Early-stage Parkinson's disease detection based on optical flow and video vision transformes

Anas Filali Razzouki, Laetitia Jeancolas, Graziella Mangone, Sara Sambin, Alizé Chalançon, Manon Gomes, Stéphane Lehéricy, Jean-Christophe Corvol, Marie Vidailhet, Isabelle Arnulf, et al.

Early-Stage Parkinson's Disease Detection Based on Optical Flow and Video Vision Transformer

Anas Filali Razzouki[1], Laetitia Jeancolas[2], Graziella Mangone[3], Sara Sambin[3], Alizé Chalançon[3], Manon Gomes[3], Stéphane Lehéricy[3], Jean-Christophe Corvol[3], Marie Vidailhet[3], Isabelle Arnulf[3], Mounim A. El-Yacoubi[1], Dijana Petrovska-Delacrétaz[1]

[1] Laboratoire SAMOVAR, Télécom SudParis, Institut Polytechnique de Paris, France
[2] Electrical & Computer Engineering Department, Concordia University, Montreal, Canada
[3] Sorbonne Université, Paris Brain Institute - ICM, Inserm, CNRS, APHP, Hôpital Pitié-Salpêtrière, Paris, France
Corresponding author: mounim.el_yacoubi@telecom-sudparis.eu

***Abstract*—Hypomimia, a symptom of Parkinson's disease (PD), is marked by reduced facial movements and loss of face emotional expressions. This study focuses on identifying hypomimia in individuals with early-stage PD using optical-flow-based video vision transformer. Our study included video recordings from 109 PD and 45 healthy control (HC) subjects with an average of two videos per person (294 videos in total). The participants asked to speak freely while being recorded. To extract typical facial muscle movements from subjects, we computed the optical flow (OF) from the videos. Video vision transformer is then used to infer feature representations from OF and RGB modalities, input to a Random Forest (RF) classifier to classify PD vs. HC. We obtained classification scores up to 83% in terms of balanced accuracy (BA) and an area under the curve (AUC) of 84% at subject level. The results are promising for identifying hypomimia in the early stages of PD, and this research could lead to the possibility of continuous monitoring of hypomimia outside of hospital settings via telemedicine.**

Keywords: Hypomimia-based Parkinson's Disease Detection; Optical Flow; Video Vision Transformer.

1. Introduction

Parkinson's disease (PD) is the second most prevalent neurodegenerative disorder, affecting 1% of the population over the age of 60 [1]. It impacts the central nervous system through the deterioration of dopaminergic neurons in the substantia nigra [2], leading to fundamental motor symptoms like bradykinesia, stiffness, and tremors at rest. Individuals with PD may also suffer from non-motor symptoms such as depression, anxiety, and dysautonomia [3]. These motor symptoms usually appear several years after the onset of neuropathology [4], with approximately 60% of dopaminergic neurons lost by the time of diagnosis [5]. Thus, early detection of PD [6] is essential for administering treatments before substantial irreversible damage to the brain occurs.

Hypomimia, referred to as "masked face", emerges as a prevalent symptom during the early phases of PD [7]. This symptom is characterized by a significant decrease in facial mobility and a marked impairment in expressing emotions through facial expressions. It highlights the extensive neuromuscular deficits associated with PD, severely impacting the ability of patients to communicate non-verbally and presenting considerable obstacles in their social engagement [8]. Neurologists assess hypomimia using the MDS-UPDRS scale [9] by directly observing the patient's face during the examination. The neurologist looks for reduced movement of facial muscles, which might be evident in a reduced frequency, amplitude of blinking and parted lips.

Outside of clinical environments, the detection of hypomimia was explored by using cameras to record patients' faces and analyzing these videos through expert evaluators [10], [11], [12] or by computer vision techniques [13]. Electromyography (EMG) [14] [15] and optoelectronic systems [16] [17], which measure facial muscle activity, have also been utilized. Prior to recording, participants are introduced to various scenarios. Some scenarios evoked their emotional side by asking them to voluntarily produce facial expressions such as smiling, anger, surprise, or disgust [18] [19] upon the clinician's request or by imitating emotive faces displayed on a screen. Other scenarios induced spontaneous emotions, as in studies [20] [21] [10] where participants watched a movie provoking natural emotional responses. Certain scenarios focused on non-emotional facial activities, requesting participants to engage in speech-related tasks such as free speech [22] or repeating syllables [23], or to perform various facial actions, including closing their eyes, moving their eyebrows, or looking downwards [24] [25].

