measure flow but also to magnify it. Some of the early methods for motion magnification involve estimating motion trajectories using Lagrangian methods which typically involve performing video registration, estimation and maximization steps, clustering of trajectories, and dense optic flow field interpolation [35]. Phase variations of a complex steerable pyramid have been found to capture a good representation of motion in video [12]. Some other approaches have shown that phase-based representations are good at magnifying subtle motions [11] which is important in our applications as many of the motions of interest (e.g., shaking). In addition, pyramid representations can be used to capture motions at different spatial scales. The recent application of DNNs to the task of motion magnification have enabled source specific magnification [13] that reduces the artifacts associated with

(a) Task 1. Finger tapping (b) Task 2. Closing and opening

(c) Task 3. Pronation and supination (d) Task 4. Holding still

Fig. 2: Participants of our study performed four different hand-motor tasks from the MDS-UPDRS [36]. a) Tapping the index finger and thumb 10 times. b) Opening and closing their hand 10 times. c) Rotating their palm 180 degrees from upward facing to downward facing. d) Holding arms and hands outwardly at a 90 degree from the body. Tasks 1-3 were repeated with both hands.

magnifying motions unrelated to the source of interest.

III. DATASET

This work uses an existing dataset collected using the PARK framework (available at www.parktest.net) [36] which allows participants to record videos of themselves and transfer them to the cloud for analysis. Participants were asked to perform seven tasks from the MDS-UPDRS, which includes two speech tasks, one facial expression task, and four motor tasks. As we are interested in motion-based analyses, we have selected only the hand-motor tasks (i.e., tasks related to hands) for our experiments. These tasks are:

1) Finger tapping: Participants tap their index finger and thumb 10 times as fast as possible (see Fig. 2a).
2) Closing and opening: Participants make a fist and then open and close the hand for 10 times as fast as possible (see Fig. 2b).
3) Pronation and supination: Participants stretch out their arms and flip their palms up and down for 10 times (see Fig. 2c).
4) Holding still: Participants stretch out their arms and hold the position for 10 seconds (see Fig. 2d).

The dataset includes videos of 345 individuals performing each of the four different hand movement tasks, totalling 1380 videos. The mean duration of the videos is 9.7 seconds (sd = 6.1). As the severity of PD-related symptoms is highly dependent on medication, participants were asked to self-report the time of their last intake. For part of our analysis, we considered participants who took their medication between the previous 45 (kick-in) and 180 (wore-off) minutes as participants with PD under the influence of medication.

All participants were 50 years old or older. The mean age of the PD participants was 66.8 (sd = 8.1), and the non-PD participants was 63.3 (sd = 5.7). PD recruitment was done via online forums, in-person clinic visits, support groups, mailing lists from the department of Neurology at University of Rochester, and Michael J. Fox foundation. We sent emails to those who were already diagnosed with PD. For non-PD we recruited participants from the local hospitals, Facebook ads., and Amazon Mechanical Turks. Each participant was compensated with a $50 Amazon gift card. After performing the tasks participants completed several surveys regarding their demographics, platform usability, and medication intake. We analyzed those participants who mentioned their age 50 or more. Table I provides a summary of the population.

TABLE I: The Demographic Composition of our Dataset.

	Non-PD	**PD**	
		With Med	W/O Med
N	139	87	119
Female/Male	91/48	56/31	33/86
Age (mean/std)	63.3(5.7)	66.6(11.6)	66.9(4.5)
Country(US/other)	122/17	83/4	94/25
Years since diagnosed (mean/std)	N/A	8.7(5.1)	6.4(9.2)

IV. METHODS

This section describes the proposed framework which includes the segmentation, feature representation, and classification tasks. To normalize the input to the system, all of the videos were preprocessed to a fixed frame rate of 15 fps and a resolution of 256×256 pixels. Figure 1 shows an overview of the proposed framework.

A. Segmentation

When inspecting the videos, we noticed that they contained information that was not associated with the relevant hand-motor task. For instance, several participants were talking to other people (e.g., caregivers, family members), adjusting the camera or their sitting positions, and/or interacting with the recording interface (e.g., pressing "Start" and "Stop" buttons, reading instructions). Even when the participants were performing the relevant task, the hand often occupied a small region of each frame. As a result, a large amount of non-relevant hand-motor information was also captured in the videos. To help find relevant information, we applied different segmentation techniques.

