Exploring Facial Metric Normalization For Within- and Between-Subject Comparisons in a Multimodal Health Monitoring Agent

OLIVER ROESLER, HARDIK KOTHARE, WILLIAM BURKE, MICHAEL NEUMANN, JACKSON LISCOMBE, ANDREW CORNISH, DOUG HABBERSTAD, DAVID PAUTLER, DAVID SUENDERMANN-OEFT, and VIKRAM RAMANARAYANAN, Modality.AI, Inc., USA

The use of facial metrics obtained through remote web-based platforms has shown promising results for at-home assessment of facial function in multiple neurological and mental disorders. However, an important factor influencing the utility of the obtained metrics is the variability within and across participant sessions due to position and movement of the head relative to the camera. In this paper, we investigate two different facial landmark predictors in combination with four different normalization methods with respect to their effect on the utility of facial metrics obtained through a multimodal assessment platform. We analyzed 38 people with Parkinson's disease (pPD) and 22 healthy controls who were asked to complete four interactive sessions, a week apart from each other. We find that metrics extracted through MediaPipe clearly outperform metrics extracted through OpenCV and Dlib in terms of test-retest reliability and patient-control discriminability. Furthermore, our results suggest that using the inter-caruncular distance to normalize all raw visual measurements prior to metric computation is optimal for between-subject analyses, while raw measurements (without normalization) can also be used for within-subject comparisons.

CCS Concepts: • **Applied computing → Health care information systems**; **Health informatics**.

Additional Key Words and Phrases: Multimodal dialogue system, Remote patient monitoring, Normalization, Parkinson's disease

ACM Reference Format:
Oliver Roesler, Hardik Kothare, William Burke, Michael Neumann, Jackson Liscombe, Andrew Cornish, Doug Habberstad, David Pautler, David Suendermann-Oeft, and Vikram Ramanarayanan. 2022. Exploring Facial Metric Normalization For Within- and Between-Subject Comparisons in a Multimodal Health Monitoring Agent. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '22 Companion), November 7–11, 2022, Bengaluru, India.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3536220.3558071

1 INTRODUCTION

The SARS-COV-2 pandemic [17] has underscored the need for remote patient monitoring not only to improve ease and frequency of access to health care but also to enhance our understanding of the patients' conditions through the rich data captured in the natural environment of their homes to tailor their treatment [16, 20]. This could improve outcomes for individual patients while at the same time substantially decreasing healthcare costs. Previous work has shown promising results for the use of facial metrics as digital biomarkers for a variety of neurological and mental health conditions, such as Amyotrophic Lateral Sclerosis (ALS) [1], Depression [18], or Parkinson's Disease (PD) [10]. However, an important factor influencing the utility of the obtained facial metrics is the variability within and across sessions due to camera and head movement. Most previous studies (e.g. [7, 14]) tried to reduce the variability by normalizing all

Manuscript submitted to ACM

metrics through the inter-caruncular or interocular distance[1] in pixels leading to unitless metrics. However, none of the studies explained why a particular normalization method was used nor reported results for both normalized as well as unnormalized metrics to evaluate the influence of normalization on the utility of facial metrics. Furthermore, there exist, to the best of our knowledge, no studies that focus on the evaluation and comparison of different normalization approaches for different scenarios, like between-subject and within-subject analyses.

Therefore, in this paper, we take a first step towards filling this gap by investigating four different normalization methods in combination with two different facial landmark predictors regarding their effect on the utility of facial metrics obtained through a multimodal assessment platform. More specifically, we utilize the rich multimodal data collected to answer the following research questions regarding facial metrics normalization:

(1) How does normalization by inter-caruncular distance or iris diameter (with both unitless and millimeter conversions) impact:
   (a) test-retest reliability (within subject)?
   (b) effect sizes (between subjects)?
(2) How do different facial landmark predictors compare with respect to the influence of normalization on facial metric utility?
(3) Can iris diameter estimation using MediaPipe FaceMesh be used to compute accurately the inter-caruncular distance in millimeters?

To this end, 38 people with PD (pPD) and 22 controls were recruited in an ongoing study and were asked to complete four interactive sessions, a week apart from each other. For each session facial metrics were automatically extracted in real-time while participants were guided through a battery of standard tasks designed to elicit speech and facial behaviors by a virtual conversational agent.

