This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2020.3019242, IEEE Journal of Biomedical and Health Informatics

A New Dataset for Facial Motion Analysis in Individuals with Neurological Disorders

Andrea Bandini, *Member, IEEE*, Sia Rezaei, Diego Guarín, Madhura Kulkarni, Derrick Lim, Mark I. Boulos, Lorne Zinman, Yana Yunusova, and Babak Taati

***Abstract*—We present the first public dataset with videos of oro-facial gestures performed by individuals with oro-facial impairment due to neurological disorders, such as amyotrophic lateral sclerosis (ALS) and stroke. Perceptual clinical scores from trained clinicians are provided as metadata. Manual annotation of facial landmarks is also provided for a subset of over 3300 frames. Through extensive experiments with multiple facial landmark detection algorithms, including state-of-the-art convolutional neural network (CNN) models, we demonstrated the presence of bias in the landmark localization accuracy of pre-trained face alignment approaches in our participant groups. The pre-trained models produced higher errors in the two clinical groups compared to age-matched healthy control subjects. We also investigated how this bias changes when the existing models are fine-tuned using data from the target population. The release of this dataset aims to propel the development of face alignment algorithms robust to the presence of oro-facial impairment, support the automatic analysis and recognition of oro-facial gestures, enhance the automatic identification of neurological diseases, as well as the estimation of disease severity from videos and images.**

***Index Terms*—Algorithmic bias, dataset, face alignment, oro-facial impairment, amyotrophic lateral sclerosis, stroke.**

I. Introduction

MANY neurological diseases – e.g., stroke, amyotrophic lateral sclerosis (ALS), Parkinson's disease (PD), etc. – affect the oro-facial musculature with major impairments to speech, swallowing, and oro-motor abilities, as well as expression of emotions [1]–[3]. A timely and accurate assessment of oro-facial impairments can contribute to the overall disease diagnosis and lead to early interventions and improved quality of life. The objective analysis of facial kinematics can support the oro-facial structural and functional assessment as well as provide outcome measures to track treatment progress in neurological disorders [4], [5].

Currently, oro-facial assessment relies either on clinical evaluations performed by experts (i.e., cranial nerve examination) or on the use of sensor-based techniques (e.g., opto-electronic tracking methods, electromagnetic articulography). However, subjective assessments show reduced reliability [6] and sensor-based techniques require expensive instrumentation, prohibiting the translation of such technology into everyday clinical practice [7]. These drawbacks limit effective disease progression and treatment recovery monitoring.

Computer vision research can help improve clinical assessment. The study of the human face through computer vision techniques for clinical purposes has thrived over the past few years, with many applications in neurology, speech-language pathology, and psychiatry [8]–[16]. The availability of efficient and accurate face alignment approaches constitutes an important step towards the development of marker-less and intelligent tools for healthcare applications. Recent studies reported that simple and clinically interpretable measures (e.g., velocity, acceleration, range of motion) extracted from lip and jaw movements allow detecting the bulbar symptoms of ALS and oro-facial impairment in individuals with PD and post-stroke (PS) [10], [11], [17], [18]. However, since most of the available datasets used to train these algorithms do not include images of individuals with neurological disorders and oro-facial impairments, there might be a degradation of the landmark localization performance when impaired and non-standard facial movements are presented. The presence of an algorithmic bias was recently demonstrated in a number of works [19]–[21]. Specifically, state-of-the-art landmark localization performance was shown to perform worse in older adults with dementia [19], [21] as compared to cognitively intact older adults. A similar bias was reported in face alignment accuracy for individuals with facial palsy [15], [20]. Retraining or fine-tuning of the models with data from the clinical group of interest can help reduce this bias [15], [20], [21].

Since the major obstacle for obtaining good landmark localization performance in clinical populations is the limited availability of annotated training data, we released the first dataset of facial videos of individuals with ALS and PS accompanied

This work was supported in part by: AGE-WELL NCE Inc., a member of the Networks of Centres of Excellence program; Canadian Partnership for Stroke Recovery (Heart and Stroke Foundation); NIH (grants R01DC009890 and R01DC013547); National Sciences and Engineering Research Council (NSERC, Canada); and Michael J. Fox Foundation for Parkinson's Research. *Yana Yunusova and Babak Taati contributed equally to this work. Corresponding author: Yana Yunusova (yana.yunusova@utoronto.ca).*

AB, SR, and DG are with KITE – Toronto Rehab – University Health Network (UHN), Toronto, ON, Canada (e-mail: andrea.bandini@uhn.ca).

MK is with Hurvitz Brain Sciences Program, Sunnybrook Research Institute (SRI), Toronto, ON, Canada.

DL was with the Department of Speech Language Pathology, Rehabilitation Sciences Institute, University of Toronto, Toronto, ON, Canada.

MIB is with Sunnybrook Health Sciences Centre and the Department of Medicine, Division of Neurology, University of Toronto, Toronto, ON, Canada.

LZ is with Hurvitz Brain Sciences Program, SRI; the L.C. Campbell Cognitive Neurology Research Unit, SRI; and Department of Medicine, Division of Neurology, University of Toronto, Toronto, ON, Canada.

YY is with Rehabilitation Sciences Institute - Department of Speech Language Pathology, University of Toronto; Hurvitz Brain Sciences Program, SRI; and KITE – Toronto Rehab – UHN, Toronto, ON, Canada (e-mail: yana.yunusova@utoronto.ca).

BT is with KITE – Toronto Rehab – UHN; the Institute of Biomedical Engineering, University of Toronto; the Department of Computer Science, University of Toronto; and the Vector Institute for Artificial Intelligence, Toronto, ON, Canada (e-mail: babak.taati@uhn.ca).

by the clinical scores and the ground truth location of 68 facial landmarks on more than 3300 representative image frames. The availability of this data aims to foster the development of novel and robust approaches for face alignment and oro-facial assessments that can be used to track and analyze facial movements in these clinical populations. This dataset is expected to facilitate further development of state-of-the-art automatic assessments of neurological disorders. Moreover, another aim of this study was to estimate the extent of face alignment bias in individuals with neurological diseases affecting the oro-facial function, such as ALS and stroke. To detect the presence of bias, we evaluated multiple pre-trained face alignment models across the range of disease severity. Further experiments were also conducted by fine-tuning the best pre-trained model on subsets of patients' data to evaluate how the use of frames from target populations might help alleviate this issue. To summarize, the main contributions of this paper included:

- We release *Toronto NeuroFace*[1], the first dataset of 261 videos, clinical scores per video, and more than 3300 annotated frames of faces from individuals with ALS and PS as well as age-matched healthy control (HC) subjects, while performing oro-facial tasks typical of the clinical assessment.
- For the first time, we analyzed the problem of face alignment bias in neurological disorders affecting the oro-facial functions, such as stroke and ALS.
- Finally, we reported results of experiments linking the face alignment error and clinical disease metrics using pre-trained and fine-tuned face alignment algorithms. These experiments allowed us to quantitatively demonstrate the benefits of using data from target populations when developing specific face alignment applications.