1) Temporal: To help detect frames in which the person was performing the relevant hand-motor task, we implemented a Convolutional Neural Network (CNN) based classifier that detected relevant frames. For example, a frame where a hand is visible would likely be associated with a task and thus should receive a higher probability score than a frame without a hand present. To train the classifier, we used a semi-supervised approach. In particular, we created a dataset consisting 55,200 images from the videos. For the negative class (where participants were not performing the task), we took the first 10 and last 10 frames from the videos. For the positive class, we took 20 frames from the middle of the videos. The first three and last three frames of these 20 frames were manually inspected to ensure that they belonged to the positive class. Therefore, from each video we obtained a total of 40 images which we used to

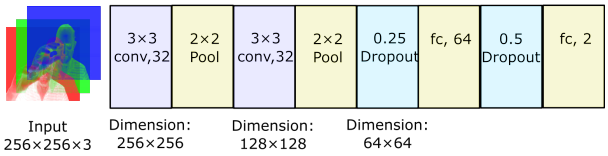


Fig. 3: Architecture of the neural network to temporally segment relevant hand-motor movements. conv = convolutional layer, pool = max. pool layer, fc = fully connected layer.

train a CNN classifier that detects relevant hand-motor tasks. While there are many potential choices for the network, we selected a simplified version of the VGG16 architecture [37] which worked adequately in practice (see Fig. 3). A different classifier was trained for each of the different tasks. We then used a stochastic gradient descent optimizer from the Keras API[1] to train the classifier. When a video of a particular task is given as an input, the classifier outputs the probability of the positive class for each frame. To enforce temporal smoothness, we applied a running mean of the output from 10 consecutive frames. For the remainder of the analysis, we only considered frames for which the likelihood was above a certain threshold (0.5 in our case).

2) Spatial: To help minimize non-relevant background information, we used the body segmentation tool from Microsoft Teams[2]. This body segmentation tool separates the upper body from background information using a CNN encoder-decoder architecture (see Fig. 1). For faster computation, this model utilizes the shortcut method [38], residual connections [39], low dimensional embedding [40], and filter grouping [41], [42]. As a post-processing step, this tool performs blob analysis, temporal and spatial smoothing. The network was trained using CNTK[3] and Tensorflow[4]. The data used for the training of their model included image-mask pairs of people in different poses in front of a green screen, and web-search images with alpha channels. For additional data, certain augmentation methods such as mirroring, scaling, and rotating were utilized.

B. Feature Representations

After performing the temporal and spatial segmentation, we extracted feature representations from each of the video, and computed their variation over time using the fast Fourier transform (FFT). To understand the potential value of different representations, we implemented and compared three types of representations.

1) Unmagnified Raw Pixels (Pixel): Our baseline feature representation is unmagnified raw pixels. This involves constructing a timeseries for each pixel across the length of the video, $\mathbf{x}_{i,j}$, where $\mathbf{x}_{i,j,t}$ is the value of pixel at position i, j at time t. We then compute the frequency components for each pixel vector $\mathbf{x}_{i,j}$ using the FFT. The resulting frequency spectra for all pixels are then averaged.

2) Phase-based Magnified Features (Phase): Our next baseline feature representation is a phase-based representation that has been shown to be effective at capturing subtle motions, such as those related to heartbeats or respiration [12]. In this case, we magnified phase variations of a complex steerable pyramid over time. The complex steerable pyramid is a filter bank that breaks each frame of the video *C(t)* into complex-valued sub-bands corresponding to different scales and orientations. In particular, the basis functions of this transformation are scaled and oriented Gabor-like wavelets with both cosine- and sine-phase components that can separate the amplitude of local wavelets from their phase. The phases are temporally bandpass filtered to isolate specific temporal frequencies relevant to our application (cutoffs: 0.5-2.5 Hz) and remove DC components. These cut-off frequencies were chosen heuristically, based on the frequencies of motions observed when watching a set of videos from the dataset. Finally, we compute the frequency components for each of the magnified phase features and averaged them.

