Estimation of Orofacial Kinematics in Parkinson's Disease: Comparison of 2D and 3D Markerless Systems for Motion Tracking

Diego L. Guarin[1], Aidan Dempster[1], Andrea Bandini[1], Yana Yunusova[1,2,3] and Babak Taati[1,4,5]

[1] KITE — Toronto Rehabilitation Institute — University Health Network, Toronto, ON, Canada.
[2] Department of Speech Language Pathology, University of Toronto, Toronto, ON, Canada.
[3] Hurvitz Brain Sciences Program, Sunnybrook Research Institute, Toronto, ON, Canada.
[4] Institute of Biomaterials and Biomedical Engineering, University of Toronto, Toronto, ON, Canada.
[5] Department of Computer Science, University of Toronto, Toronto, ON, Canada.

***Abstract*— Orofacial deficits are common in people with Parkinson's disease (PD) and their evolution might represent an important biomarker of disease progression. We are developing an automated system for assessment of orofacial function in PD that can be used in-home or in-clinic and can provide useful and objective clinical information that informs disease management. Our current approach relies on color and depth cameras for the estimation of 3D facial movements. However, depth cameras are not commonly available, might be expensive, and require specialized software for control and data processing. The objective of this paper was to evaluate if depth cameras are needed to differentiate between healthy controls and PD patients based on features extracted from orofacial kinematics. Results indicate that 2D features, extracted from color cameras only, are as informative as 3D features, extracted from color and depth cameras, differentiating healthy controls from PD patients. These results pave the way for the development of a universal system for automatic and objective assessment of orofacial function in PD.**

I. INTRODUCTION

Orofacial symptoms are common in Parkinson's disease (PD) [1], and the severity of orofacial motor disorders may be correlated with the severity of PD as measured by clinical scales [2], [3]. Thus, there has been great interest in developing techniques to objectively evaluate orofacial movements in PD. Researchers have introduced multiple approaches to measure the kinematic characteristics of the tongue, jaw, and lips, such as position and velocity, during speech production [4]–[7]. These methods provide objective information on how PD affects normal function and have the potential to improve the clinical management of the disease. However, the clinical utility of these techniques is limited as they depend on expensive and difficult-to-use systems, such as optical motion tracking systems, electromagnetic articulography, and electromyography.

Bandini *et al.* introduced an alternative approach to estimate orofacial kinematics that uses a color and depth cameras. The technique employs a computer vision model to automatically estimate the position of facial landmarks – defining the location of the eyebrows, eyes, nose, mouth, and jawline – in videos acquired with the color camera. The information provided by the depth camera is subsequently used to reconstruct the 3D *real-word* locations of the facial landmarks [8]–[10]. Results demonstrated that this approach was able to differentiate between healthy controls and PD patients based only on the information provided by a few short video segments of subjects performing a task commonly applied during neurological examination [11].

This approach represents an important step towards an objective, automatic, cost-effective, and universally available system for assessment of orofacial deficits in PD. Such a system would be available for in-clinic or in-home use with little or no intervention from movement disorder specialists, and would provide clinically useful information that informs disease management. However, the clinical utility of the system proposed by Bandini *et al.* is limited by the necessity of a depth camera, which might not be available in the clinic or at home. Thus, a truly universally available system should rely only on color cameras, which are already available on mobile phones, tablets, and personal computers.

To our knowledge, only one work analyzed the ability of 2D and 3D orofacial kinematic features to differentiate between healthy controls and PD patients. Bandini *et al.* analyzed one speech task – repetition of the syllable /pa/ – and found that only 3D features were significantly different between healthy controls and PD patients [8]. This paper builds upon the work of Bandini *et al.* and analyzes the motion of the lips during execution of speech and non-speech tasks from 8 PD patients and 12 healthy controls. Kinematics parameters were extracted in i) 3D using a combination of color and depth cameras, and ii) 2D using only a color camera, with the objective of evaluating if 2D information were sufficient for assessing of orofacial function in PD patients and healthy controls.

II. MATERIALS AND METHODS

A. Participants

Twenty participants were recruited for this study: eight patients with Parkinson's disease (1 female, age 66.9 ± 20.5 years), and twelve age-matched healthy controls (8 female, age 72.9 ± 12.5 years). The study was approved by the University of Toronto's Research Ethics Board. Participants signed an informed consent form according to the requirements of the Declaration of Helsinki.

