Automatic Facial Landmark Localization in Clinical Populations - Improving Model Performance with a Small Dataset

Diego L Guarin (✉ diego.guarin@uhn.ca)

Toronto Rehabilitation Institute <https://orcid.org/0000-0001-6077-7063>

Babak Taati

Toronto Rehabilitation Institute

Tessa Hadlock

Massachusetts Eye and Ear Infirmary

Yana Yunusova

University of Toronto

Research

Keywords: Automatic Facial Analysis, Facial Landmarks Localization, Machine Learning, Computer Vision

Posted Date: March 30th, 2020

DOI: <https://doi.org/10.21203/rs.3.rs-19405/v1>

License: This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

RESEARCH1

Automatic Facial Landmark Localization in2
Clinical Populations – Improving Model3
Performance with a Small Dataset4
5
6
7
8

Diego L. Guarin^{1,2,3*}, Babak Taati^{1,4,5}, Tessa Hadlock² and Yana Yunusova^{1,6,7}9

10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33

Correspondence:

Diego.Guarin@uhn.ca

KITE — Toronto Rehabilitation
Institute, 550 University Av.,
Toronto, Canada

Full list of author information is
available at the end of the article

Abstract

Background: Automatic facial landmark localization in videos is an important first step in many computer vision applications, including the objective assessment of orofacial function. Convolutional neural networks (CNN) for facial landmarks localization are typically trained on faces of healthy and young adults, so model performance is inferior when applied to faces of older adults or people with diseases that affect facial movements, a phenomenon known as algorithmic bias. Fine-tuning pre-trained CNN models with representative data is a well-known technique used to reduce algorithmic bias and improve performance on clinical populations. However, the question of how much data is needed to properly fine-tune the model remains.

Methods: In this paper, we fine-tuned a popular CNN model for automatic facial landmarks localization using different number of manually annotated photographs from patients with facial palsy and evaluated the effects of the number of photographs used for model fine-tuning in the model performance by computing the normalized root mean squared error between the facial landmarks positions predicted by the model and those provided by manual annotators. Furthermore, we studied the effect of annotator bias by fine-tuning and evaluating the model with data provided by multiple annotators.

Results: Our results showed that fine-tuning the model with as little as 8 photographs from a single patient significantly improved the model performance on other individuals from the same clinical population, and that the best performance was achieved by fine-tuning the model with 320 photographs from 40 patients. Using more photographs for fine-tuning did not improve the model performance further. Regarding the annotator bias, we found that fine-tuning a CNN model with data from one annotator resulted in models biased against other annotators; our results also showed that this effect can be diminished by averaging data from multiple annotators.

Conclusions: It is possible to remove the algorithmic bias of a depth CNN model for automatic facial landmark localization using data from only 40 participants (total of 320 photographs). These results pave the way to future clinical applications of CNN models for the automatic assessment of orofacial function in different clinical populations, including patients with Parkinson's disease and stroke.

Keywords: Automatic Facial Analysis; Facial Landmarks Localization; Machine Learning; Computer Vision

¹Background

²Clinical assessment of orofacial function in patients with neurological diseases such²
³as dementia, Parkinson's disease, stroke, and facial palsy is typically performed³
⁴in face-to-face interviews with trained clinicians or using laboratory-based tests⁴
⁵[1]. However, clinical evaluations are subjective, can be unreliable, and suffer from⁵
⁶poor interrater agreement [2]. Moreover, laboratory tests, such as electromyography,⁶
⁷magnetic resonance imaging, and facial motion tracking technology, are expensive,⁷
⁸time-consuming, and their availability is limited [1].⁸

⁹New methods and algorithms based on recent advancements in machine learning⁹
¹⁰(ML) and computer vision (CV) have been developed to assess orofacial function¹⁰
¹¹objectively and extract clinically useful information automatically from videos and¹¹
¹²photographs of patients [3–10]. These techniques track orofacial movements with¹²
¹³out using facial markers or specialized equipment. Different from laboratory tests,¹³
¹⁴these techniques are inexpensive, easy-to-use, readily available, and can be used¹⁴
¹⁵at-home. Our team have used these methods to evaluate the effectiveness of surgi-¹⁵
¹⁶cal treatments for Bell's Palsy [4, 5, 8], and to extract biomarkers for detection and¹⁶
¹⁷assessment of orofacial deficits in Parkinson's disease [10, 11], amyotrophic lateral¹⁷
¹⁸sclerosis [6], and stroke [7]. These techniques have the potential to revolutionize the¹⁸
¹⁹assessment of orofacial deficits as they would allow frequent, cost-effective, objec-¹⁹
²⁰tive in-home monitoring of disease progression and treatment effectiveness across²⁰
²¹various neurological diseases affecting the orofacial musculature.²¹