3) Deep Neural Network based Magnified Features (Deep Mag): Traditional phase-based magnification uses frequency properties to separate the target signal from noise. All frequencies in the band of interest (0.5-2.5 Hz) will be magnified equally. However, if the signal of interest is at a similar frequency of another noisy signal, the phase-based magnification approach would magnify both and cause numerous artifacts [13] (see Fig. 5). In this work, we applied a component specific deep neural network based magnification. Recent work proposed the use of neural networks for the task of motion magnification [13]. We apply the same source specific neural network-based architecture and leverage a pre-trained network from [13] to provide initial network weights. Fig. 4 shows the architecture of the neural network.

As this approach is supervised, we fine tuned the network on data from the tasks used in our system. In particular, we captured 18 videos from seven individuals (not featured in the PARK dataset) performing the four tasks. To capture ground truth labels for the gross arm motion, participants wore a wearable sensor[5] on both hands that captured 3-axis accelerometer data. To aggregate the three axis, we took the absolute values across them and then computed the addition of both hands. Using this approach, any hand movement in the video corresponded to changes in the ground truth signal and, therefore, the model is able to magnify motion components associated with hands or arms movements.

Similar to the phase-based approach, we magnified phase variations of a complex steerable pyramid over time. The basis functions of this transformation are scaled and oriented Gabor-like wavelets with both cosine- and sine-phase components. As can be seen in Fig. 5 this did not produce the same artifacts as the phase-based approach.

[1]https://keras.io/
[2]https://products.office.com/en-us/microsoft-teams/group-chat-software
[3]https://github.com/microsoft/CNTK
[4]https://www.tensorflow.org/
[5]https://www.empatica.com/research/e4/

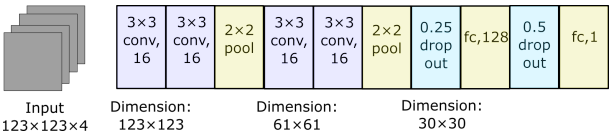


Fig. 4: Architecture of the neural network to extract magnified features. The input of the network is a complex steerable pyramid generated from a recording. conv = convolutional layer, pool = max pool layer, fc = fully connected layer.

C. Classification

After extracting frequency information from each of the feature representations, we then used Support Vector Machines (SVMs) with Radial Basis Function kernel to discriminate the different classes. In particular, we considered a 2-class problem in which we discriminate PD and non-PD participants, and a 3-class problem in which we discriminate PD with treatment, PD without treatment, and non-PD. The hyper-parameters of the SVM were selected using a linear grid search approach ($C = [0.01 - 20], \gamma = [0.001 - 5]$) with step size = 0.1 and maximizing the development set accuracy. After applying the classifier we computed the metrics for evaluation. All the metrics were computed from 10-fold cross validation over five iterations.

V. RESULTS

This section provides the accuracy scores when using each of the four different tasks and the different feature representations. We then investigate the best performing combination of segmentation steps and feature representations via an ablation approach.

Table II shows the accuracy of the two-class classification (PD vs. non-PD) where the columns "None" represents no segmentation, "Temp." represents only temporal segmentation, "Spat." represents spatial segmentation, and "Temp. & Spat." represents both applied in sequence. The bar plots above the tables show the same accuracy scores for a better visualization. The best accuracy scores are highlighted in bold (see Table II). In our dataset 59.7% participants had PD and therefore a naive baseline accuracy that always predicts the most frequent class would be of 59.7%. As can be seen, the worst performing feature representation is the raw unmagnified pixels which yielded a prediction performance close to the baseline. Considering the unmagnified pixel feature representation, the best accuracy (69.8%) was obtained for Task 1 with temporal and spatial segmentation. The phase-based magnification feature representation had better accuracy than the pixel feature representation (78.3%) in Task 2 with temporal and spatial segmentation. The best performance was obtained from the deep neural network-based magnification representation with temporal and spatial segmentation. This model yielded the highest accuracy for all the tasks, yielding 81.1% for Task 1, 82.5% for Task 2, 77.0% for Task 3, and 76.4% for Task 4. Additionally, it should be noted that temporal and spatial segmentation yielded better performance in all of the feature representations compared to each of the segmentations alone. This