The remainder of the paper is structured as follows: Section 2 describes the employed multimodal dialogue system. The collected data, evaluated normalization methods, and the performed analyses are presented in Sections 3 and 4. Finally, Section 5 concludes this paper.

2 SYSTEM

The virtual dialog agent, Tina, is powered by the Modality platform, a cloud-based multimodal dialogue system [23] that conducts on-demand automated screening interviews through a HIPAA-compliant, secure screening portal over smartphone app or web browser to monitor disease progression and facilitate the development of treatment plans. During the conversation, Tina engages patients in a mixture of structured speaking exercises and open-ended questions to elicit speech and facial behaviors, while she can instruct patients to complete standard survey instruments such as the Parkinson's disease questionnaire (PDQ-39), at the end of the conversation. During each call, analytics modules automatically extract a variety of audio (e.g., speaking rate, duration) and facial (e.g., range and speed of movement of lips and jaw) metrics in real-time and store them in a database together with meta-information of the interaction, like captured participant responses, call duration, or completion status [21]. This information can be accessed by clinicians during and after the interaction through an easy-to-use dashboard, which provides a high-level overview of the interaction and a detailed breakdown of individual interaction turns.

[1]The inter-caruncular or interocular distance is the distance between the inner canthi of the eyes (see Figure 1 for a visual illustration).

Table 1. Participant demographics: Age, MoCA scores, and years since diagnosis are presented as: median; mean (standard deviation).

Group	Sex	Age (years)	MoCA score	Years since diagnosis
Controls	18F/4M	64.5; 63.14 (10.92)	28; 27.55 (1.88)	-
pPD	19F/19M	70; 67.00 (9.19)	27; 26.03 (3.58)	5; 7.58 (6.17)

3 DATA

This study includes data from 60 participants (Table 1) collected between November 2020 and January 2022. All participants were recruited through the Purdue Motor Speech Lab at Purdue University and informed consent was obtained from all participants, after explaining the nature of the study and what it involved. Inclusion criteria for pPD were: between 30 and 85 years of age, a diagnosis of idiopathic PD, availability of a device with a microphone and a camera, internet access, no hearing and vision loss (self-reported) and fluency in English. Exclusion criteria were: diagnosis of a neurological disease other than PD; a history of head and neck cancer/surgery, voice disorder, pulmonary disease, smoking (in the past 5 years), more than moderate cognitive impairment as indicated by a Montreal Cognitive Assessment (MoCA) [19] score of less than 10. Controls were age- and sex-matched. Participants were asked to complete four sessions, a week apart from each other. Some participants completed fewer or more than the suggested number of sessions, resulting in a total of 265 sessions. The conversational callflow required participants to do the following speaking exercises: (a) sustained vowel (held steady /A/ , up-or-down pitch glide /i/), (b) read speech: speech intelligibility test (SIT) sentences, sentences that elicited variation in intonational prosody, rainbow passage, (c) story retells and (d) spontaneous speech (Spont) on any topic of their choice with a few topics suggested on the screen. At the end of each session, participants completed the Parkinson's Disease Questionnaire (PDQ-39) [6] and the Communicative Participation Item Bank short form (CPIB-S) [3].

3.1 Facial Metrics Extraction and Normalization

Facial metrics were calculated for each turn in three steps using two different face detectors and facial landmark predictors, which have been chosen for this study because they are commonly used, are open-source, and allow commercial use:

(1) Face detection to determine the (x, y)-coordinates of one or more faces for every input frame
   (a) using the face detector in the *dnn* module of OpenCV (`https://opencv.org/`), which uses a Single Shot Detector architecture [15].
   (b) using MediaPipe Face Detection, which is based on BlazeFace [4].
(2) Facial landmark extraction
   (a) using the Dlib facial landmark detector, which uses an ensemble of regression trees [13] to extract 68 facial landmarks according to MultiPIE [9].
   (b) using MediaPipe Face Mesh [12], which uses a residual neural network architecture to extract 478 facial landmarks[2].
(3) Facial metrics calculation, which uses 14 facial landmarks to compute metrics like the speed and acceleration of articulators (jaw, lower lip), surface area of the mouth, and eyebrow raises. The 14 landmarks used for their

[2]Originally, MediaPipe Face Mesh only extracted 468 facial landmarks, however, in 2020 it was extended with an attention mechanism by Grishchenko et al. [8] to extract 10 additional facial landmarks including eight iris landmarks.