Computer vision techniques for detecting hypomimia from face videos can be categorized into two groups. The first extracts handcrafted motion features from facial landmarks [18] [22] [26] or facial action units (AUs) [23] [25] [27]. These features are then utilized as input for shallow classifiers or statistical tests to distinguish between subjects with PD and HC. The second extracts features automatically from raw signals using Deep Neural Networks (DNNs) [28] [29] [30] [31], which are subsequently inputted into shallow

classifiers or directly into the DNN for classification.

Our approach lies in the second category, which has shown encouraging results recently. For instance, Gomez *et al.* [28] modified a CNN initially intended for face recognition to identify AUs in the EmotioNet dataset. Subsequently, they fine-tuned the model for classifying PD vs. HC, achieving an accuracy of 87.3%. Rajnoha *et al.* [29] used a pretrained CNN to extract embedding features, used as input to a decision tree to classify PD vs. HC, achieving an accuracy of 67.33% on a dataset of 100 subjects (50 PD and 50 HC). Huang *et al.* [32] used a generative adversarial network to create six facial expression images from a neutral image of each of the 95 PD patients in the study. These synthesized images emulated a "non-PD" scenario for these individuals. Subsequently, a deep feature extraction classifier was trained on the original facial expressions of PD patients, their synthesized counterparts, and standard facial expression images from additional public datasets to classify PD vs. HC. They obtained a 99% accuracy using 5-fold cross-validation. The RGB images in these studies represent only the video spatial component, and thus the dynamics is lost. To consider the latter, Su *et al.* [33] incorporated dynamic data by merging OF with RGB to construct a dual-stream CNN, trained on a dataset of 47 PD and 39 HC. They reached an accuracy of 99% on the test set comprising 30% of the data. Valenzuela *et al.* [30] proposed a spatio-temporal 3D CNN to classify PD vs. HC. The dataset comprised 16 PD and 16 HC. The model was trained at different temporal resolutions, achieving the highest accuracy of 91% with input sequences of 14 frames. Huang *et al.* [31] used a 3D attention-based convolution layer to extract features, processed then by LSTM to capture contextual information for end-to-end PD vs. HC classification. The dataset comprised patients various stages of PD (stage 1: 54, stage 2: 145, stage 3: 75, stage 4: 26, according to the Hahn-Yahr scale) and 77 HC subjects. They discriminated stage 2 PD patients from HC with an 88% accuracy and accomplished a 60% accuracy in multi-classification, distinguishing between stage 1 PD, stage 2 PD patients, and HC.

Despite their promising results, the studies above are usually conducted on small datasets. Also, except [31], PD patients were not in their early stage. Furthermore, some of these works sometimes optimize certain hyperparameters on the test set, which may introduce bias to the results.

We propose a study for identifying hypomimia in early-stage PD patients through the analysis of their facial videos. Participants in the study were asked to speak freely while being recorded. To characterize the facial muscles' movement from the videos, we calculated the optical flow (OF). Subsequently, we derived embedding features from both RGB and OF images using a video vision transformer model (VideoMAE) and a CNN model (ResNet34), which serves as our baseline. For classification, we used a Random Forest (RF) that processes the four resulting feature representations, namely ResNet34-RGB, ResNet34-OF, VideoMAE-RGB, and VideoMAE-OF, to classify PD vs. HC. To the best of our knowledge, this is the first application of a video vision transformer to OF modality for detecting hypomimia.

Our paper is structured as follows: Section 2 introduces the database, the OF extraction process, and the video vision transformer model. Section 3 highlights the main results. Section 4 details the findings and their analysis. The paper concludes by discussing future work in Section 5

2. Methods

2.1. Database

Data collection was conducted in accordance with the ICEBERG protocol (clinicaltrials.gov, NCT02305147) guidelines, a continuous longitudinal study carried out at the Paris Brain Institute (ICM). The objective of this research is to identify and validate biomarkers for PD. Eligibility for the study was determined based on the following criteria: (1) Participants diagnosed with idiopathic Parkinson's Disease according to the standards of the United Kingdom Parkinson's Disease Society Brain Bank [34], with the condition not exceeding four years in duration; and (2) Control subjects were required to have no neurological disorders.