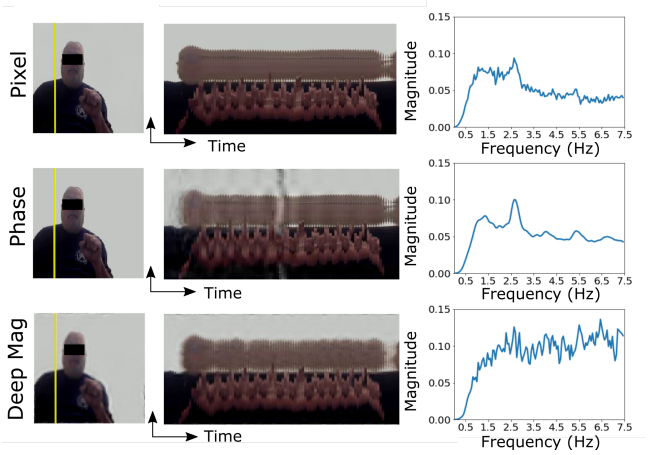


Fig. 5: Scanline images after applying segmentation with three feature representations. On the right column the frequency spectrum generated from each video is shown. Notice how the phase magnification algorithm results in temporal artifacts in the scanline.

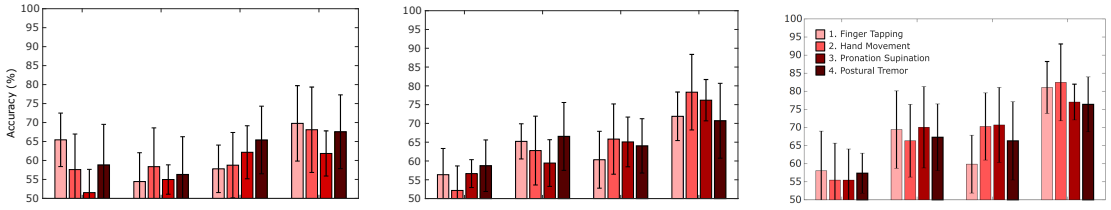
indicates that our temporal and spatial segmentations were able to effectively remove non-relevant information from the data and that both of them contributed to solving the problem.

Table III shows the accuracy of the three-class classification (non-PD, PD with medication, and PD without medication). In this case, the non-PD group is the largest, thus predicting the most likely class yields an accuracy of 40.2%. With pixel feature representation, the classifier performs as good as predicting the most likely class. Phase-based magnification feature representation showed better accuracy than raw pixel feature representation. The best accuracy with phase-based magnification feature representation was 58.8% for Task 3 with temporal and spatial segmentation. Similar to the two-class classification, the deep neural network-based magnification feature representation yielded the best accuracy when considering both temporal and spatial segmentation. For each of the tasks, the model yielded an accuracy of 58.7% for Task 1, 57.9% for Task 2, 62.3% for Task 3, and 61.8% for Task 4. On average the performance was around 50% greater than the baseline.

Table IV shows additional performance metrics of the best performing condition (temporal and spatial segmentation with Deep Mag). We further combine the four tasks with a later-fusion approach to have a model that leverages all the predictions. Specifically, we trained a unique linear SVM model with the predicted labels from each of the tasks. With a leave-one-subject-out cross validation, the accuracy increased to 91.8% and 73.5% for the two-class and three-class classification problems, respectively. Upon further investigating the linear model feature weights, we found that the Task 2 and Task 3 yielded the highest weights, highlighting their contribution in prediction.

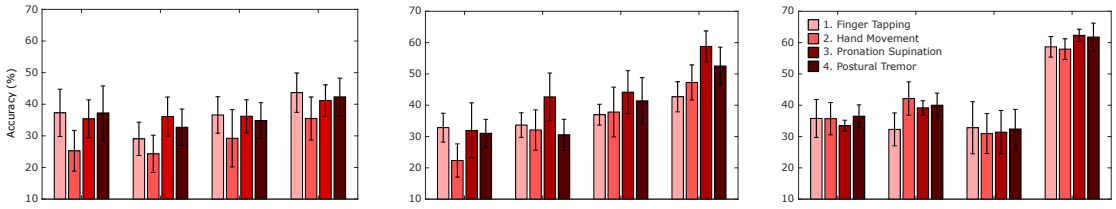
To better understand the difference in performance across feature representations, Fig 5 shows an example scanline image of a video after applying temporal and spatial seg-

TABLE II: TWO-CLASS CLASSIFICATION: Average accuracy (%) and standard error bars.