²²The technology behind these methods and algorithms known as facial alignment²²
²³(FA) – a set of ML models and algorithms for automatic localization of facial land-²³
²⁴marks in digital images and videos – has been an active research topic for more²⁴
²⁵than twenty-five years [12]. Convolutional Neural Networks (CNN) have recently²⁵
²⁶emerged as the mainstream approach for FA; in particular, the Facial Alignment²⁶
²⁷Network (FAN), a type of CNN, demonstrated superior performance in FA for²⁷
²⁸photographs taken in realistic scenarios under challenging illumination, pose, and²⁸
²⁹expressions [13–18]. We recently demonstrated that FAN provides superior perfor-²⁹
³⁰mances when applied to clinical populations as compared to other popular algo-³⁰
³¹rithms for FA [19]. However, challenges remain when applying algorithms for facial³¹
³²analysis to a population with diversity in gender, and ethnicity [20], and to patients³²
³³with diversity in age, physical/cognitive ability, and orofacial deficits [9, 21]. We³³

¹and others have shown that publicly available, i.e., pre-trained, FA algorithms and¹
²models perform better in healthy, young males, with light skin tones as compared²
³to patients with neurological diseases [8, 9, 18, 21].³

⁴ Recently, we retrained classical FA algorithms such as active appearance mod-⁴
⁵els [21], and cascade of regression trees [9], and fine-tuned the FAN model using⁵
⁶transfer learning [18, 21] to improve their landmark localization accuracy in clini-⁶
⁷cal populations including elderly subjects with dementia and patients with Bell's⁷
⁸palsy. Our results demonstrated that using patient-specific databases significantly⁸
⁹improved the models' performance. We obtained these results after including more⁹
¹⁰than 700 images for model training in the case of patients with dementia [18, 21],¹⁰
¹¹and more than 1500 images in the case of unilateral facial palsy [9].¹¹

¹² The clinical application of these results is limited as developing such large¹²
¹³databases from patients with various - and often rare - diseases for model fine-¹³
¹⁴tuning is difficult, expensive, and oftentimes impossible. Thus, the first objective of¹⁴
¹⁵this paper was to evaluate to what extent the size of the training set (i.e., number¹⁵
¹⁶of disease-specific photographs) used to fine-tune CNN models for FA influences the¹⁶
¹⁷model performance. Based on results on fine-tuning CNN models for image clas-¹⁷
¹⁸sification [22], we hypothesized that a relatively small number of standard clinical¹⁸
¹⁹photographs would be sufficient to significantly boost the performance of the FAN¹⁹
²⁰model for FA.²⁰

²¹ Moreover, a recent study showed that ML models trained with manually anno-²¹
²²tated databases might be influenced by the annotator, and their performance may²²
²³vary depending on the annotator [23]. The annotator bias has important conse-²³
²⁴quences for the real-world application of ML models and has been an active re-²⁴
²⁵search topic in the area of Natural Language Processing [24–26]. However, to our²⁵
²⁶knowledge, the problem of annotator bias has not been studied in the context of FA.²⁶
²⁷ Thus, the second objective of this paper was to estimate the effects of annotator²⁷
²⁸bias in the model performance and evaluate approaches to reduce the bias. Based²⁸
²⁹on previous results from natural language processing [25], we hypothesized that by²⁹
³⁰averaging the landmark positions provided by multiple annotators, it was possible³⁰
³¹to reduce the effects of annotator bias, and that this reduction was less meaningful³¹
³²when the inter-annotators agreement was higher.³²

Materials and Methods

Databases

Two different datasets were used in this study for fine-tuning and evaluating the FA algorithm; in total, 250 patients and 10 healthy subjects participated in the creation of these datasets. The first dataset consists of 1600 facial photographs from 200 patients with unilateral facial paralysis [9]. The dataset contains facial photographs of 135 females and 65 males aged between 7 and 89 years.

During the creation of this datasets, a clinical photographer captured high-resolution photographs (1080×720 pixels) of a standardized series of eight facial expressions used to evaluate facial mimetic function [27]; these expression corresponds to rest, eyebrow elevation, light effort eye closure, full effort eye closure, light effort smile, full effort smile, pucker, and lip depression. Photographs of patients were obtained while they were looking straight at the camera and under the same illumination and background conditions.

The 1600 facial photographs were randomly divided into three non-overlapping groups; all images of the same patient were included in the same group. The groups were: (1) model training (160 patients, corresponding to $N=1280$ photographs), (2) model validation (20 patients, corresponding to $N=160$ photographs), and (3) model testing (20 patients, corresponding to $N=160$ photographs).