For five years, participants engaged in yearly visits to ICM, where they underwent a variety of assessments, including neurological exams, evaluations of motor and cognitive capabilities, collection of biological samples, brain MRI imaging, and audio-visual recordings. Subjects with PD received pharmacological treatment, and their faces were recorded in the ON phase. Informed consent was obtained from all participants prior to research activities, and the study was approved by the ethics committee (IRBParis VI, RCB: 2014-A00725-42) in accordance with local guidelines.

The ICEBERG Video-Feb2023 dataset considered in this study includes 294 videos: 203 videos for PD (126 men, 77 women) and 91 videos of HC (58 men, 33 women). Details regarding age and clinical assessments in the OFF state (including Hoehn and Yahr stage, MDS-UPDRS III score, MDS-UPDRS III face item, and MoCa score) at the first video visit are outlined in Table 1. During the recording session, lasting between 15 to 20 minutes, participants are instructed to perform 29 vocal tasks, with guidance provided through a user interface. In this study, we investigated a monologue task where participants were asked to speak freely. As this task involves natural speech, it mimics real-life communication and may potentially help detecting hypomimia in such contexts. The recording equipment used was the Logitech C922 Pro Stream Webcam, model number 195, with built-in encoding and compression capabilities. It operates at a frame rate of 24 frames per second (fps) and offers a resolution of 1920x1080 pixels.

2.2. Preprocessing

2.2.1. Face extraction

We used OpenSeeFace [35], a facial tracker, to extract the face in each frame. This system utilized facial landmarks for adjustments by accurately delineating the face's bounding box. Its face detection relies on the RetinaFace model [36], while its identification of facial landmarks is based on the MobileNetV3 model, trained on the LS3D-W dataset [37].

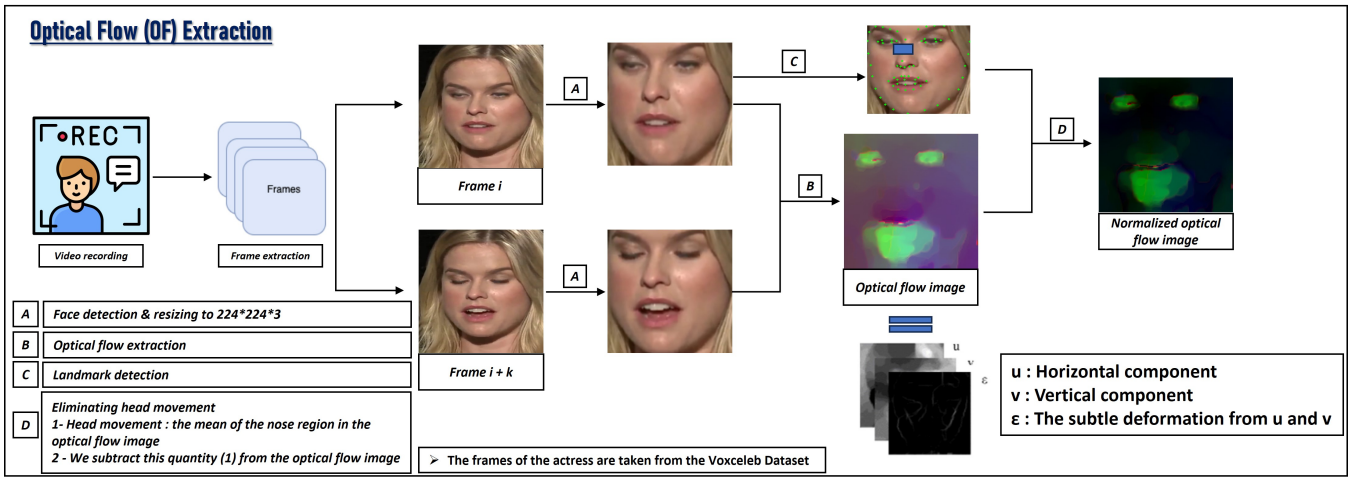

Figure 1. Optical flow (OF) extraction process

TABLE 1. INFORMATION RELATED TO THE ICEBERG VIDEO-FEB2023 DATASET RECORDINGS.