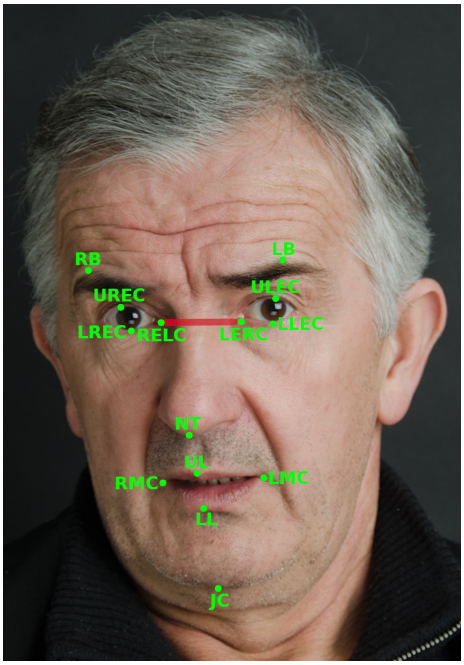

Fig. 1. Illustration of the 14 facial landmarks used to calculate the five representative facial metrics used in this study and the inter-caruncular distance (shown in red) between the inner canthi of the eyes (RELC and LERC).

calculation are illustrated in Figure 1. For ease of readability, the rest of this paper will focus on the following five metrics as a representative sample of facial metrics.

- **eyebrow_vpos_nt_max**: Maximum vertical eyebrow displacement, calculated as the difference between the vertical positions of RB and NT, and LB and NT.
- **eye_open_max**: Maximum eye opening, calculated as the Euclidean distances between UREC and LREC as well as ULEC and LLEC.
- **S_max**: Maximum total mouth surface calculated as the sum of the two triangles with the vertices RMC, UL, LL and LMC, UL, LL.
- **width_max**: Maximum width of the mouth, calculated as the Euclidean distance between RMC and LMC.
- **vLL_abs_avg**: Average speed of the lower lip (LL).

Four different normalization methods were applied to account for differences in camera distances. Table 2 provides an overview of the applied normalization methods. For the metrics obtained using OpenCV and Dlib only unitless normalization using the inter-caruncular distance could be applied (Method I) because no landmarks for the iris were available, which are required to convert pixels to millimeters. For the metrics obtained using MediaPipe (Methods II through V), unitless normalization was obtained by either dividing facial metrics in pixels by the inter-caruncular distance between the participant's eyes in pixels (Method II) or the mean horizontal iris diameter[3] in pixels (Method IV). Additionally, for the metrics extracted via MediaPipe, normalization was also done by dividing the metrics in pixels by the mean iris diameter in pixels and multiplying the result by 11.7 mm (Method V) because various studies

[3]The mean horizontal iris diameter was calculated across both eyes and all frames of a turn.

Table 2. Overview of the four investigated normalization methods applied to the metrics calculated with landmarks extracted either by Dlib or by MediaPipe Face Mesh. In the provided equations, α and β represent the unnormalized and normalized metrics, respectively. For area-based metrics, expressed in squared pixels, all terms on the right side of the equations were raised to the power of 2.

	Face detector	Landmarks predictor	Distance	Unit	Equation
I	OpenCV SSD face detector [15]	Dlib facial landmark detector [13]	Inter-caruncular distance (ICD)	px	$\beta = \frac{\alpha}{ICD}$
II	MediaPipe Face Detection [4]	MediaPipe Face Mesh [12]			
III				mm	$\beta = \frac{\alpha}{ICD} * (\frac{ICD}{ID} * 11.7\text{mm})$
IV			Iris diameter (ID)	px	$\beta = \frac{\alpha}{ID}$
V				mm	$\beta = \frac{\alpha}{ID} * 11.7\text{mm}$

have shown that the average horizontal iris diameter of the human eye is about 11.7 mm across a wide swathe of the population [2, 5, 11, 22]. Finally, normalization was done by dividing facial metrics in pixels by the inter-caruncular distance in pixels and multiplying it with the inter-caruncular distance in millimeters (Method III), which was in turn obtained by dividing the inter-caruncular distance in pixels by the iris diameter in pixels and multiplying it with the iris diameter in mm.