The remainder of the paper is organized as follows: Section II summarizes the existing datasets for face alignment and the application of face alignment algorithms in clinical conditions; Section III describes in detail the data collection and pre-processing steps involved in building the *Toronto NeuroFace* dataset; Section IV provides a review of the face alignment algorithms approaches for our experiments; Sections V and VI describe the experiments and results performed with the pre-trained and fine-tuned face alignment models, respectively; and, finally, Sections VII and VIII conclude with a discussion of the results.

II. Related work

In this section, we summarize some of the recent advancements on video-based analysis of facial movements and expressions for clinical applications, with an overview of the existing datasets.

A. Automatic face analysis for clinical applications

The analysis of facial movements and expressions for healthcare applications is a fast-growing area of research, which has seen important advancements over the recent years [22]. Some of the applications are: the recognition of pain from facial images and videos [8], [9], [23]; the automatic analysis of the oro-facial dynamics in patients with neurological disorders (e.g., PD, stroke, ALS, Alzheimer's disease – AD, etc.) [10]–[12], [17], [18], [24]; and the automatic detection of symptoms related to psychiatric conditions, such as depression and schizophrenia [13], [16], [25]. Regardless of the specific condition, the overall aim is to provide accurate, objective, and standardized information to clinicians associated with facial kinematics and dynamics, in order to improve the current assessment practices and evaluate treatment effects.

In many cases, facial landmark detection is used as the basis of the processing pipeline, in order to extract robust spatio-temporal features of gestures and expressions that, in turn, can be used to infer the clinical condition of interest [8], [9], [11], [12], [25]. Among the face alignment approaches, the most widely used in this field are: active appearance models (AAM) [26], [27], supervised descent method (SDM) [28], and ensemble of regression trees (ERT) [29]. Recently, state-of-the-art deep-learning-based approaches, such as the face alignment network (FAN) [30], have been applied in patients with dementia and facial paralysis [14], [15], [19], [21], demonstrating higher localization accuracy than traditional face alignment approaches.

Other authors [13], [24] did not use facial landmark representations, but relied on deep-learned features to study the facial dynamics. Wang *et al.* [24] implemented different deep-learning architectures (3DCNN and multi-stream CNN) to extract spatio-temporal features from the whole face region, in order to classify different facial activities in patients with AD. Another approach [13] implemented a VGG16 [31] to detect facial action units (AUs) from specific face areas and used the AUs as low level representation for estimating schizophrenia severity. However, when the goal is the analysis of facial kinematics in neurological conditions that affect gestures and movements (i.e., stroke, PD, ALS, etc.), the facial representation via landmark detection would be preferred, since it allows the extraction of clinically interpretable outcome measures that can be related to the presence and severity of symptoms [11].

B. Existing datasets

To further improve the performance of automatic assessment systems and promote their translation into clinical practice, large public datasets with facial videos, images, and clinical metadata (e.g., diagnosis, clinical scores, etc.) are needed. Not only will the availability of this data promote the development of accurate approaches, but it will also unify the efforts made by different researchers towards solving problems in the clinical domain. Although many face alignment datasets have been published in the past 10 years [32]–[36], only a few were developed and published for healthcare applications (e.g. pain [37], [38] and facial paralysis [20], [39]).

To the best of our knowledge, none of the existing datasets include facial images and videos of individuals with oro-facial impairment due to disorders of the nervous system (such as stroke and ALS) accompanied by the ground truth facial

[1]Access to the *Toronto NeuroFace* dataset can be requested at slp.utoronto.ca/faculty/yana-yunusova/speech-production-lab/datasets/

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2020.3019242, IEEE Journal of Biomedical and Health Informatics

landmarks and clinical metadata at the same time. The lack of training data from specific clinical conditions might cause a bias in the face alignment performance [14], [19], [21], similar to what happens with race and sex biases in face recognition models [40]–[42].

III. Dataset description

In this section we provide details about participants, data collection procedures, clinical assessment of the recorded videos, and manual annotation of facial landmarks conducted on a subset of frames.

A. Participants

Thirty-six participants were recruited for this study: 11 patients with ALS (4 male, 7 female), 14 patients PS (10 male, 4 female), and 11 HC subjects (7 male, 4 female). All participants were cognitively unimpaired (Montreal Cognitive Assessment score $\geq$ 26) [43] and passed a hearing screening. Patients with ALS were diagnosed according to the El Escorial Criteria for the World Federation of Neurology [44]. Nine participants had spinal symptoms at onset, whereas two participants presented bulbar onset ALS. The ALS severity with respect to the effect on daily function was evaluated using the ALS Functional Rating Scale – Revised (ALSFRS-R) [45]. The demographic and clinical summary for the participants is reported in Table I. The study was approved by the Research Ethics Boards at the Sunnybrook Research Institute and UHN: Toronto Rehabilitation Institute. All participants signed informed consent according to the requirements of the Declaration of Helsinki, allowing inclusion into a shareable database.

	Age (years)	Duration (months)	ALSFRS-R
HC	63.2 $\pm$ 14.3	–	–
ALS	61.5 $\pm$ 8.0	49.6 $\pm$ 31.6	34.8 $\pm$ 5.0
PS	64.7 $\pm$ 14.7	19.4 $\pm$ 34.2	–

TABLE I
Demographic and clinical information for the three participant groups. Duration: months from the date of symptom onset (ALS) or from stroke (PS). (ALS: Amyotrophic Lateral Sclerosis; PS: Post-Stroke; HC: Healthy Control; ALSFRS-R: ALS Functional Rating Scale - Revised).

B. Tasks and experimental setup

Each subject was asked to perform a set of speech and non-speech tasks commonly used during a clinical oro-facial examination [46], [47]. They included: 10 repetitions of the sentence "Buy Bobby a Puppy" at a comfortable speaking rate and loudness (*BBP*); repetitions of the syllable /pa/ as fast as possible in a single breath (*PA*); repetitions of the syllables /pataka/ as fast as possible in a single breath (*PATAKA*); puckering of the lips (e.g., pretend to blow a candle 5 times and pretend to kiss a baby 5 times - *BLOW* and *KISS*); maximum opening of the jaw 5 times (*OPEN*); pretending to smile with tight lips 5 times (*SPREAD*); making a big smile 5 times (*BIGSMILE*); and raising the eyebrows 5 times (*BROW*).