	PD		HC	
Biological sex	Male	Female	Male	Female
No. of videos	126	77	58	33
No. of subjects	70	39	26	19
Age (years)	64.2 ± 9.4	65.6 ± 8.6	63.4 ± 9.5	63.1 ± 8.5
Hoehn & yahr	1.9 ± 0.3	1.9 ± 0.3	0 ± 0	0 ± 0
MDS-UPDRS III total	33.9 ± 6.9	28.9 ± 8.3	3.9 ± 2.7	5.5 ± 3.3
MDS-UPDRS III face item	1.1 ± 0.5	0.9 ± 0.4	0 ± 0	0 ± 0
MoCA	26.5 ± 2.6	28.1 ± 1.7	27.6 ± 2	27.9 ± 1.7

2.2.2. Optical flow (OF) extraction

We adapted the technique in [38] to calculate the OF. First, the facial region in each frame is cropped and resized to 224 × 224 pixels. Subsequently, three OF components are computed between two frames, by considering a step k, i.e. by considering, for each frame index i, the OF between frame F_{i+k} and frame F_i. k, a parameter for capturing meaningful facial muscle movements, can be tuned according to the frame rate and the task scenario. We set $k = 7$ as, in our previous work on AUs [23], it was the optimal value for capturing subtle movements that discriminate PD from HC. Horizontal and vertical components, u and v are computed using TV-L1 OF method [39]. Optical strain (ϵ), derived from infinitesimal strain theory, measures the slight facial deformations from OF components [40]. To eliminate global head motion, we take the landmark position of the nose region with a five-pixel margin in each frame F_i. We then subtract the movement associated with this region from the OF image. Figure 1 shows the pipeline for extracting the OF components. The three elements ($\mathbf{u}$, $\mathbf{v}$, and ϵ) constitute the input data for model training and inference.

2.3. DNN models

2.3.1. ResNet34

ResNet34 is a derivative of the ResNet family [41], engineered with 34 layers, with residual blocks that leverage skip connections. The architecture comprises 33 convolutional layers for feature extraction, with a max-pooling layer with a 3x3 kernel size for down-sampling, followed by an average pooling layer before the final fully connected layer.

2.3.2. VideoMAE

VideoMAE [42] is a transformer-based architecture that employs a self-supervised learning approach specifically designed for video data. It focuses on a simple yet effective design principle: masking and reconstructing video cubes. The architecture comprises an asymmetric encoder-decoder framework where the encoder in this architecture processes only the unmasked video cubes, while the decoder is tasked with reconstructing the original video from this input.

2.4. Feature extraction

We used the VideoMAE-base model, which underwent pre-training on the Kinetics-400 dataset [43] through a self-supervised process over 1600 epochs. To extract features, the VideoMAE encoder was applied, generating 768 embedding features from an input chunk (a subset of frames), either RGB or OF, with dimensions $(16, 3, 224, 224)$. Here, 16 represents the number of images (chunk size), and $(3, 224, 224)$ specifies the color channels and spatial dimensions of each frame for RGB, or the three components $(\mathbf{u}, \mathbf{v}, \epsilon)$ as described in subsection 2.2.2, when using OF.

To compare the video vision transformer model with a baseline CNN, ResNet34 trained on the ImageNet dataset. This resulted in the generation of 512 embedding features from each image, whether it was RGB or OF.

2.5. Classification

Given a feature representation, we used RF to classify PD vs. HC. We evaluated RF using stratified nested cross validation (CV) with five folds in the inner and outer loops. The stratification split is based on clinical condition and biological sex to ensure equal representation of PD/males, PD/females, HC/males, HC/females across the training, validation, and test sets. Note that same-subject videos are exclusively included in one of the three sets.

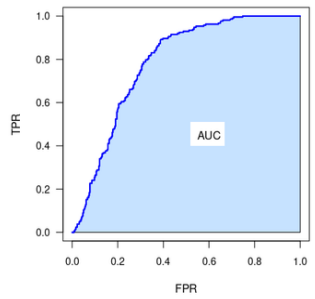

Figure 2. ROC curves and area under curve (AUC)

RF training is conducted at the frame level when extracting embedding features with ResNest34, and at the chunk level with VideoMAE. Consequently, RF inference yields a classification score (ranging from 0 to 1) for PD and HC for each frame or chunk. The mean classification scores across the video frames or chunks is then used to classify PD vs. HC. To obtain a classification score for each subject, we average the classification scores from all video visits related to that subject. Finally, balanced accuracy (BA) and area under the curve (AUC) metrics are employed to evaluate model performance at both the visit and subject levels.

The RF hyperparameters were optimized by performing an inner loop of CV with five folds in each training fold of the outer loop. The best hyperparameters were selected according to the highest average BA on the validation folds.