4 ANALYSES

4.1 Effect of normalization

This section compares the effect of the different normalization methods described in Section 3.1 regarding their influence on the test-retest reliability of facial metrics and their effect sizes. Figure 2 shows within-subject correlations between sessions for five facial metrics. The results show that metrics obtained via OpenCV and Dlib have lower test-retest reliability than metrics obtained using MediaPipe Face Detection and Media Pipe Face Mesh. Similarly, we find that the effect sizes for metrics obtained using OpenCV and Dlib are lower than for the metrics obtained via MediaPipe (Figure 3). Also, MediaPipe metrics normalized using the inter-caruncular distance outperform all other normalization methods, including no normalization, in terms of higher effect sizes, suggesting that this normalization method is optimal for between subject comparisons. The test-retest reliability of MediaPipe metrics normalized using the inter-caruncular distance is also higher than all other normalization methods. Counterintuitively, we found that the raw values without normalization also had high test-retest reliability for this dataset, in some cases higher than other normalization methods. While this could be due to the nature of our data collection, where a researcher checked that all participants adhered and compliance to task instructions, we cannot claim that this observation might generalize to all datasets. Important to note is also that only for two of the metrics the correlations are statistically significant, while for the other three metrics the correlations for most normalization methods are not significant (Figure 2). Furthermore, when looking at patients and controls separately, the latter have a higher test-retest reliability than the former and all normalization methods involving MediaPipe show statistically significant correlations for controls, while for patients less correlations are statistically significant than for all participants. It is also interesting that, for both test-retest reliability and effect sizes, millimeter metrics and unit-less metrics normalized using the iris diameter are always the same (though worse than the inter-caruncular distance), indicating that the ground truth approximation of the iris diameter in millimeters, i.e. 11.7 mm, is relatively accurate so that the results mostly depend on the accurateness of the predicted iris diameter.

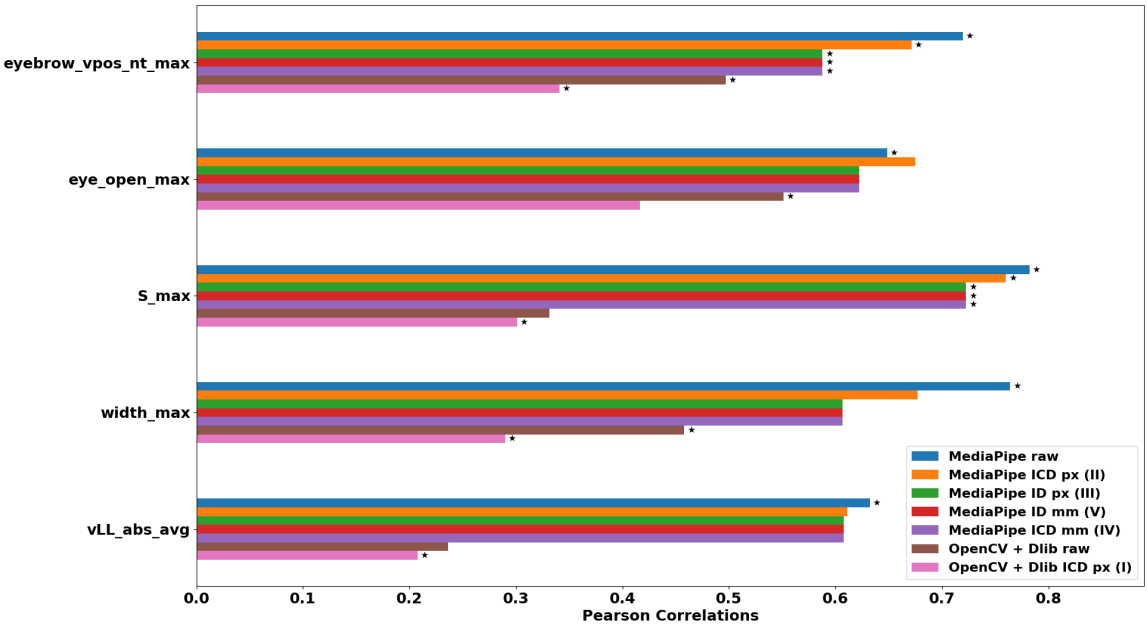

Fig. 2. Overview of test-retest reliability measured as the average Pearson's correlation coefficient across all pairs of sessions for five facial metrics that represent a representative sample of the set of extracted facial metrics with r >= 0.5 for both predictors and all four normalization methods (see Table 2). Statistically significant correlations ($p < 0.05$) are marked with a star.