Participants' faces were video-recorded using the Intel® RealSense™ SR300 camera. During the tasks, participants were seated in front of the camera, with a face-camera distance between 30 and 60 cm. A continuous light source was placed behind the SR300 to illuminate the face uniformly. For each task we collected a separate video recording composed of a pair of color (RGB) and depth videos. Experiments and results reported in this paper only consider the color videos, but both video modalities are released in the dataset. Both streams were stored at approximately 50 frames per second and 640 $\times$ 480 pixels of image resolution. A total of 261 video recordings were included in the dataset: 80 from HC subjects, 76 from patients with ALS, and 105 from patients PS.

C. Clinical oro-motor examination

Two trained speech-language pathologists watched the video recordings and rated the above tasks based on: symmetry, range of motion (ROM), speed, variability, and fatigue of facial movements. They judged each of the above aspects on a 5-point Likert scale with 1 indicating normal function and 5 indicating severe dysfunction. For each video a total score was also computed as sum of the 5 sub-scores. The average scores between the two raters are reported in Table II. The inter-rater agreement was found to be fair to moderate according to the weighted Cohen's kappa statistic (κ). This is in line with results previously reported in the literature [48]. Fair to moderate inter-rater agreement reflects the subjective nature of the clinical scores and is one of the motivating factors towards developing vision-based objective assessment systems. The average scores between the two raters were used in all the experiments reported in this paper. Kruskal-Wallis test showed statistically significant differences in the clinical measures among the 3 groups. A post-hoc Wilcoxon rank-sum test showed a small yet statistically significant increase of the scores in both ALS and PS as compared to HC subjects (see Table II). These results indicate that impairments were present in the two clinical groups when compared to HC subjects. In the majority of participants, these impairments were mild to moderate in their severity.

D. Manual annotation of face landmarks

A set of 3306 frames (1015 HC, 920 ALS, and 1371 PS) were extracted from the above videos and considered for the experiments. On these frames, the ground truth positions of 68 facial landmarks were annotated following the Multi-PIE 2D configuration [33]. For each non-speech task, we considered 3 frames per repetition: 1) beginning of the gesture (i.e., rest position); 2) peak of the gesture (e.g., maximum jaw opening, maximum lip puckering or spread, etc.); and the midpoint between 1 and 2. For the speech tasks, the selection of frames was carried out based on the visemes: 5 frames for each *BBP* repetition (/b/ of *Buy*, /o/ of *Bobby*, /a/ between *Bobby* and *puppy*, /p/ and /y/ of *puppy*); 3 frames for *PA* (maximum lip compression of /p/, maximum lip opening of /a/ and midpoint between /p/ and /a/); and 3 frames for *PATAKA* (maximum lip compression of /p/, maximum opening after /p/, and midpoint between /pa/ and /ta/). These criteria were adopted to cover

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2020.3019242, IEEE Journal of Biomedical and Health Informatics

	Symmetry	ROM	Speed	Variability	Fatigue	Total
HC	1.38 ± 0.22	1.16 ± 0.12	1.17 ± 0.12	1.19 ± 0.16	1.04 ± 0.05	5.39 ± 0.48
ALS	1.73 ± 0.22**	1.60 ± 0.68*	1.50 ± 0.46*	1.65 ± 0.45***	1.50 ± 0.18***	7.11 ± 1.05***
PS	2.26 ± 0.71**	1.86 ± 0.38***	1.63 ± 0.45*	1.78 ± 0.57**	1.23 ± 0.18**	7.98 ± 1.33***
KW	$H(2) = 14.40$, $p < .001$	$H(2) = 15.93$, $p < .001$	$H(2) = 8.03$, $p = .018$	$H(2) = 12.21$, $p = .002$	$H(2) = 22.66$, $p < .001$	$H(2) = 21.50$, $p < .001$
κ	0.57	0.59	0.41	0.61	0.33	

TABLE II

MEAN VALUE AND STANDARD DEVIATION OF CLINICAL SCORES. POST-HOC WILCOXON RANK-SUM TESTS BETWEEN HC AND ALS/ PS SUBJECTS: $^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$. (*KW* = KRUSKAL-WALLIS TEST; κ = WEIGHTED COHEN'S KAPPA STATISTIC TO MEASURE THE INTER-RATER AGREEMENT FOR EACH SCORE).

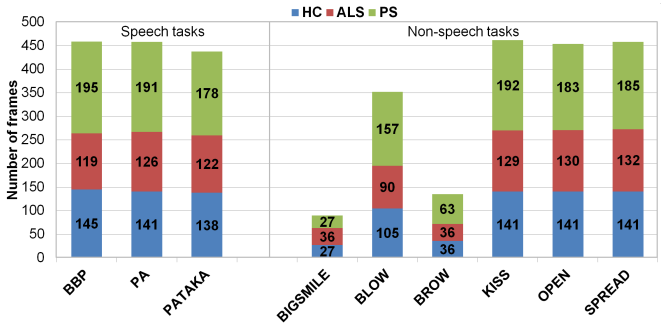

Fig. 1. Distribution of frames per task.

a wide range of facial gestures and movements required to perform the above tasks. Figure 1 shows the distribution of frames for each task and group.

A second rater, blinded to the first rater's annotation, marked the 68 facial landmarks on a subset of 515 frames (15.6 % of the annotated frames). To measure the inter-rater agreement, the point-to-point Euclidean distance normalized by the diagonal of the face bounding box was computed (nRMSE). The face bounding box was obtained using the maximum and minimum coordinates of the annotated landmarks. For all the frames annotated by the two raters, the nRMSE was lower than 5% (90.1% was below 2%), with an average nRMSE of $1.36 \pm 0.46\%$.

IV. METHODS

In this section we describe the face alignment algorithms tested on the *Toronto NeuroFace* dataset. We also provide details about the pre-training of these algorithms as well as metrics to evaluate the localization performance on the 3 groups of interest.

A. Face alignment

The detection of facial landmarks is composed of two steps: 1) face detection – to find a region of interest (ROI) within the image where a face might be located; and 2) face alignment – to locate the facial landmarks within the ROI. Since our aim was to estimate the extent of face alignment bias in individuals with neurological disorders, we used the ground truth bounding box – obtained using the ground truth landmarks – in all the experiments.

Five face alignment approaches were implemented: AAM [26], [27], constrained local model (CLM) [49], ERT [29], SDM [28], and FAN [30].