	Pixel				Phase				Deep Mag			
Task	None	Temp.	Spat.	Temp. &Spat.	None	Temp.	Spat.	Temp. &Spat.	None	Temp.	Spat.	Temp. &Spat.
1	65.4	54.4	57.8	69.8	56.4	65.2	60.3	71.9	58.1	69.4	59.9	**81.1**
2	57.6	58.4	58.8	68.1	52.2	62.8	65.8	78.3	55.5	66.3	70.3	**82.5**
3	51.5	55.0	62.2	61.9	56.6	59.5	65.1	76.2	55.5	70.1	70.7	**77.0**
4	58.8	56.3	65.4	67.6	58.7	66.5	64.0	70.7	57.4	67.3	66.3	**76.4**

TABLE III: THREE-CLASS CLASSIFICATION: Average accuracy (%) and standard error bars.



	Pixel				Phase				Deep Mag			
Task	None	Temp.	Spat.	Temp. &Spat.	None	Temp.	Spat.	Temp. &Spat.	None	Temp.	Spat.	Temp. &Spat.
1	37.3	29.1	36.6	43.7	32.8	33.7	37.0	42.7	35.8	32.3	32.8	**58.7**
2	25.3	24.3	29.2	35.5	22.4	32.1	37.8	47.3	35.7	42.1	30.9	**57.9**
3	35.4	36.1	36.2	41.2	32.0	42.7	44.2	58.8	33.5	39.2	31.4	**62.3**
4	37.2	32.7	34.8	42.3	31.0	30.5	41.4	52.5	36.5	40.0	32.4	**61.8**

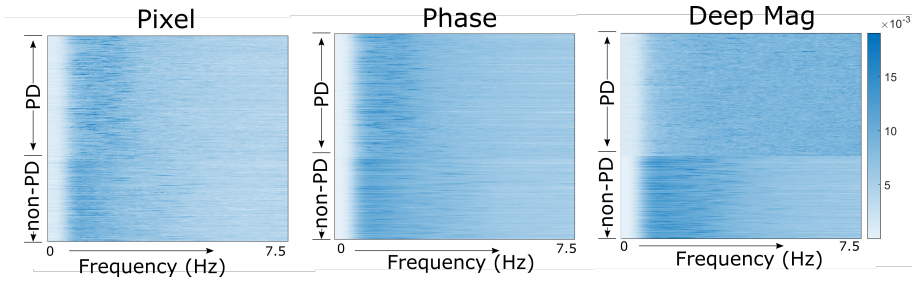


Fig. 6: Frequency distribution of PD and non-PD features after segmentation. Darker color indicates higher power.

mentation. On the right column of Fig. 5 the corresponding frequency components are shown. As can be seen, the Deep Mag approach magnifies the high frequency components of the motion. In this case, PD is characterized by non-rhythmic, involuntary, and random movements, which should be visible in the high frequency components. Since Deep Mag is magnifying the higher frequency components, it is expected to have better accuracy than phase-based magnification and no-magnification. To further investigate how Deep Mag magnification is helping the classification process, Fig. 6 shows the distribution of the frequencies of PD and non-PD participants. From looking at the pixel and phase-based feature representations, it is clear that the frequencies of PD and non-PD are slightly different but difficult to quantify. In contrast, the Deep Mag feature representation shows readily observable differences between PD and non-PD frequencies. In particular, the high frequency components of the PD participants are magnified and spread.

VI. DISCUSSION AND FUTURE WORK

Leveraging recent advancements in remote sensing and machine learning, this work explores the possibility of recognizing hand-motor symptoms of people with PD. We have proposed a framework composed of multiple deep neural networks to address some of the main challenges associated with real-life data recordings. We implemented spatial and temporal segmentation algorithms that allow

TABLE IV: Performance metrics for the best performing model (deep neural network-based magnification feature representation with temporal and spatial segmentation)

	Task	F1	Accuracy	Recall	Precision
Two-Class Classification	1	87.2	81.1	92.5	82.4
	2	87.8	82.5	97.3	80.0
	3	84.2	77.0	86.4	82.0
	4	84.1	76.4	88.9	79.8
Three-Class Classification	1	56.3	58.6	57.0	55.7
	2	59.8	57.9	56.6	63.3
	3	65.1	62.3	61.7	68.9
	4	64.6	61.8	61.4	68.1

finding relevant information, as well as motion amplification algorithms to help capture subtle motions. Furthermore, we have evaluated the proposed framework with a dataset of 1380 video recordings and studied a 2-class classification problem considering PD and non-PD, and a more challenging but clinically relevant 3-class problem considering PD with medication, PD without medication, and non-PD.