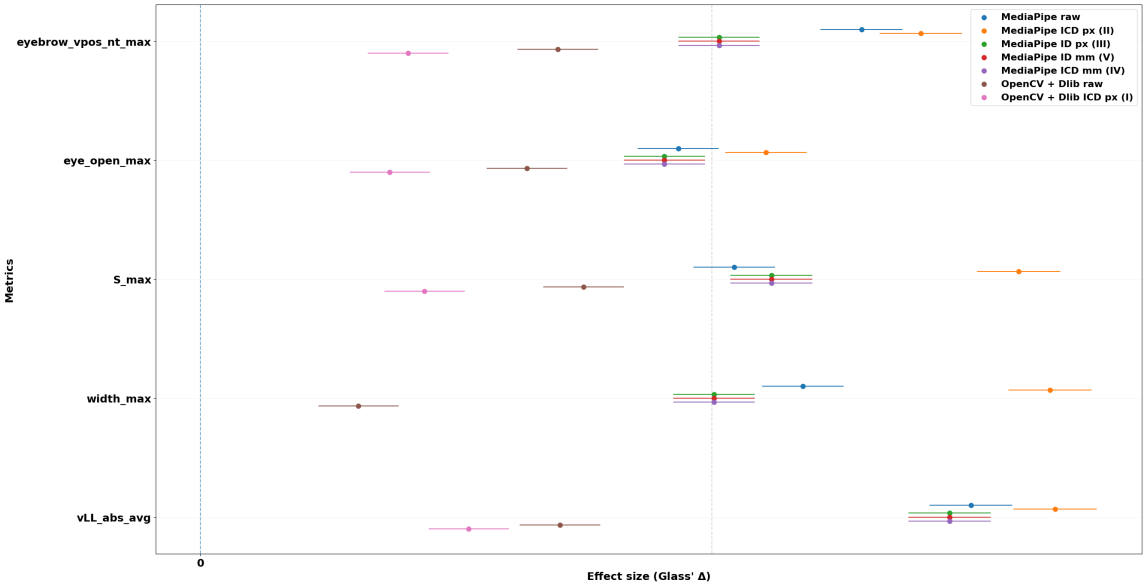

Fig. 3. Effect sizes for five facial metrics that represent a representative sample of the set of extracted facial metrics for both predictors and all four normalization methods (see Table 2). Each metric shows a statistically significant difference between controls and pPD at $\alpha = 0.05$.

Table 3. Comparison between the inter-caruncular distance in mm obtained via MediaPipe and the manually obtained ground truth.

MAE	absolute error std	mean relative error	relative error std	RMSE
3.59 mm	3.24 mm	10.01%	8.10%	4.84 mm

4.2 Accuracy of iris diameter estimation

Providing metrics in millimeters instead of pixels or without unit increases their interpretability for researchers and clinicians. However, to be useful the conversion must be accurate. Therefore, participants were asked at the beginning of each session to hold a ruler to their forehead. Afterwards, the inter-caruncular distance in millimeters was manually determined by looking at one of the corresponding video frames. These ground truth values were then compared to the millimeter values of the inter-caruncular distance calculated using the facial landmarks obtained via MediaPipe Face Mesh. The results in Table 3 show that the error between the predicted iris diameter and the ground truth is relatively high with a RMSE of 4.84 mm and a relative error of 10.01%. However, the results should be taken with a grain of salt because the ground truth values were manually obtained through visual inspection of the recorded frames and for some sessions the available frames were relatively blurry. This hypothesis is supported by the large standard deviation of 3.24 mm (8.10%), which illustrates that the accuracy of the estimated iris diameter varies strongly across sessions.