2.6. Evaluation metrics

Balanced Accuracy (BA) is defined as the average of sensitivity (Se) (the rate of correct classifications among subjects with PD) and specificity (Sp) (the rate of correct classifications among control subjects).

$$Specificity = \frac{TrueNegatives}{TrueNegatives + FalsePositives}$$

$$Sensitivity = \frac{TruePositives}{TruePositives + FalseNegatives}$$

$$Balanced\ accuracy = \frac{Sensitivity + Specificity}{2}.$$

The Area Under the Curve (AUC) is a metric used to assess model performance in binary classification tasks. It quantifies the quality of a model's predictions by measuring the area beneath the receiver operating characteristic (ROC) curve. This curve, illustrated in figure 2, plots the true positive rate (TPR) against the false positive rate (FPR) across various threshold settings. In other terms, the AUC represents a numeric evaluation of the model's capacity to differentiate between positive and negative classes, according to the ranking of their predicted probabilities.

3. Results

Table 2 presents the test set performance of the RF classifier distinguishing between PD and HC under both RGB and OF modalities, utilizing two different feature extractors: VideoMAE and ResNet34. With the RGB modality, the VideoMAE feature extractor achieved a BA of 77% and an AUC of 82% at the subject level. This performance significantly surpasses that of the ResNet34 feature extractor, which produced results comparable to random chance, with both BA and AUC approximately 50%. The VideoMAE feature extractor outperforms the ResNet34 in detecting hypomimia, as the former is tailored to process temporal information, while the latter is tailored to process static information. The VideoMAE model processes sequences of 16 video frames, enabling it to capture dynamic changes in the face during speech over time. This is crucial for detecting hypomimia, which involves subtle and gradual changes in facial muscle activity. In contrast, ResNet34 processes only a single static image, lacking the capacity to analyze temporal patterns and motion cues.

Using the OF modality, the VideoMAE and ResNet34 features extractors appears to yield similar results, ranging from 77% to 79% in terms of BA and 82% to 83% in terms of AUC at subject level. These results indicate that when using OF instead of RGB for detecting hypomimia in the context of free speech, ResNet34 is now effective and achieves results comparable to those of VideoMAE, due to its ability now to process temporal dynamics. OF maps the movement and changes in pixels between sequential frames, providing vital insights into the subtle facial movements that are crucial for diagnosing hypomimia.

When comparing the OF and RGB modalities with VideoMAE, we found that both exhibited similar results at the subject level, achieving a similar AUC of 82% and a BA ranging from 77% to 79%. When combining the OF and RGB modalities by averaging the classification scores from both approaches for each subject, we attained a BA of 83% and an AUC of 84%, which corresponds to an error rate reduction of 19% for BA and 11% for AUC.

4. Discussion

We have introduced an automated method for detecting hypomimia in early-stage PD using facial video recordings. Participants in the study were asked to speak freely while being recorded. Our technique involves measuring facial muscle movements by computing the OF. RGB and OF are then inputted into a video vision transformer model (VideoMAE) and a CNN model (ResNet34) to extract embedding features. The resulting four feature representations are subsequently used as input for a RF classifier to classify PD vs. HC.

Using the VideoMAE feature extractor, we achieved a BA of 83% and an AUC of 84% by integrating OF and RGB. Our study's findings reaffirmed hypomimia as a prevalent symptom in early-stage PD and demonstrated the capability to detect hypomimia through natural speech patterns. In a similar speech task, these results are consistent with prior studies. The first study [22] achieved an accuracy of 78% and an AUC of 87% by using geometric features from facial landmarks. The second [29] extracted embedding features from the first RGB frame and used a decision tree to classify PD vs. HC, achieving an accuracy of 67%.

Our findings highlight the significance of using dynamic

TABLE 2. CLASSIFICATION RESULTS FOR EACH MODALITY ACROSS DIFFERENT FEATURE EXTRACTOR MODELS USING RF CLASSIFIER.