Across different experiments, adding temporal and spatial segmentation consistently improved performance. This suggests that they effectively removed non-relevant information which helped boost the signal-to-noise ratio in the resulting motion signals. These segmentation algorithms can be considered as an attention mechanism for the classifier. Considering the standard deviations in the results, however, it is difficult to assess which task may be more informative. To address this, we built an additional model that combined the predictions when considering the different tasks and showed that Task 2 (closing and opening) and Task 3 (pronation and supination) were the most relevant. Furthermore, performance was further increased to 91.8% and 73.5% for the 2-class and 3-class problems, respectively.

Among all the feature representations the deep neural network based magnification worked the best. This was due to the fact that deep magnification magnified the subtle motions in the high frequency range (> 3.5 Hz) (see Fig. 6). Also, the deep magnification model was trained to magnify the spatially relevant movements in a video, unlike the phase magnification, which has no spatial perception. Since the subtle high frequency movements are an indicator of PD, deep magnification were able represent that adequately.

The training data for deep magnification was collected from healthy and younger participants (ages between 22 to 40). This is a limitation of the model, since PD-related tremors were not present in the training samples. This younger data indicates that the deep learning-based motion magnification was not targeting PD-related movements. It simply magnified any motion that was task-related and, hence, the PD-related random motions were present, they were magnified. In the future, we plan to augment the training data for deep magnification with PD participants, as well as for the general, older population.

We used the accuracy score to select and compare the best combination of methods. Although the accuracy score may not be the best metric to describe the performance of a model, since our dataset is not hugely imbalanced, we used it to find the best combination of different approaches. In addition to the accuracy, the precision, recall and F1 scores for the best model showed strong performance.

For part of the analysis, this paper has considered the classification of PD and non-PD. However, PD is a progressive disease without a strong binary separation. However, we believe the proposed framework can be useful in detecting some of the most extreme cases. Our long-term goal is to provide a more continuous score of PD severity which would be consistent. To help address this, we have shown some results considering non-PD and PD on and off medication. However, the sample size for each of the groups was significantly reduced limiting the potential generalization of our findings. Furthermore, we only collected information of the type of their medication at baseline which was *carbidopa-levodopa* in our dataset. Although *carbidopa-levodopa* can reduce PD-related symptoms, it can also induce other side-effects such as *dyskinesia*, which may again show symptoms and could have had an impact on our results.

The framework presented in this paper is an initial exploration and should not be used as a clinical diagnostic tool. However, we feel that the framework could be useful for clinicians who are unable to see all of their patients everyday in order to obtain an objective assessment of the patients' tremors. The framework could look for gradual progression of subtle tremors and, if and when appropriate, suggest a referral to see a neurologist. In addition to PD, we believe the the techniques described in this paper could potentially be relevant to develop tools to identify other potential movement disorders (e.g., essential tremor and Huntington's disease).

VII. CONCLUSION

This paper proposes and evaluates a novel framework for detecting PD from online video recordings. The pipeline addresses some of the issues of real-life noisy recordings by segmenting and magnifying the relevant parts of the videos. With a systematic evaluation, we have shown how different types of segmentation as well as feature representations can help improve the classification. We are looking forward to a future when similar frameworks will help facilitate more frequent and remote monitoring systems for not only PD patients but also other potential movement disorders.

VIII. ACKNOWLEDGMENTS

We would like to thank Taylor Myers, Abdullah Al Mamun, and Taylan Sen for their important role in the data collection, and Mary Czerwinski for providing insightful suggestions and guidance to perform the research.

REFERENCES

[1] Parkinson's Foundation. Statistics, 2018.
[2] Dennis L Kasper, Eugene Braunwald, Anthony S Fauci, Stephen L Hauser, Dan L Longo, J Larry Jameson, et al. Harrison's manual of medicine. 2005.
[3] Kelvin L Chou. Diagnosis and differential diagnosis of parkinson disease. *Waltham (MA): UpToDate*, 2017.
[4] Chauncey Spears. 10 early signs of parkinson's disease, 2019.