5 DISCUSSION

This study illustrates the benefits for between-subjects analyses of facial metrics normalization in comparison to the use of raw values. We also found that the accuracy of facial landmarks used to calculate facial metrics has a stronger influence on the utility of the metrics than any of the investigated normalization methods for both between- and within-subject analyses. More specifically, the obtained results show that MediaPipe Face Mesh clearly outperforms OpenCV and Dlib for all metrics and independent of the applied normalization method suggesting a higher facial landmark prediction accuracy. Normalizing Mediapipe metrics using inter-caruncular distance proved to be the best normalization method for between-subjects analysis. While it was also the best normalization method for within-subject analysis, the test-retest reliability values suggested that in our specific dataset, raw values performed as well or better. This could be used when participants adhere to instructions very well and always sit in the same position relative to the camera, as in our case. However, we cannot claim that this result generalizes to all datasets, even when there are limited controls on task performance and compliance. We will examine the generalizability of this observation in future work. Overall, the results confirm that normalization through the inter-caruncular distance in pixels, as used in many previous studies without explanation (see Section 1), provides a benefit for between-subject analyses. Furthermore, the results also show that metrics can be converted to millimeters by utilizing MediaPipe Face Mesh for iris diameter estimation with an acceptably small reduction in effect sizes and test-retest reliability, when human interpretation of metrics, e.g. by a clinician, is desired. In future work, we will compare the metrics normalization methods evaluated in this study with facial landmarks normalization, i.e. the the facial landmarks would be normalized before the metrics are computed, which could also be used for non-pixel based metrics like facial action units.

6 ACKNOWLEDGMENTS

We would like to thank our wonderful collaborators Andrew Exner, Sandy Snyder, and Jessica Huber from the Purdue Motor Speech Lab at Purdue University for many useful discussions and all their help with the recruitment of participants, data collection and hand annotation of facial distances.

REFERENCES

[1] A. Bandini, J. R. Green, B. D. Richburg, and Y. Yunusova. 2018. Automatic detection of orofacial impairment in stroke. In *Interspeech*. Hyderabad, India.

[2] M. Baumeister, E. Terzi, Y. Ekici, and T. Kohnen. 2004. Comparison of manual and automated methods to determine horizontal corneal diameter. *Journal of Cataract & Refractive Surgery* 30, 2 (2004), 374–380. https://doi.org/10.1016/j.jcrs.2003.06.004

[3] Carolyn Baylor, Kathryn Yorkston, Tanya Eadie, Jiseon Kim, Hyewon Chung, and Dagmar Amtmann. 2013. The Communicative Participation Item Bank (CPIB): Item bank calibration and development of a disorder-generic short form. (2013).

[4] V. Bazarevsky, Y. Kartynnik, A. Vakunov, K. Raveendran, and M. Grundmann. 2019. BlazeFace: Sub-millisecond Neural Face Detection on Mobile GPUs. *CoRR* abs/1907.05047 (2019). arXiv:1907.05047 http://arxiv.org/abs/1907.05047

[5] J.P.G. Bergmanson and J.G. Martinez. 2017. Size does matter: what is the corneo-limbal diameter? *Clinical and Experimental Optometry* 100, 5 (2017), 522–528. https://doi.org/10.1111/cxo.12583

[6] Donald M Bushnell and Mona L Martin. 1999. Quality of life and Parkinson's disease: translation and validation of the US Parkinson's Disease Questionnaire (PDQ-39). *Quality of Life Research* 8, 4 (1999), 345–350.

[7] Luis F. Gomez, Aythami Morales, Juan R. Orozco-Arroyave, Roberto Daza, and Julian Fierrez. 2021. Improving Parkinson Detection using Dynamic Features from Evoked Expressions in Video. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 1562–1570. https://doi.org/10.1109/CVPRW53098.2021.00172

[8] I. Grishchenko, A. Ablavatski, Y. Kartynnik, K. Raveendran, and M. Grundmann. 2020. Attention Mesh: High-fidelity Face Mesh Prediction in Real-time. *CoRR* abs/2006.10962 (2020). arXiv:2006.10962 https://arxiv.org/abs/2006.10962

[9] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker. 2010. Multi-PIE. In *Proceedings of the International Conference on Automatic Face and Gesture Recognition (FG)*, Vol. 28. 807–813.