		Visit Based				Subject Based			
Modality	Model	Se	Sp	BA	AUC	Se	Sp	BA	AUC
RGB	MVAE	72%	73%	73%	78%	78%	76%	77%	82%
	ResNet34	77%	21%	49%	48%	81%	16%	48%	47%
OF	MVAE	68%	83%	76%	78%	71%	87%	79%	82%
	ResNet34	70%	80%	75%	80%	72%	82%	77%	83%
Combine OF and RGB	MVAE	72%	76%	74%	80%	79%	87%	**83%**	**84**%

features in free speech tasks for detecting hypomimia. Unlike static feature representations extracted from a single RGB image with ResNet34, which proved ineffective in detecting hypomimia, previous research by Gomez et *al* [44] on patients showing emotions demonstrated different results. In that study, static features extracted from Apex images using a pretrained face recognition model reached PD classification accuracies between 66.00% and 77.36%, varying with the type of facial expression. The study differed from ours in the scenario type as it focused on facial expression of emotions employing Apex images that capture the most pronounced expressions at the peak of a sequence, which could be effective in identifying reduced expressivity in people with PD. However, the authors found that accuracies for dynamic features, quantified using facial landmarks, were lower, ranging from 66.00% to 71.15%. Nevertheless, when both static and dynamic features were incorporated, the results showed improvements in system performance, with 13.46% increase in PD detection accuracy.

We also showed that the video vision transformer is effective in detecting hypomimia when taking as input directly the RGB images, confirming the findings of [31] regarding PD severity assessment in the context of emotional expression. The core feature of this model is the attention mechanism, which analyzes each patch of an image in the context of its relevance to other patches, both within the same frame and across multiple frames. Thus, it allows the model to process different information sub-spaces simultaneously, which is crucial for integrating diverse cues like motion and spatial relationships over time. As a result, this model achieves comparable outcomes to those obtained when using a ResNet34 to analyze features extracted from OF, which are specifically designed to capture facial movements over time.

To better capture subtle facial muscle movements, we proposed the VideoMAE model to extract features from OF images, which demonstrated results comparable to previous methods. Nevertheless, by integrating this approach with the RGB modality where the VideoMAE model also serves as feature extractor, we have achieved the best results at the subject level, reaching a BA of 83% and an AUC of 84%.

Our study has limitations due to the relatively small dataset, although it is larger than existing ones. It also has distribution imbalances between PD and HC, and in biological sex within the dataset, despite our verification of the non-significance of the sex confounding effect.

5. Conclusion

We have introduced a novel method for detecting hypomimia in subjects with early stage PD, characterizing the movement of facial muscles through an optical-flow-based video vision transformer. Our approach has achieved classification scores of 83% for BA and 84% for AUC at the subject level. This work demonstrates the valuable benefit of including OF and video vision transformer into the framework of PD early diagnosis or monitoring. In the future, we will investigate integrating optical-flow-based video vision transformer with our approaches based on AUs [23]. Furthermore, we plan to use these facial features to estimate the severity of hypomimia using the MDS-UPDRS3 face item, and to predict additional clinical measures such as MDS-UPDRS3, Hahn Yahr stage, and rigidity score. Additionally, given that our dataset encompasses data collected over five years, we will utilize these longitudinal data for stratification, by categorizing patients into distinct groups. We will also explore a multimodal approach by considering other modalities harnessed for neurodegenerative disease assessment, such as as voice [45], and, handwriting [46].

Acknowledgment

The authors thank V. Maheo for her contributions to the acquisitions, Dr Habib Benali for his support and all the subjects for their participation in this research.

Funding

This work was supported by DIGIPD (01KU2110), a project supported by the Fund for Scientific Research (F.R.S.–FNRS), the Agence Nationale de la Recherche (ANR), the Federal Ministry of Education and Research (BMBF), the Federal Ministry of Health (BMG), and the National Research Fund (FNR), under the frame of ERA PerMed. It was also supported by grants from ANR Nucleipark, DHOS-Inserm, France Parkinson, Ecole des NeuroSciences de Paris (ENP), Fondation pour la Recherche Médicale (FRM), and the Investissements d'Avenir, IAIHU-06 (Paris Institute of Neurosciences – IHU), ANR-11-INBS-0006, Fondation d'Entreprise EDF, BIOGEN Inc., Fondation Saint Michel, Fondation Thérèse and René Planiol, JPND Control-PD (ANR-21-JPW2-0005-05), Unrestricted support for Research on Parkinson's disease from Energipole and Société Française de Médecine Esthétique.

References

[1] O.-B. Tysnes and A. Storstein, "Epidemiology of parkinson's disease," *Neural Transm*, vol. 124, pp. 901–905, 2017.

[2] L. Darden, "Mechanisms and models," *Cambridge Companion to Philos. Biol.*, vol. 39, 2007.

[3] F. Salawu, A. Danburam, and A. Olokoba, "Non-motor symptoms of Parkinson's disease: diagnosis and management." *Nigerian J. Med*, 2010.