1) Generative methods – AAM: AAM is a linear statistical model of the shape and appearance of the face. It can generate several instances of shape and appearance models by varying a small number of parameters. To fit an AAM, the shape and appearance parameters are estimated to generate a model instance that best fits the test face [27]. AAM is a well-known early generative method used for face alignment [50].

2) Discriminative methods – CLM, ERT, and SDM: Unlike AAM, discriminative methods learn a set of discriminative functions to directly infer the landmark position from the facial appearance. CLM is a part-based approach and learns an independent local appearance model for each face point [49]. Geometrical constraints are then imposed using a shape model over the local appearance models. CLM is considered more robust to partial occlusions and lighting changes than AAM [50]. ERT and SDM are part of a family of discriminative approaches called cascaded regression methods. These algorithms learn a regression function to estimate the shape of the face step-by-step. Starting from an initial shape (e.g., average shape), they sequentially refine it through trained regressors. The shape increment is regressed using shape-indexed features (i.e., features extracted in the current shape estimate). The main difference between ERT and SDM is in the nature of the regression function used: SDM uses a linear regression to estimate the shape updates starting from scale-invariant feature transform (SIFT) features, whereas ERT employs tree-based regression [28], [29], [50]. Speed and accuracy made the cascaded regression methods state-of-the-art in face alignment before the advent of deep learning.

3) Deep learning methods – FAN: The FAN is a deep-learning approach for face alignment based on a stack of four hourglass networks [51]. This network architecture, originally proposed for human-pose estimation, was re-adapted for solving face alignment problems, by replacing the bottleneck block of each hourglass with the hierarchical, parallel and multi-scale block proposed in [52]. The FAN estimates 68 facial landmarks via heatmap regression from the RGB input images, and it showed state-of-the-art performance on most available face alignment datasets [30]. This architecture has also been generalized to solve 3D face alignment problems (i.e., 3D-FAN). However, considering that our dataset was collected using frontal face positions, we implemented only the 2D

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2020.3019242, IEEE Journal of Biomedical and Health Informatics

Method	Overall nRMSE (%)
AAM	2.29 ± 0.99
CLM	2.97 ± 1.37
ERT	2.02 ± 0.56
FAN	**1.80 ± 0.34**
SDM	2.20 ± 0.62

TABLE III
OVERALL NRMSE FOR EACH PRE-TRAINED MODEL. THE LOWEST NRMSE IS HIGHLIGHTED IN BOLD.

version (i.e., 2D-FAN).

B. Pre-training and error metrics

The Menpo implementations of AAM, CLM, ERT, and SDM were used [53]. These four algorithms were trained on the 300-W training set (~4000 images) [35], [54], [55], a widely-used dataset for 2D face alignment. For the FAN, we used the pre-trained 2D-FAN model[2] trained on the 300W-LP-2D dataset (~60k images), which was obtained by extending the 300W with synthetically generated images [30], [56].

Landmark localization performance between HC subjects and ALS/ PS patients was compared in terms nRMSE between the estimated landmarks and the ground truth annotations. For each frame, the nRMSE was calculated as the point-to-point Euclidean distance normalized by the diagonal of the bounding box [30]. Comparison among the models and groups was also performed by computing the percentage of frames with an error lower than a pre-defined threshold.

V. EXPERIMENTS WITH PRE-TRAINED MODELS

In this section, we analyze landmark localization errors to investigate: 1) the existence of bias in performance across the different groups, and 2) the relationship between the localization error and disease severity.

A. Overall error performance

The mean and standard deviation of the nRMSE values are reported in Table III. These values were obtained by running the pre-trained models on the whole set of 3306 annotated frames. A non-parametric Friedman test was conducted to test for differences between the errors obtained with the five models. This test was preferred to a one-way ANOVA with repeated measures, since the data was not normally distributed. Friedman's test showed a significant difference among the five groups (p <0.001). A Tukey's honestly significant difference test for multiple comparisons indicated significant differences among all pairs. Thus, the lowest landmark localization error was the one produced by FAN, followed by ERT. The higher localization accuracy of FAN can also be seen from the convergence curves of Figure 2, where the nRMSE for FAN was lower than 3% in 99.97% of the frames, versus 95.74% of ERT, 92.98% of SDM, 87.39% of AAM, and 68.51% of CLM.

[2]The 2D-FAN model used for these tests was downloaded from https://github.com/1adrianb/2D-and-3D-face-alignment

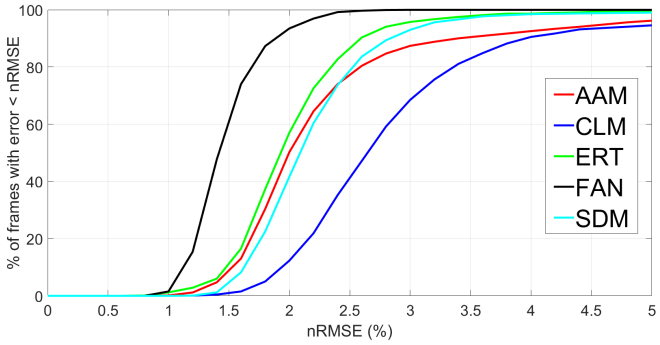

Fig. 2. Convergence curves for the pre-trained face alignment models, showing the nRMSE (%) vs. the percentage of frames with landmark localization error lower than the corresponding nRMSE threshold.

B. Error analysis across groups

To analyze the presence of a bias in the landmark localization performance, for each model we computed the nRMSE obtained on ALS and PS patients and compared it with the nRMSE obtained in HC subjects. Mean values and standard deviations of the nRMSE obtained on HC subjects (1015 frames), patients with ALS (920 frames), and patients PS (1371 frames) are reported in Table IV. Kruskall-Wallis test showed significant differences among the three groups for all the approaches. Post-hoc Wilcoxon rank-sum tests showed that the nRMSE in the two clinical groups of interest (ALS and PS) was significantly higher than in the HC group. Thus, although FAN showed excellent overall results, there was a bias in its face alignment performance.

C. Error analysis with respect to disease severity

To investigate the relationship between nRMSE and disease severity, we computed the Spearman's correlation coefficients between the average nRMSE obtained on each video and the corresponding clinical scores, averaged between the two clinician raters. Table V shows the results for the Symmetry and ROM scores; no significant correlations were found for the other aspects of the perceptual assessment. A weak, yet significant, positive correlation was found with the symmetry score in the PS videos for FAN, ERT, and SDM (see Table V), suggesting that the nRMSE can increase with the severity of the facial asymmetry. Moreover, significant correlations (weak negative) were found between the nRMSE and the ROM score in both ALS and PS participants. In this case, the negative correlation denotes lower errors in individuals with higher impairment severity, namely reduced oro-facial movements.