[5] San Francisco University of California. Parkinson's disease clinic and research center: Parkinson's disease medications, 2014.

[6] Movement Disorder Society Task Force. The unified parkinson's disease rating scale (updrs): Status and recommendations. *Movement Disorders*, 18(7):738–750, July 2003.

[7] Evanthia E Tripoliti, Alexandros T Tzallas, Markos G Tsipouras, George Rigas, Panagiota Bougia, Michael Leontiou, Spiros Konitsiotis, Maria Chondrogiorgi, Sofia Tsouli, and Dimitrios I Fotiadis. Automatic detection of freezing of gait events in patients with parkinson's disease. *Computer methods and programs in biomedicine*, 110(1):12–26, 2013.

[8] Walter Maetzler, Josefa Domingos, Karin Srulijes, Joaquim J Ferreira, and Bastiaan R Bloem. Quantitative wearable sensors for objective assessment of parkinson's disease. *Movement Disorders*, 28(12):1628–1637, 2013.

[9] Alberto J Espay, Paolo Bonato, Fatta B Nahab, Walter Maetzler, John M Dean, Jochen Klucken, Bjoern M Eskofier, Aristide Merola, Fay Horak, Anthony E Lang, et al. Technology in parkinson's disease: challenges and opportunities. *Movement Disorders*, 31(9):1272–1282, 2016.

[10] Daniel McDuff, Rana El Kaliouby, and Rosalind Picard. Crowdsourced data collection of facial responses. In *Proceedings of the 13th international conference on multimodal interfaces*, pages 11–18. ACM, 2011.

[11] HY Wu, M Rubinstein, E Shih, J Guttag, F Durand, and WT Freeman. Web page: Eulerian video magnification for revealing subtle changes in the world, 2012.

[12] Neal Wadhwa, Michael Rubinstein, Frédo Durand, and William T Freeman. Phase-based video motion processing. *ACM Transactions on Graphics (TOG)*, 32(4):80, 2013.

[13] Weixuan Chen and Daniel McDuff. Deepmag: Source specific motion magnification using gradient ascent. *arXiv preprint arXiv:1808.03338*, 2018.

[14] Max A Little, Patrick E McSharry, Eric J Hunter, Jennifer Spielman, Lorraine O Ramig, et al. Suitability of dysphonia measurements for telemonitoring of parkinson's disease. *IEEE transactions on biomedical engineering*, 56(4):1015–1022, 2009.

[15] GRAEME Macphee. Diagnosis and differential diagnosis of parkinson's disease. *Parkinson's disease in the older patient*, pages 41–75, 2008.

[16] R Frail, JI Godino-Llorente, N Saenz-Lechon, Vıctor Osma-Ruiz, and Corinne Fredouille. Mfcc-based remote pathology detection on speech transmitted through the telephone channel. *Proc Biosignals*, 2009.

[17] Ayyoob Jafari. Classification of parkinson's disease patients using nonlinear phonetic features and mel-frequency cepstral analysis. *Biomedical Engineering: Applications, Basis and Communications*, 25(04):1350001, 2013.

[18] Achraf Benba, Abdelilah Jilbab, and Ahmed Hammouch. Detecting patients with parkinson's disease using mel frequency cepstral coefficients and support vector machines. *International Journal on Electrical Engineering and Informatics*, 7(2):297, 2015.

[19] Alexandros T. Tzallas, Markos G. Tsipouras, Georgios Rigas, Dimitrios G. Tsalikakis, Evaggelos C. Karvounis, Maria Chondrogiorgi, Fotis Psomadellis, Jorge Cancela, Matteo Pastorino, María Teresa Arredondo Waldmeyer, Spiros Konitsiotis, and Dimitrios I. Fotiadis. Perform: A system for monitoring, assessment and management of patients with parkinson's disease. *Sensors (Basel)*, 14(11):21329–21357, November 2014.

[20] Luca Lonini, Andrew Dai, Nicholas Shawen, Tanya Simuni, Cynthia Poon, Leo Shimanovich, Margaret Daeschler, Roozbeh Ghaffari, John A Rogers, and Arun Jayaraman. Wearable sensors for Parkinson's disease: which data are worth collecting for training symptom detection models. *npj Digital Medicine*, 1:64, 2018.