[4] C. H. Hawkes, K. Del Tredici, and H. Braak, "A timeline for parkinson's disease," *Parkinsonism Relat Disord*, vol. 16, 2010.

[5] B. R. Haas, T. H. Stewart, and J. Zhang, "Premotor biomarkers for parkinson's disease-a promising direction of research," *T. Neurodegener.*, vol. 1, 2012.

[6] H. Fröhlich *et al.*, "Leveraging the potential of digital technology for better individualized treatment of parkinson's disease," *FIN*, 2022.

[7] J. Jankovic, "Parkinson's disease: Clinical features and diagnosis," *JNNP*, 2008.

[8] S. D. Gunnery, B. Habermann, M. Saint-Hilaire, C. A. Thomas, and L. Tickle-Degnen, "The relationship between the experience of hypomimia and social wellbeing in people with Parkinson's disease and their care partners," *Parkinsons Dis*, 2016.

[9] M. D. S. T. F. on Rating Scales for Parkinson's Disease, "The unified parkinson's disease rating scale (updrs): status and recommendations," *Mov Disord*, vol. 18, no. 7, pp. 738–750, 2003.

[10] G. Simons, M. C. S. Pasqualini, V. Reddy, and J. Wood, "Emotional and nonemotional facial expressions in people with parkinson's disease," *JINS*, vol. 10, no. 4, pp. 521–535, 2004.

[11] Cannavacciuolo *et al.*, "Facial emotion expressivity in patients with parkinson's and alzheimer's disease," *J.Neural Transm*, vol. 131, no. 1, pp. 31–41, 2024.

[12] Ricciardi *et al.*, "Emotional facedness in parkinson's disease," *J.Neural Transm*, vol. 125, pp. 1819–1827, 2018.

[13] E. Bianchini *et al.*, "The story behind the mask: A narrative review on hypomimia in parkinson's disease," *Brain Sciences*, vol. 14, no. 1, p. 109, 2024.

[14] J. Kang, D. Derva, D.-Y. Kwon, and C. Wallraven, "Voluntary and spontaneous facial mimicry toward other's emotional expression in patients with parkinson's disease," *PloS one*, vol. 14, 2019.

[15] S. R. Livingstone, E. Vezer, L. M. McGarry, A. E. Lang, and F. A. Russo, "Deficits in the mimicry of facial expressions in parkinson's disease," *Front Psychol*, vol. 7, 2016.

[16] M. Bologna *et al.*, "Altered kinematics of facial emotion expression and emotion recognition deficits are unrelated in parkinson's disease," *Front Neurol*, vol. 7, p. 230, 2016.

[17] L. Marsili *et al.*, "Bradykinesia of posed smiling and voluntary movement of the lower face in parkinson's disease," *Parkinsonism Relat. Disord*, vol. 20, no. 4, pp. 370–375, 2014.

[18] A. Bandini *et al.*, "Analysis of facial expressions in parkinson's disease through video-based automatic methods," *J. Neurosci. Methods*, vol. 281, pp. 7–20, 2017.

[19] D. Bowers *et al.*, "Faces of emotion in parkinsons disease: micro-expressivity and bradykinesia during voluntary facial expressions," *J. Int. Neuropsychol. Soc*, vol. 12, 2006.

[20] N. Vinokurov, D. Arkadir, E. Linetsky, H. Bergman, and D. Weinshall, "Quantifying hypomimia in parkinson patients using a depth camera," in *Int. Symp. Perv. Comput. Paradigms Ment. Health*. Springer, 2015.

[21] P. Wu *et al.*, "Objectifying facial expressivity assessment of parkinson's patients: preliminary study," *Comput Math Methods Med*, vol. 2014, 2014.

[22] M. Novotny, T. Tykalova, H. Ruzickova, E. Ruzicka, P. Dusek, and J. Rusz, "Automated video-based assessment of facial bradykinesia in de-novo parkinson's disease," *NPJ Digit. Med.*, vol. 5, no. 1, pp. 1–8, 2022.

[23] A. Filali Razzouki *et al.*, "Early-stage parkinson's disease detection based on action unit derivatives," in *JETSAN*, 2023.

[24] K. Clawson, L. S. Delicato, S. Garfield, and S. Young, "Automated representation of non-emotional expressivity to facilitate understanding of facial mobility: Preliminary findings," in *IntelliSys*. IEEE, 2017, pp. 779–785.