VI. EXPERIMENTS WITH FINE-TUNED FAN

In this section, we report on how face alignment accuracy and clinical bias changed when data from the *Toronto NeuroFace* dataset were used for fine-tuning a face alignment algorithm. We investigated the case of FAN, since the results from Section V showed its higher localization accuracy for this dataset. Specifically, we conducted two experiments: 1) fine-tuning the FAN separately within each group (HC, ALS, and

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2020.3019242, IEEE Journal of Biomedical and Health Informatics

	HC	ALS	PS	Kruskal-Wallis test
AAM	2.27 ± 0.97	2.18 ± 0.82	2.39 ± 1.11***	$H(2) = 12.07$, $p = .002$
CLM	2.78 ± 1.22	2.75 ± 0.94	3.25 ± 1.64***	$H(2) = 85.11$, $p < .001$
ERT	1.85 ± 0.46	1.92 ± 0.46***	2.20± 0.62***	$H(2) = 320.06$, $p < .001$
FAN	1.66 ± 0.22	1.87 ± 0.38***	1.86± 0.36***	$H(2) = 305.39$, $p < .001$
SDM	2.09 ± 0.51	2.12 ± 0.48*	2.33 ± 0.73***	$H(2) = 108.76$, $p < .001$

TABLE IV
NRMSE OBTAINED IN THE THREE GROUPS ANALYZED. POST-HOC WILCOXON RANK-SUM TESTS BETWEEN HC AND ALS/ PS SUBJECTS: * $p < .05$; ** $p < .01$; *** $p < .001$.

		ALS		PS	
		Symm.	ROM	Symm.	ROM
AAM	ρ	0.038	**-0.23**	0.10	**-0.21**
	p val	0.74	**0.04**	0.30	**0.03**
CLM	ρ	-0.12	-0.02	0.15	**-0.22**
	p val	0.30	0.89	0.11	**0.03**
ERT	ρ	-0.09	**-0.31**	**0.23**	**-0.29**
	p val	0.44	**0.007**	**0.02**	**0.003**
FAN	ρ	-0.12	0.12	**0.23**	0.01
	p val	0.31	0.32	**0.02**	0.91
SDM	ρ	-0.17	-0.12	**0.27**	-0.19
	p val	0.14	0.29	**0.006**	0.05

TABLE V
CORRELATION BETWEEN ERROR AND ORO-FACIAL IMPAIRMENT SEVERITY (SYMMETRY AND ROM). SIGNIFICANT CORRELATIONS ARE REPORTED IN BOLD.

PS) using a leave-one-subject-out cross-validation (LOSO); and 2) fine-tuning the FAN with data from one group and testing it on the other two groups (leave-two-groups-out cross-validation – LTGO). These tests were conducted to investigate how different combinations of data for fine-tuning affect the clinical bias detected of the pre-trained model.

For all tests, the final hourglass model of the pre-trained FAN was fine-tuned for 50 epochs using RMSProp optimizer. The learning rate was initialized to 1e-5 and was decayed by a factor of 0.5 every 10 epochs. Similar to the previous section, we used the nRMSE to compare the landmark localization error among the three groups and its correlation with the clinical perceptual evaluation.

A. Fine-tuned error across groups

The nRMSE values for the fine-tuned FAN are shown in Tables VI and VII.

1) LOSO results: Fine-tuning the FAN with data from the target group lowered the nRMSE in patients with ALS and PS (see Figure 3) and the nRMSE obtained for these two groups was lower than the error obtained in HC subjects with the pre-trained model (see Table IV). However, fine-tuning the FAN made the nRMSE decrease in HC subjects too, with values always significantly lower than the fine-tuned nRMSE of ALS and PS patients. Thus, despite the improved landmark localization accuracy, a clinical bias still remains.

To gain more insight into the effect of fine-tuning the FAN, we quantified the reduction of error as the difference between the pre-trained and fine-tuned nRMSE (ΔE). In the LOSO test, this reduction was slightly larger in patients with ALS and PS than HC subjects (see Table VI).

2) LTGO results: Even when the FAN was fine-tuned with data from one group and tested on the other two groups, the error in HC subjects remained lower than patients with ALS and PS. Moreover, looking at the average ΔE values reported in Table VII, we can observe the following trends:

- Fine-tuning using HC data caused a decrease of nRMSE in ALS and PS groups similar to ΔE obtained in HC subjects in the LOSO test.
- Fine-tuning using ALS data caused a decrease of nRMSE in HC and PS groups lower than ΔE obtained in patients with ALS in the LOSO test.
- Fine-tuning using PS data caused a decrease of nRMSE in HC and ALS groups lower than ΔE obtained in patients PS in the LOSO test.

These results suggested that, at least in the two clinical groups, data from the same population were needed when fine-tuning the network.

B. Fine-tuned error vs. disease severity

To further explore how the algorithmic bias changed after fine-tuning the FAN, we computed the Spearman's correlation coefficient between the fine-tuned nRMSE (average value for each video) and the corresponding clinical score (average between the two raters). Previously, the nRMSE obtained with pre-trained FAN showed a positive correlation with the symmetry score in individuals in the PS group ($\rho = 0.23$, p = 0.017, Table V). After fine-tuning the FAN on the PS data (LOSO test), this correlation decreased and was no longer statistically significant ($\rho = 0.10$, p = 0.30). A smaller decrease of correlation was obtained when the FAN was fine-tuned using data from patients with ALS ($\rho = 0.19$, p = 0.06) and from HC subjects ($\rho = 0.17$, p = 0.08). This result further confirmed that fine-tuning the FAN with data from the population of interests may have important benefits in reducing the clinical bias due to the presence of neurological diseases and oro-facial impairment.

VII. DISCUSSION

In this paper, we proposed and described a novel dataset and baseline results for facial landmark localization with state-of-the-art face alignment models in patients with ALS and PS. To the best of our knowledge, this is the first dataset that includes videos and images of facial gestures captured from individuals with these conditions alongside relevant clinical scores. The dataset is intended to be made available to the research community to foster future release of similar datasets

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2020.3019242, IEEE Journal of Biomedical and Health Informatics

	HC	ALS	PS	Kruskal-Wallis test
nRMSE	1.34 ± 0.24	1.53 ± 0.33***	1.53 ± 0.33***	$H(2) = 274.24$, $p < .001$
ΔE	0.31 ± 0.21	0.34 ± 0.21**	0.33 ± 0.19*	$H(2) = 9.8$, $p = .007$

TABLE VI
RESULTS OBTAINED BY FINE-TUNING FAN WITHIN EACH GROUP (LOSO) . POST-HOC WILCOXON RANK-SUM TESTS BETWEEN HC AND ALS/ PS SUBJECTS: * $p < .05$; ** $p < .01$; *** $p < .001$.