DLG is supported by the Michael J. Fox Foundation for Parkinson's Research

B. Experimental setup

Participants were seated in front of an Intel RealSense D400, consisting of a registered depth and color camera pair, with a face-to-camera distance of 30-50 cm [8]. A continuous light source was placed adjacent to the camera to provide uniform illumination. Participants were asked to look at the camera and were recorded during the execution of standard neurological assessment tasks. A video composed of color (RGB) and depth information was recorded for each task. Both streams were recorded at $\sim$30 frames per second at VGA resolution (640 x 480 pixels). A total of 80 videos were included in the analysis, 48 from healthy controls, and 32 from patients. An audio file was also recorded for the duration of the procedure; a flashlight with a clicker was used to synchronize audio and video recordings.

C. Experimental procedure

Participants were asked to perform a set of speech and non-speech tasks commonly used during neurological evaluation of orofacial performance [12]. Based on previous results reported in experiments involving people with Parkinson's disease [8], amyotrophic lateral sclerosis [12], and stroke [10], four tasks were considered in this study. These task included repetition of the sentence '*Buy Bobby a Puppy*' 5 times at a comfortable rate and loudness (BBP); repetition of the syllable */pa/* as fast as possible on a single breath (PA); making a big smile showing teeth 5 times (BIGSMILE); and maintaining a neutral facial expression, eyes open, and mouth closed for 20 s (REST). Participant were encouraged to take breaks between tasks to prevent fatigue.

D. Pre-processing

Data pre-processing consisted of three steps applied in sequence: (1) Task segmentation by repetition, (2) face alignment, and (3) reconstruction of 3D information. Preprocessing was performed with a custom scrip written in Python.

1) Task segmentation: All tasks, except REST, were manually segmented into individual repetitions by a trained observer; the observer identified the beginning and end of each repetition using the audio or video recordings. The end results of this procedure were a set of videos, each containing a single repetition of the task.

2) Facial alignment: The Facial Alignment Network (FAN), an open-source, deep-neural network-based framework for automatic facial detection and localization of facial landmarks [13] was used to localize the face and the position of 68 facial landmarks in each video frame. Landmarks outlined the superior border of the brow, the free margin of the upper and lower eyelids, the nasal midline, the nasal base, the mucosal edge and vermillion-cutaneous junction of the upper and lower lips, and the lower two-thirds of the face [14].

3) Reconstruction of 3D information: Color and depth streams were aligned using the camera intrinsic information. Afterwards, the real world coordinates for each landmark were computed.

E. Orofacial features

1) Orofacial properties: Feature selection was based in previous studies with PD patients. Features were computed from five mouth properties that describe the mouth length and shape during task execution. The properties correspond to the vertical and horizontal mouth opening, the areas of the left and right side of the mouth, and the total mouth area. These properties were computed based on the position of the estimated 2D ($[x_p, y_p]$) and 3D ($[x_w, y_w, z_w]$) landmarks for each video frame as follows:

- Vertical mouth opening (TB), computed as the euclidean distance between the landmarks localized at the top and bottom vermillion borders at the midline;
- Horizontal mouth opening (WM), computed as the euclidean distance between the landmarks localized at the left and right oral commissures;
- Left mouth area ($AreaLeft$), computed as the area of a triangle formed by the landmarks localized at the top and bottom vermillion borders at the midline and the left oral commissure;
- Right mouth area ($AreaRigh$), computed as the area of a triangle formed by the landmarks localized at the top and bottom vermillion borders at the midline and the right oral commissure; and
- Overall mouth area ($Area$), computed as the sum of left and right mouth areas.

2) Normalization of 2D and 3D orofacial properties: The mean values of the mouth properties computed during the REST task were used as normalization factors for each subject. Only the middle 5 s segment of the REST task were used to compute the normalization factors. The normalization factors were computed by estimating the mouth properties for every frame during the 5 s window (150 video frames) and extracting mean values.