[10] Diego L. Guarin, Aidan Dempster, Andrea Bandini, Yana Yunusova, and Babak Taati. 2020. Estimation of Orofacial Kinematics in Parkinson's Disease: Comparison of 2D and 3D Markerless Systems for Motion Tracking. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*. 540–543. https://doi.org/10.1109/FG47880.2020.00112

[11] H. Hashemi, M. Khabazkhoob, M. H. Emamian, M. Shariati, A. Yekta, and A. Fotouhi. 2015. White-to-white corneal diameter distribution in an adult population. *Journal of Current Ophthalmology* 27, 1 (2015), 21–24. https://doi.org/10.1016/j.joco.2015.09.001

[12] Y. Kartynnik, A. Ablavatski, I. Grishchenko, and M. Grundmann. 2019. Real-time Facial Surface Geometry from Monocular Video on Mobile GPUs. *CoRR* abs/1907.06724 (2019). arXiv:1907.06724 http://arxiv.org/abs/1907.06724

[13] V. Kazemi and J. Sullivan. 2014. One Millisecond Face Alignment with an Ensemble of Regression Trees. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Columbus, USA.

[14] H. Kothare, O. Roesler, W. Burke, M. Neumann, J. Liscombe, A. Exner, S. Snyder, A. Cornish, D. Habberstad, D. Pautler, D. Suendermann-Oeft, J. Huber, and V. Ramanarayanan. 2022. Speech, Facial and Fine Motor Features for Conversation-Based Remote Assessment and Monitoring of Parkinson's Disease. In *Proceedings of the 44th International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. Glasgow.

[15] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. 2016. SSD: Single Shot MultiBox Detector. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 21–37.

[16] Lakmini P Malasinghe, Naeem Ramzan, and Keshav Dahal. 2019. Remote patient monitoring: a comprehensive study. *Journal of Ambient Intelligence and Humanized Computing* 10, 1 (2019), 57–76.

[17] Francesco Motolese, Alessandro Magliozzi, Fiorella Puttini, Mariagrazia Rossi, Fioravante Capone, Keren Karlinski, Alit Stark-Inbar, Ziv Yekutieli, Vincenzo Di Lazzaro, and Massimo Marano. 2020. Parkinson's disease remote patient monitoring during the COVID-19 lockdown. *Frontiers in neurology* 11 (2020).

[18] M. Nasir, A. Jati, P. G. Shivakumar, S. N. Chakravarthula, and P. Georgiou. 2016. Multimodal and Multiresolution Depression Detection from Speech and Facial Landmark Features. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge (AVEC)*. Amsterdam, The Netherlands, 43–50.

[19] Ziad S Nasreddine, Natalie A Phillips, Valérie Bédirian, Simon Charbonneau, Victor Whitehead, Isabelle Collin, Jeffrey L Cummings, and Howard Chertkow. 2005. The Montreal Cognitive Assessment, MoCA: a brief screening tool for mild cognitive impairment. *Journal of the American Geriatrics Society* 53, 4 (2005), 695–699.

[20] Spyros Papapetropoulos, Georgia Mitsi, and Alberto J Espay. 2015. Digital health revolution: is it time for affordable remote monitoring for Parkinson's disease? *Frontiers in neurology* 6 (2015), 34.

[21] Vikram Ramanarayanan, Oliver Roesler, Michael Neumann, David Pautler, Doug Habberstad, Andrew Cornish, Hardik Kothare, Vignesh Murali, Jackson Liscombe, Dirk Schnelle-Walka, et al. 2020. Toward Remote Patient Monitoring of Speech, Video, Cognitive and Respiratory Biomarkers Using Multimodal Dialog Technology.. In *INTERSPEECH*. 492–493.

[22] F. Ruefer, A. Schroeder, and C. Erb. 2005. White-to-White Corneal Diameter. *Cornea* 24, 3 (2005), 259–261. https://doi.org/10.1097/01.ico.0000148312.01805.53

[23] D. Suendermann-Oeft, A. Robinson, A. Cornish, D. Habberstad, D. Pautler, D. Schnelle-Walka, F. Haller, J. Liscombe, M. Neumann, M. Merrill, O. Roesler, and R. Geffarth. 2019. NEMSI: A Multimodal Dialog System for Screening of Neurological or Mental Conditions. In *Proceedings of ACM International Conference on Intelligent Virtual Agents (IVA)*. Paris, France.