Fine-tuning on		HC	ALS	PS	Wilcoxon rank-sum test
HC group	nRMSE	–	1.57 ± 0.34	1.55 ± 0.32	Z = -0.06, $p = 0.94$
	ΔE	–	0.30 ± 0.23	0.31 ± 0.18	Z = -0.28, $p = 0.78$
ALS group	nRMSE	1.39 ± 0.22	–	1.55 ± 0.32	Z = -14.47, $p < .001$
	ΔE	0.27 ± 0.20	–	0.31 ± 0.18	Z = -5.04, $p < .001$
PS group	nRMSE	1.35 ± 0.22	1.58 ± 0.33	–	Z = -16.18, $p < .001$
	ΔE	0.31 ± 0.22	0.30 ± 0.26	–	Z = 1.22, $p = 0.22$

TABLE VII
RESULTS OBTAINED BY FINE-TUNING THE FAN WITH DATA FROM ONE GROUP AND TEST IT ON THE OTHER TWO GROUPS (LTGO).

and the development of novel face alignment approaches robust to the presence of oro-facial impairments. This dataset will facilitate the development of novel and intelligent systems for the automatic assessment of motor speech disorders and oro-facial impairments. In addition to landmark localization, the availability of rich metadata such as the diagnosis, clinical perceptual assessment, and type of oro-facial gesture will allow researchers to use this dataset for multiple purposes, including automatic classification of neurological diseases, estimation of clinical scores, and analysis of facial gestures in clinical populations.

In this work, we also demonstrated that even the presence of mild to moderate oro-facial impairment can cause a bias in the face alignment accuracy when the algorithms are not trained with data from the target populations. This bias translated to higher landmark localization errors in individuals with ALS and PS, and there was a statistically significant positive correlation between the nRMSE and the severity of facial asymmetry in patients PS. These results added further evidence to the presence of a bias in the face alignment accuracy in clinical groups, as recently demonstrated in [15], [19]–[21].

A comparison of our results with those in [21] reveals much smaller nRMSE values obtained with our data. This difference can be explained by two main factors. First, we used the ground truth bounding boxes as the face detector (i.e., ideal case), since our aim was to investigate the performance of the face alignment step exclusively. Secondly, our recording setting was highly controlled and standardized, with uniform

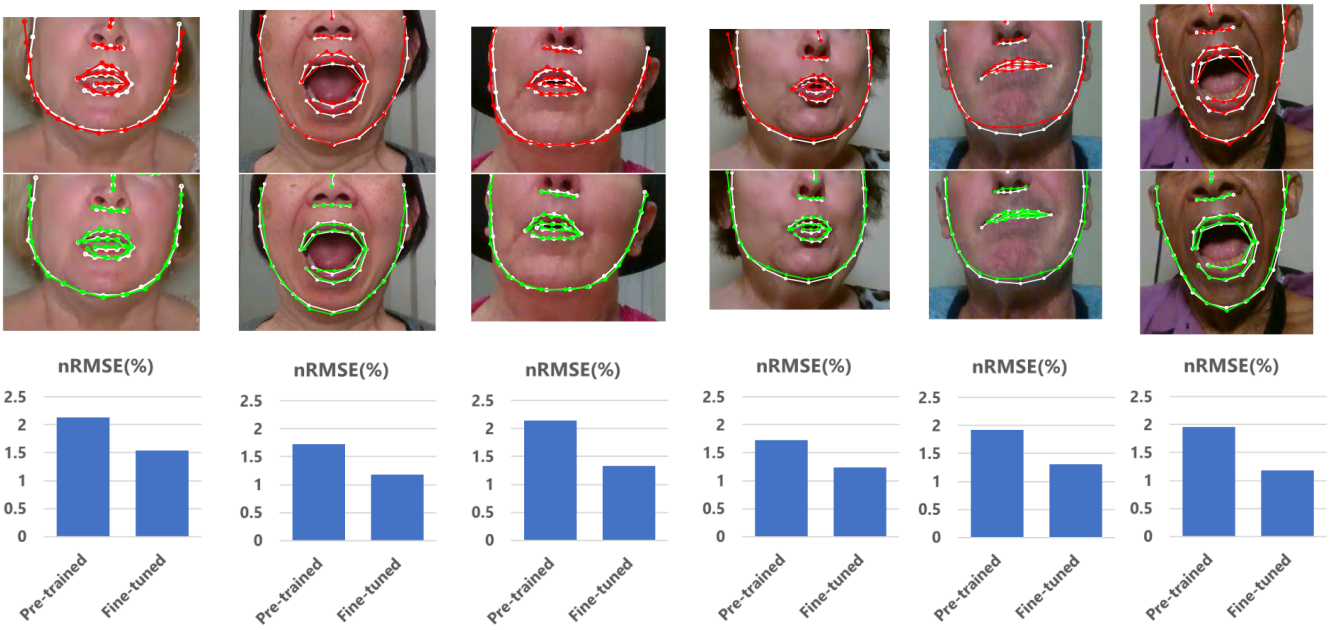

Fig. 3. Comparison between pre-trained FAN (top row) and fine-tuned FAN (middle row). Bar plots (bottom row) show the nRMSE values corresponding to the above sample frames. In these examples, we show how the fine-tuning can improve the landmark localization accuracy of facial contour and mouth regions. White: ground truth landmarks; Red: facial landmarks obtained with pre-trained FAN; Green; landmarks obtained with fine-tuned FAN.

face illumination, short and consistent camera-face distance, and frontal face recordings. This standardization helped improve the quality of the video recording, which facilitated the performance of the face alignment algorithms (see Figure 3). Nevertheless, the pre-trained models were not robust enough to the presence of the oro-facial impairments. Future studies will be devoted to investigate the face alignment bias in uncontrolled situations, such as video-recordings collected in home environments, since one of the end goals of developing intelligent systems for oro-facial assessment is to design automated tools for monitoring patients remotely.