3) Extraction of 2D and 3D features: Thirteen orofacial kinematic features were extracted from each repetition of '*Buy Bobby a Puppy*', */pa/*, and big smile, these included:

- ΔTB, computed as the difference between the maximum and minimum values of TB
- Max velocity TB, computed as maximum value of the first derivative of TB
- Min velocity TB, computed as minimum value of the first derivative of TB
- Max acceleration TB, computed as maximum value of the second derivative of TB
- Min acceleration TB, computed as minimum value of the second derivative of TB
- ΔWM, computed as the difference between the maximum and minimum values of WM
- Max velocity WM, computed as maximum value of the first derivative of WM
- Min velocity WM, computed as minimum value of the first derivative of WM
- Max acceleration WM, computed as maximum value of the second derivative WM

TABLE I

STANDARDIZED MEAN DIFFERENCE (SMD) BETWEEN HEALTHY CONTROLS (HC) AND SUBJECTS WITH PARKINSON'S DISEASE (PD).

		3D Landmarks			2D Landmarks		
Task	Feature	HC	PD	SMD	HC	PD	SMD
BBP	ΔTB	1.7 ± 0.9	1.1 ± 0.3	**0.90**	1.2 ± 0.4	0.91 ± 0.3	**0.84**
	Max velocity TB (1/s)	36.2 ± 27.5	17.2 ± 6.1	**0.86**	19.2 ± 6.2	14.1 ± 4.7	**0.89**
	Min velocity TB (1/s)	-30.7 ± 26.1	-16.8 ± 5.5	0.67	-20.0 ± 7.8	-15.7 ± 5.6	0.69
	Max acceleration TB (1/s^2)	1836.2 ± 1605.1	868.9 ± 346.1	0.76	1041.1 ± 430.9	771.7 ± 329.8	0.68
	Min acceleration TB (1/s^2)	-2312 ± 2204.5	-880.4 ± 365.4	0.75	-1032.3 ± 383.7	-761.9 ± 306.3	0.76
	$\Delta Area$	1.7 ± 1.1	1.2 ± 0.4	0.61	1.2 ± 0.3	0.9 ± 0.3	**0.80**
	$CCC\ Area$	0.8 ± 0.2	0.7 ± 0.2	0.18	0.6 ± 0.2	0.4 ± 0.3	0.65
BIGSMILE	ΔWM	0.3 ± 0.0	0.2 ± 0.1	**0.85**	0.3 ± 0.0	0.2 ± 0.1	**0.84**
	Min velocity WM (1/s)	-3.4 ± 0.9	-2.8 ± 0.8	0.66	-3.3 ± 0.9	-2.7 ± 0.9	0.68
	$\Delta Area$	1.7 ± 0.7	1.5 ± 0.7	0.25	1.4 ± 0.5	1.0 ± 0.4	0.61
	$CCC\ Area$	0.9 ± 0.1	0.8 ± 0.2	0.55	0.8 ± 0.2	0.5 ± 0.3	**1.24**

Bold values indicate a *large* difference between groups (SMD>0.8)

- Min acceleration WM, computed as minimum value of the second derivative of WM
- Mean $Area$, computed as the mean value of $Area$
- $\Delta Area$, computed as the difference between the maximum and minimum values of $Area$
- $CCC\ Area$, computed as the concordance correlation coefficient between $AreaLeft$ and $AreaRight$

The concordance correlation coefficient measures the agreement between two signals [15], and was used as a measure of symmetry between left and right mouth movements.

F. Statistical analysis

Each task was analyzed independently. The ability of each feature to differentiate between healthy controls and PD patients was evaluated using the standardized mean difference (SMD) [16], computed as

$$SMD = \frac{\mu_1 - \mu_2}{\sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}}$$

were μ_1, μ_2, s_1, s_2, n_1, and n_2 are the mean, standard deviation, and number of elements for the feature computed for healthy controls and PD patients respectively. An SMD value lower than 0.5 indicates a *small* difference between groups, an SMD value of 0.5 or larger indicates a *medium* difference between groups, and an SMD value of 0.8 or larger indicates a *large* difference between groups [16], [17]. This analysis was performed independently for features obtained from 3D and 2D landmarks.