[21] Andrea Bandini, Silvia Orlandi, Hugo Jair Escalante, Fabio Giovannelli, Massimo Cincotta, Carlos A Reyes-Garcia, Paola Vanni, Gaetano Zaccara, and Claudia Manfredi. Analysis of facial expressions in parkinson's disease through video-based automatic methods. *Journal of neuroscience methods*, 281:7–20, 2017.

[22] Lacramioara Dranca, Lopez de Abetxuko Ruiz de Mendarozketa, Alfredo Goñi, Arantza Illarramendi, Irene Navalpotro Gomez, Manuel Delgado Alvarado, and María Cruz Rodríguez-Oroz. Using kinect to classify parkinsons disease stages related to severity of gait impairment. *BMC bioinformatics*, 19(1):471, 2018.

[23] E Ray Dorsey, Alistair M Glidden, Melissa R Holloway, Gretchen L Birbeck, and Lee H Schwamm. Teleneurology and mobile technologies: the future of neurological care. *Nature Reviews Neurology*, 14(5):285, 2018.

[24] E Ray Dorsey, Vinayak Venkataraman, Matthew J Grana, Michael T Bull, Benjamin P George, Cynthia M Boyd, Christopher A Beck, Balaraman Rajan, Abraham Seidmann, and Kevin M Biglan. Randomized controlled clinical trial of virtual house calls for parkinson disease. *JAMA neurology*, 70(5):565–570, 2013.

[25] S. Arora, V. Venkataraman, A. Zhan, S. Donohue, K.M. Biglan, E.R. Dorsey, and M.A. Little. Detecting and monitoring the symptoms of parkinson's disease using smartphones: A pilot study. *Parkinsonism and Related Disorders*, 21(6):650–653, June 2015.

[26] Abdulwahab Sahyoun, Karim Chehab, Osama Al-Madani, Fadi Aloul, and Assim Sagahyroon. Parknosis: Diagnosing parkinson's disease using mobile phones. In *IEEE 18th Int. Conf. e-Health Networking, Applicat. and Services*, pages 387–392, 2016.

[27] E. Ray Dorsey and Eric J. Topol. State of telehealth. *New England Journal of Medicine*, 375(2):154–161, 2016. PMID: 27410924.

[28] CA Beck, DB Beran, KM Biglan, and et al. National randomized controlled trial of virtual house calls for parkinson disease. *Neurology*, 89(11):1152–1161, September 2017.

[29] Zdenka Uhríková and Václav Hlaváč. Periodic motion detection on patient with motion disorders. In *Computer-Based Medical Systems, 2008. CBMS'08. 21st IEEE International Symposium on*, pages 90–92. IEEE, 2008.

[30] Natasa K. Orphanidou, Abir Hussain, Robert Keight, Paulo Lishoa, Jade Hind, and Haya Al-Askar. Predicting freezing of gait in parkinsons disease patients using machine learning. In *2018 IEEE Congress on Evolutionary Computation, CEC 2018, Rio de Janeiro, Brazil, July 8-13, 2018*, pages 1–8, 2018.

[31] Andrea Bandini, Silvia Orlandi, Hugo Jair Escalante, Fabio Giovannelli, Massimo Cincotta, Carlos A. Reyes-Garcia, Paola Vanni, Gaetano Zaccara, and Claudia Manfredi. Analysis of facial expressions in parkinson's disease through video-based automatic methods. *Journal of Neuroscience Methods*, 281:7 – 20, 2017.

[32] A. Y. Meigal, K. S. Prokhorov, N. A. Bazhenov, L. I. Gerasimova-Meigal, and D. G. Korzun. Towards a personal at-home lab for motion video tracking in patients with parkinson's disease. In *2017 21st Conference of Open Innovations Association (FRUCT)*, pages 231–237, Nov 2017.

[33] Berthold KP Horn and Brian G Schunck. Determining optical flow. *Artificial intelligence*, 17(1-3):185–203, 1981.

[34] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2758–2766, 2015.

[35] Ce Liu, Antonio Torralba, William T Freeman, Frédo Durand, and Edward H Adelson. Motion magnification. *ACM transactions on graphics (TOG)*, 24(3):519–526, 2005.

[36] Anonymous. Anonymous for peer review.

[37] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[38] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *CoRR*, abs/1505.04597, 2015.

[39] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[40] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[41] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258, 2017.

[42] Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6848–6856, 2018.