As expected, fine-tuning the FAN on data from the *Toronto NeuroFace* dataset improved the landmark localization accuracy. However, despite the improved accuracy, a clinical bias still existed after fine-tuning, with errors significantly lower in HC subjects as compared to ALS and PS participants. This is consistent with recent findings from Asgarian *et al.* [19], where fine-tuning could not reduce the gap in face alignment error between older adults with and without dementia. This result can be explained by the presence of two main types of domain shifts in this problem: the first one is the difference between the original training data and our dataset; the second one is the presence of oro-facial impairment. Although it is difficult to delineate how much of this error reduction depended on each of these two types of domain shifts, our results suggested that the former was prevalent. In fact, the error reduction in HC subjects was in most cases comparable to the one obtained in the two clinical groups, and it can be explained by the composition of our dataset, which included older adults who are generally not well represented in the dataset used for pre-training. Moreover, recent work [15] suggested that the number of patients needed to remove clinical bias in individuals with facial palsy had to be at least 40 or higher. Although our dataset included different populations and tasks and thus a direct comparison cannot be made, the sizes of our two clinical groups were 3 to 4 times smaller than the proposed sample size. Thus, future work will focus on expanding this dataset.

Despite the small sample sizes, however, some evidence from our experiments also suggested that fine-tuning might have some effect – although small – on decreasing the bias. In fact, correlation with facial asymmetry decreased and error decrease in clinical groups can be slightly higher as compared to HC subjects. Thus, future research will also focus on understanding if an optimal composition of dataset for fine-tuning exists and if different clinical conditions require different types of data – in addition to larger datasets – for removing the algorithmic bias.

VIII. Conclusion

In this work, we developed the first dataset with facial images and videos from individuals with oro-facial impairment due to stroke and ALS, as well as videos from age-matched healthy control subjects. Our experiments demonstrated that, even in case of standardized experimental setup (e.g., frontal face, uniform illumination, short distance from the camera) and mild to moderate oro-facial impairment due to neurological diseases, a bias in the face alignment accuracy occurred. We also demonstrated that fine-tuning the face-alignment algorithm on the target dataset improved the landmark localization accuracy, but only had a mild effect on removing the algorithmic bias. Thus, more efforts should be made by the research community to publish new datasets with images and videos from clinical populations, in this particular case neurological diseases affecting the oro-facial functions.

In addition to new investigations on algorithmic bias in face alignment, future work will focus on using this dataset to improve the automatic identification of neurological disorders and the estimation of disease severity from videos and images of oro-facial gestures.

The paucity of available datasets with facial images from clinical populations remains the main issue that hinders the development of robust face alignment algorithms able to deal with the large inter- and intra-group variability present in clinical conditions affecting the oro-facial musculature. The availability of novel datasets in the field can foster the development of accurate approaches for the automatic assessment of neurological diseases and oro-facial impairments.

Acknowledgment

The authors would like to thank Hoda Nabavi and Zakia Hussain for their valuable assistance in data annotation. The authors would also like to thank all the participants and their families involved in the study.

References

[1] M. Bologna, G. Fabbrini, L. Marsili, G. Defazio, P. D. Thompson, and A. Berardelli, "Facial bradykinesia," *J Neurol Neurosurg Psychiatry*, vol. 84, no. 6, pp. 681–685, 2013.

[2] H. L. Flowers, F. L. Silver, J. Fang, E. Rochon, and R. Martino, "The incidence, co-occurrence, and predictors of dysphagia, dysarthria, and aphasia after first-ever acute ischemic stroke," *Journal of communication disorders*, vol. 46, no. 3, pp. 238–248, 2013.

[3] S. E. Langmore and M. E. Lehman, "Physiologic deficits in the orofacial system underlying dysarthria in amyotrophic lateral sclerosis," *Journal of Speech, Language, and Hearing Research*, vol. 37, no. 1, pp. 28–37, 1994.

[4] P. Rong, Y. Yunusova, J. Wang, and J. R. Green, "Predicting early bulbar decline in amyotrophic lateral sclerosis: A speech subsystem approach," *Behavioural Neurology*, vol. 2015, 2015.

[5] Y. Yunusova, J. R. Green, M. J. Lindstrom, L. J. Ball, G. L. Pattee, and L. Zinman, "Kinematics of disease progression in bulbar als," *Journal of communication disorders*, vol. 43, no. 1, pp. 6–20, 2010.

[6] K. M. Allison, Y. Yunusova, T. F. Campbell, J. Wang, J. D. Berry, and J. R. Green, "The diagnostic utility of patient-report and speech-language pathologists' ratings for detecting the early onset of bulbar symptoms due to als," *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration*, vol. 18, no. 5-6, pp. 358–366, 2017.

[7] A. Bandini, J. R. Green, J. Wang, T. F. Campbell, L. Zinman, and Y. Yunusova, "Kinematic features of jaw and lips distinguish symptomatic from presymptomatic stages of bulbar decline in amyotrophic lateral sclerosis," *Journal of Speech, Language, and Hearing Research*, vol. 61, no. 5, pp. 1118–1129, 2018.

[8] A. Ashraf, A. Yang, and B. Taati, "Pain expression recognition using occluded faces," in *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*. IEEE, 2019, pp. 1–5.

[9] A. B. Ashraf, S. Lucey, J. F. Cohn, T. Chen, Z. Ambadar, K. M. Prkachin, and P. E. Solomon, "The painful face–pain expression recognition using active appearance models," *Image and vision computing*, vol. 27, no. 12, pp. 1788–1796, 2009.

[10] 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.

[11] 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 *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 2018, pp. 150–157.

[12] 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.

[13] M. Bishay, P. Palasek, S. Priebe, and I. Patras, "Schinet: Automatic estimation of symptoms of schizophrenia from facial behaviour analysis," *IEEE Transactions on Affective Computing*, 2019.

[14] D. L. Guarin, J. Dusseldorp, T. A. Hadlock, and N. Jowett, "A machine learning approach for automated facial measurements in facial palsy," *JAMA facial plastic surgery*, vol. 20, no. 4, pp. 335–337, 2018.

[15] D. L. Guarin, B. Taati, T. Hadlock, and Y. Yunusova, "Automatic facial landmark localization in clinical populations–improving model performance with a small dataset."

[16] S. Song, L. Shen, and M. Valstar, "Human behaviour-based automatic depression analysis using hand-crafted statistics and deep learned spectral features," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 2018, pp. 158–165.

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

[18] D. Guarin, A. Dempster, A. Bandini, Y. Yunusova, and B. Taati, "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)(FG)*, pp. 705–708.

[19] A. Asgarian, S. Zhao, A. B. Ashraf, M. Erin Browne, K. M. Prkachin, A. Mihailidis, T. Hadjistavropoulos, and B. Taati, "Limitations and biases in facial landmark detection d an empirical study on older adults with dementia," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 28–36.