III. RESULTS

Table I shows the mean $\pm$ standard deviation of estimated features as well as the SMD between healthy controls (HC) and Parkinson's disease (PD) patients. The table only shows features with a *medium* ($0.5 \leq$ SMD < 0.8) or *large* (SMD ≥ 0.8) difference between groups when extracted with 3D or 2D landmarks.

These results show that only features extracted from the tasks BBP and BIGSMILE demonstrated *medium* or *large* difference between healthy controls and PD when estimated with 3D or 2D landmarks. In contrast, features extracted from the speech task PA demonstrated only *small* (SMD < 0.5) difference between groups and are not presented in the table.

A. 3D vs. 2D landmarks

Table I shows that features computed from 3D landmarks that demonstrated a *medium* or *large* difference between healthy controls and PD patients showed essentially the same behavior when computed from 2D landmarks. In contrast, area related features demonstrated *medium* or *large* difference between groups when computed from 2D landmarks and only *small* or *medium* difference between groups when computed from 3D landmarks.

B. Relevant features

1) BBP: Features extracted form 3D and 2D landmarks demonstrating a *large* difference between healthy controls and PD include the mouth vertical range of motion (ΔTB), and its maximum velocity (Max velocity TB).The overall change in the mouth area ($\Delta Area$) also demonstrated large difference between groups, but only when extracted from 2D landmarks.

2) BIGSMILE: The only feature extracted form 3D and 2D landmarks demonstrating a *large* difference between healthy controls and PD was the mouth horizontal range of motion (ΔWM). The concordance correlation coefficient between left and right mouth areas ($CCC\ Area$) also demonstrated large difference between groups, but only when extracted from 2D landmarks.

IV. DISCUSSION

A universal system for automatic assessment of PD should be based on readily available technology to reach a large number of patients despite geographical and socioeconomic limitations. Standard color cameras are available in mobile phones, tablet, and personal computers, which are readily accessible in most clinics. Thus, standard color cameras represent an ideal technology over which to develop a system

for remote, automated, and objective assessment of PD symptoms.

Previously developed systems automatically identified orofacial deficits in PD based on information provided by color and depth cameras. Depth cameras are available on commercial grade products such as Microsoft Kinect, Azure Kinect, or Intel RealSense. However, these cameras cost as much as $400USD, are not commonly available in clinics, and require an accompanying computer or laptop with specialized software for data acquisition and processing. Thus, the objective of this study was to compare orofacial kinematic features estimated during the execution of speech and non-speech tasks from information provided by standard color cameras (2D landmarks), and a combination of color and deep cameras (3D landmarks), and evaluate the ability of these features to distinguish between healthy controls and PD patients.

Our results demonstrated that 2D landmarks information was at least as successful as 3D landmark information in differentiating between healthy controls and PD patients; this observation was validated by the fact that for 3D features with SMD larger or equal than 0.5, their corresponding 2D counterpart also showed a SMD larger or equal than 0.5. By contrast, we observed that some 2D-based features demonstrated a large difference between healthy controls and PD only when extracted from 2D landmarks. We will study this observation in detail in our future work.

Finally, we observed that the most relevant features in 3D and 2D to distinguish between healthy controls and PD patients were related to the vertical (in BBP) and horizontal (in BIGSMILE) movements of the mouth. As it was expected, results indicate that PD patients have smaller range of motion and move slower than healthy subjects. These are cardinal orofacial symptoms of PD [3], that might be a consequence of the rigidity, bradykinesia, and akinesia associated with the disease [18].

Limitations

Herein, we used pre-trained models for face detection and for localization of facial landmarks. However, these types of models are well know for providing larger landmarks localization error when applied to elderly subjects and patients with neurological diseases [19], [20]. In our future work, we will re-train or fine-tune the network to improve its accuracy in our target population.

REFERENCES

[1] J. S. Schneider, S. G. Diamond, and C. H. Markham, "Deficits in orofacial sensorimotor function in parkinson's disease," *Annals of neurology*, vol. 19, no. 3, pp. 275–282, 1986.

[2] M. Bakke, S. L. Larsen, C. Lautrup, and M. Karlsborg, "Orofacial function and oral health in patients with parkinson's disease," *European journal of oral sciences*, vol. 119, no. 1, pp. 27–32, 2011.