[25] A. Moshkova, A. Samorodov, N. Voinova, A. Volkov, E. Ivanova, and E. Fedotova, "Studying facial activity in parkinson's disease patients using an automated method and video recording," in *FRUCT*. IEEE, 2021, pp. 301–308.

[26] J. Skibińska and J. Hosek, "Computerized analysis of hypomimia and hypokinetic dysarthria for improved diagnosis of parkinson's disease," *Heliyon*, vol. 9, no. 11, 2023.

[27] J. A. Priebe, M. Kunz, C. Morcinek, P. Rieckmann, and S. Lautenbacher, "Does parkinson's disease lead to alterations in the facial expression of pain?" *J. Neurol. Sci*, vol. 359, no. 1-2, pp. 226–235, 2015.

[28] L. F. Gomez, A. Morales, J. Fierrez, and J. R. Orozco-Arroyave, "Exploring facial expressions and action unit domains for parkinson detection," *PLoS ONE*, 2023.

[29] M. Rajnoha, J. Mekyska, R. Burget, I. Eliasova, M. Kostalova, and I. Rektorova, "Towards identification of hypomimia in parkinson's disease based on face recognition methods," *ICUMT*, 2019.

[30] B. Valenzuela, J. Arevalo, W. Contreras, and F. Martinez, "A spatio-temporal hypomimic deep descriptor to discriminate parkinsonian patients," in *EMBC*. IEEE, 2022.

[31] J. Huang *et al.*, "Parkinson's severity diagnosis explainable model based on 3d multi-head attention residual network," *Comput. Biol. Med.*, vol. 170, p. 107959, 2024.

[32] W. Huang, W. Xu, R. Wan, P. Zhang, Y. Zha, and M. Pang, "Auto diagnosis of parkinson's disease via a deep learning model based on mixed emotional facial expressions," *IEEE J-BHI*, 2023.

[33] G. Su *et al.*, "Hypomimia recognition in parkinson's disease with semantic features," *TOMM*, vol. 17, 2021.

[34] W. Gibb and A. Lees, "The relevance of the lewy body to the pathogenesis of idiopathic parkinson's disease." *JNNP*, vol. 51, no. 6, pp. 745–752, 1988.

[35] OpenSeeFace. (2017) github.com. https://github.com/emilianavt/OpenSeeFace. [Online; accessed 28-April-2024].

[36] J. Deng, J. Guo, Y. Zhou, J. Yu, I. Kotsia, and S. Zafeiriou, "Retinaface: Single-stage dense face localisation in the wild," in *arxiv*, 2019.

[37] A. Bulat and G. Tzimiropoulos, "How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks)," in *ICCV*, 2017.

[38] G.-B. Liong, J. See, and L.-K. Wong, "Shallow optical flow three-stream cnn for macro-and micro-expression spotting from long videos," in *ICIP*. IEEE, 2021, pp. 2643–2647.

[39] J. Sánchez Pérez, E. Meinhardt-Llopis, and G. Facciolo, "TV-L1 Optical Flow Estimation," *IPOL*, vol. 3, pp. 137–150, 2013.

[40] M. Shreve, J. Brizzi, S. Fefilatyev, T. Luguev, D. Goldgof, and S. Sarkar, "Automatic expression spotting in videos," *IVC*, vol. 32, no. 8, pp. 476–486, 2014.

[41] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[42] Z. Tong, Y. Song, J. Wang, and L. Wang, "Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training," *NeurIPS*, vol. 35, pp. 10 078–10 093, 2022.

[43] W. Kay *et al.*, "The kinetics human action video dataset," *arXiv*, 2017.

[44] L. F. Gomez, A. Morales, J. R. Orozco-Arroyave, R. Daza, and J. Fierrez, "Improving parkinson detection using dynamic features from evoked expressions in video," in *IEEE/CVF*, 2021, pp. 1562–1570.

[45] L. Jeancolas *et al.*, "X-vectors: new quantitative biomarkers for early parkinson's disease detection from speech," *Front Neuroinform*, vol. 15, p. 578369, 2021.

[46] M. A. El-Yacoubi, S. Garcia-Salicetti, C. Kahindo, A.-S. Rigaud, and V. Cristancho-Lacroix, "From aging to early-stage alzheimer's: Uncovering handwriting multimodal behaviors by semi-supervised learning and sequential representation learning," *Pattern Recognit*, vol. 86, 2019.