[20] D. L. Guarin, Y. Yunusova, B. Taati, J. R. Dusseldorp, S. Mohan, J. Tavares, M. M. van Veen, E. Fortier, T. A. Hadlock, and N. Jowett, "Toward an automatic system for computer-aided assessment in facial palsy," *Facial Plastic Surgery & Aesthetic Medicine*, vol. 22, no. 1, pp. 42–49, 2020.

[21] 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.

[22] J. Thevenot, M. B. López, and A. Hadid, "A survey on computer vision for assistive medical diagnosis from faces," *IEEE journal of biomedical and health informatics*, vol. 22, no. 5, pp. 1497–1511, 2017.

[23] M. Tavakolian and A. Hadid, "A spatiotemporal convolutional neural network for automatic pain intensity estimation from facial dynamics," *International Journal of Computer Vision*, pp. 1–13, 2019.

[24] Y. Wang, A. Dantcheva, J.-C. Broutart, P. Robert, F. Bremond, and P. Bilinski, "Comparing methods for assessment of facial dynamics in patients with major neurocognitive disorders," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 0–0.

[25] S. Happy, A. Dantcheva, A. Das, R. Zeghari, P. Robert, and F. Bremond, "Characterizing the state of apathy with facial expression and motion analysis," in *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*. IEEE, 2019, pp. 1–8.

[26] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, no. 6, pp. 681–685, 2001.

[27] I. Matthews and S. Baker, "Active appearance models revisited," *International journal of computer vision*, vol. 60, no. 2, pp. 135–164, 2004.

[28] X. Xiong and F. De la Torre, "Supervised descent method and its applications to face alignment," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 532–539.

[29] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1867–1874.

[30] 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.

[31] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[32] X. P. Burgos-Artizzu, P. Perona, and P. Dollár, "Robust face landmark estimation under occlusion," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 1513–1520.

[33] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-pie," *Image and Vision Computing*, vol. 28, no. 5, pp. 807–813, 2010.

[34] M. Koestinger, P. Wohlhart, P. M. Roth, and H. Bischof, "Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization," in *2011 IEEE international conference on computer vision workshops (ICCV workshops)*. IEEE, 2011, pp. 2144–2151.

[35] 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.

[36] W. Wu, C. Qian, S. Yang, Q. Wang, Y. Cai, and Q. Zhou, "Look at boundary: A boundary-aware face alignment algorithm," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2129–2138.

[37] P. Lucey, J. F. Cohn, K. M. Prkachin, P. E. Solomon, and I. Matthews, "Painful data: The UNBC-McMaster shoulder pain expression archive database," in *Face and Gesture 2011*. IEEE, 2011, pp. 57–64.

[38] S. Walter, S. Gruss, H. Ehleiter, J. Tan, H. C. Traue, P. Werner, A. Al-Hamadi, S. Crawcour, A. O. Andrade, and G. M. da Silva, "The BioVid heat pain database data for the advancement and systematic validation of an automated pain recognition system," in *2013 IEEE international conference on cybernetics (CYBCO)*. IEEE, 2013, pp. 128–131.

[39] J. J. Greene, D. L. Guarin, J. Tavares, E. Fortier, M. Robinson, J. Dusseldorp, O. Quatela, N. Jowett, and T. Hadlock, "The spectrum of facial palsy: The meei facial palsy photo and video standard set," *The Laryngoscope*.

[40] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," in *Conference on fairness, accountability and transparency*, 2018, pp. 77–91.

[41] M. Wang, W. Deng, J. Hu, X. Tao, and Y. Huang, "Racial faces in the wild: Reducing racial bias by information maximization adaptation network," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 692–702.

[42] T. Wang, J. Zhao, M. Yatskar, K.-W. Chang, and V. Ordonez, "Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 5310–5319.

[43] Z. S. Nasreddine, N. A. Phillips, V. Bédirian, S. Charbonneau, V. Whitehead, I. Collin, J. L. Cummings, and H. Chertkow, "The montreal cognitive assessment, moca: a brief screening tool for mild cognitive impairment," *Journal of the American Geriatrics Society*, vol. 53, no. 4, pp. 695–699, 2005.

[44] B. R. Brooks, R. G. Miller, M. Swash, and T. L. Munsat, "El escorial revisited: revised criteria for the diagnosis of amyotrophic lateral sclerosis," *Amyotrophic lateral sclerosis and other motor neuron disorders*, vol. 1, no. 5, pp. 293–299, 2000.

[45] J. M. Cedarbaum, N. Stambler, E. Malta, C. Fuller, D. Hilt, B. Thurmond, A. Nakanishi, B. A. S. Group, A. complete listing of the BDNF Study Group *et al.*, "The ALSFRS-R: a revised als functional rating scale that incorporates assessments of respiratory function," *Journal of the neurological sciences*, vol. 169, no. 1-2, pp. 13–21, 1999.

[46] J. R. Duffy, "Motor speech disorders: clues to neurologic diagnosis," in *Parkinson's Disease and Movement Disorders*. Springer, 2000, pp. 35–53.

[47] Y. Yunusova, J. R. Green, J. Wang, G. Pattee, and L. Zinman, "A protocol for comprehensive assessment of bulbar dysfunction in amyotrophic lateral sclerosis (als)," *JoVE (Journal of Visualized Experiments)*, no. 48, p. e2422, 2011.

[48] G. H. McCullough, R. T. Wertz, J. C. Rosenbek, R. H. Mills, K. B. Ross, and J. R. Ashford, "Inter-and intrajudge reliability of a clinical examination of swallowing in adults," *Dysphagia*, vol. 15, no. 2, pp. 58–67, 2000.

[49] D. Cristinacce and T. F. Cootes, "Feature detection and tracking with constrained local models." in *Bmvc*, vol. 1, no. 2. Citeseer, 2006, p. 3.

[50] X. Jin and X. Tan, "Face alignment in-the-wild: A survey," *Computer Vision and Image Understanding*, vol. 162, pp. 1–22, 2017.

[51] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *European conference on computer vision*. Springer, 2016, pp. 483–499.

[52] A. Bulat and G. Tzimiropoulos, "Binarized convolutional landmark localizers for human pose estimation and face alignment with limited resources," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 3706–3714.

[53] J. Alabort-i Medina, E. Antonakos, J. Booth, P. Snape, and S. Zafeiriou, "Menpo: A comprehensive platform for parametric image alignment and visual deformable models," in *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 2014, pp. 679–682.

[54] C. Sagonas, E. Antonakos, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "300 faces in-the-wild challenge: Database and results," *Image and vision computing*, vol. 47, pp. 3–18, 2016.

[55] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "A semi-automatic methodology for facial landmark annotation," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2013, pp. 896–903.

[56] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li, "Face alignment across large poses: A 3d solution," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 146–155.