[3] S.-M. Fereshtehnejad, Ö. Skogar, and J. Lökk, "Evolution of orofacial symptoms and disease progression in idiopathic parkinson's disease: Longitudinal data from the jönköping parkinson registry," *Parkinson's Disease*, vol. 2017, 2017.

[4] E. Kearney, R. Giles, B. Haworth, P. Faloutsos, M. Baljko, and Y. Yunusova, "Sentence-level movements in parkinson's disease: Loud, clear, and slow speech," *Journal of Speech, Language, and Hearing Research*, vol. 60, no. 12, pp. 3426–3440, 2017.

[5] M. Bologna, I. Berardelli, G. Paparella, L. Marsili, L. Ricciardi, G. Fabbrini, and A. Berardelli, "Altered kinematics of facial emotion expression and emotion recognition deficits are unrelated in parkinson's disease," *Frontiers in neurology*, vol. 7, p. 230, 2016.

[6] H. Ackermann, I. Hertrich, I. Daum, G. Scharf, and S. Spieker, "Kinematic analysis of articulatory movements in central motor disorders," *Movement disorders*, vol. 12, no. 6, pp. 1019–1027, 1997.

[7] K. Forrest and G. Weismer, "Dynamic aspects of lower lip movement in parkinsonian and neurologically normal geriatric speakers' production of stress," *Journal of Speech, Language, and Hearing Research*, vol. 38, no. 2, pp. 260–272, 1995.

[8] A. Bandini, S. Orlandi, F. Giovannelli, A. Felici, M. Cincotta, D. Clemente, P. Vanni, G. Zaccara, and C. Manfredi, "Markerless analysis of articulatory movements in patients with parkinson's disease," *Journal of Voice*, vol. 30, no. 6, pp. 766–e1, 2016.

[9] A. Bandini, A. Namasivayam, and Y. Yunusova, "Video-based tracking of jaw movements during speech: Preliminary results and future directions." in *INTERSPEECH*, 2017, pp. 689–693.

[10] A. Bandini, J. R. Green, B. Richburg, and Y. Yunusova, "Automatic detection of orofacial impairment in stroke." in *Interspeech*, 2018, pp. 1711–1715.

[11] A. Bandini, S. Orlandi, H. J. Escalante, F. Giovannelli, M. Cincotta, C. A. Reyes-Garcia, P. Vanni, G. Zaccara, and C. Manfredi, "Analysis of facial expressions in parkinson's disease through video-based automatic methods," *Journal of neuroscience methods*, vol. 281, pp. 7–20, 2017.

[12] A. Bandini, J. R. Green, B. Taati, S. Orlandi, L. Zinman, and Y. Yunusova, "Automatic detection of amyotrophic lateral sclerosis (als) from video-based analysis of facial movements: speech and non-speech tasks." in *13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, 2018, pp. 150–157.

[13] 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 *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1021–1030.

[14] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "300 faces in-the-wild challenge: The first facial landmark localization challenge," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2013, pp. 397–403.

[15] I. Lawrence and K. Lin, "A concordance correlation coefficient to evaluate reproducibility," *Biometrics*, pp. 255–268, 1989.

[16] J. Cohen, *Statistical power analysis for the behavioral sciences*. Routledge, 2013.

[17] S. S. Sawilowsky, "New effect size rules of thumb," *Journal of Modern Applied Statistical Methods*, vol. 8, no. 2, p. 26, 2009.

[18] L. M. De Lau and M. M. Breteler, "Epidemiology of parkinson's disease," *The Lancet Neurology*, vol. 5, no. 6, pp. 525–535, 2006.

[19] B. Taati, S. Zhao, A. B. Ashraf, A. Asgarian, M. E. Browne, K. M. Prkachin, A. Mihailidis, and T. Hadjistavropoulos, "Algorithmic bias in clinical populations—evaluating and improving facial analysis technology in older adults with dementia," *IEEE Access*, vol. 7, pp. 25 527–25 534, 2019.

[20] A. Asgarian, S. Zhao, A. B. Ashraf, M. E. Browne, K. M. Prkachin, A. Mihailidis, T. Hadjistavropoulos, and B. Taati, "Limitations and biases in facial landmark detection–an empirical study on older adults with dementia," *arXiv preprint arXiv:1905.07446*, 2019.