The second dataset consists of 480 facial photographs from 60 subjects: 50 patients with unilateral facial paralysis, and 10 healthy controls [28]. Similar to the first dataset, a clinical photographer captured high-resolution photographs (1080×720 pixels) of a standardized series of eight facial expressions used to evaluate facial mimetic function. Each patient was globally characterised by the degree of their facial paralysis as: near normal, mild, moderate, severe, and complete by an expert clinician.

In each of the photograph, the locations of 68 facial landmarks, defining the eyebrows, eyes, nose, mouth, and jawline, were manually localized by three independent annotators using Emotrics, an open-source interface for facial landmark annotation [8]. Annotators were trained together but performed the annotation independently at their own pace. Landmark coordinates of three manual annotators were averaged for each photograph to define ground-truth locations.

¹Pre-trained FAN model

²The pre-trained FAN model has been trained with 230,000 facial photographs [16].²

³The database used to train the FAN model contained a large variety of poses, expres-³

⁴sions, illumination, backgrounds, and scales so that the pre-trained model is a robust⁴

⁵and generic tool for FA [19]. The database used for model training and the result-⁵

⁶ing pre-trained model is freely available online (<https://github.com/1adrianb/face-6>

⁷alignment).

⁸ The FAN model is composed of five sequential stages; the first stage transforms⁸

⁹the input image into a set of 256 feature maps of dimensions 64x64, the remaining⁹

¹⁰four stages are a series of stacked hourglass networks that transform the features¹⁰

¹¹into a set of 68 heatmaps of dimensions 64x64; each heatmap provides the estimated¹¹

¹²position of a facial landmark [13].

¹³

¹⁴Fine-tuning the FAN model with patient's data

¹⁵To fine-tune the FAN model, we froze the provided pre-trained model's parameters¹⁵

¹⁶of the first four out of five stages and updated the parameters of the last hourglass¹⁶

¹⁷network using photographs of people with facial palsy, similar to what we did before¹⁷

¹⁸in our study of people with dementia [21]. Fine-tuning was performed in Python¹⁸

¹⁹using the PyTorch library on a Nvidia GeForce GTX 1080 graphics processing unit.¹⁹

²⁰ We used the first dataset to evaluate whether the number of disease-specific facial²⁰

²¹photographs used to fine-tune the FAN model affected the accuracy of facial land-²¹

²²mark localization in patients. Data from the model training group was divided into²²

²³ten subgroups, each subgroup containing a randomly selected group of patients; the²³

²⁴ten subgroups contained 1, 2, 3, 4, 5, 10, 20, 40, 80 and 160 patients (corresponding²⁴

²⁵to 8, 16, 24, 32, 40, 80, 160, 320, 640 and 1280 photographs respectively). Patients'²⁵

²⁶photographs in each subgroup were used to fine-tune the FAN model so that ten²⁶

²⁷new models were generated. Training was performed for 50 epochs with early stop-²⁷

²⁸ping to prevent over-fitting. The model validation group was used to determine the²⁸

²⁹best model during training

³⁰Evaluation of algorithmic bias against clinical population

³¹We used the second dataset to evaluate the algorithmic bias against clinical popu-³¹

³²lation and how fine-tuning the model with representative data mitigates the bias.³²

³³ A total of 15 subjects were selected from the database: 5 near normal, 5 with³³

1 moderate paralysis, and 5 with severe paralysis. Only photographs from one single¹
 2 expression - full effort smile - were used for evaluation of algorithmic bias. In total,²
 3 15 photographs were used; facial landmarks in each photograph were estimated us-³
 4 ing the pre-trained FAN model, and the models fine-tuned with different number⁴
 5 of patients. 5

6 Evaluation of annotator bias 6

7 We used the first dataset to estimate the effects of the annotator bias in the model⁷
 8 performance and evaluate i) whether by averaging the landmark positions provided⁸
 9 by multiple annotators it would be possible to reduce the effects of annotator bias,⁹
 10 and ii) whether this reduction was less meaningful when inter-annotators agreement¹⁰
 11 was higher. We first estimated the agreement between annotators by computing the¹¹
 12 Euclidian distance in pixels between landmarks provided by each annotator for all¹²
 13 photographs. Then, we created six additional datasets with data provided by each¹³
 14 annotator (three datasets), and by averaging the data provided by each pair of¹⁴
 15 annotators (three datasets). Similarly to the initial experiment used for fine-tuning¹⁵
 16 the FAN model with patient's data, the new datasets were split into model training¹⁶
 17 (160 patients, corresponding to $N = 1280$ images), model validation (20 patients,¹⁷
 18 corresponding to $N=160$ images), and model testing (20 patients, corresponding to¹⁸
 19 $N=160$ images). These datasets were used to create six additional models by fine-¹⁹
 20 tuning the pre-trained FAN model using all the 1280 images in the model training²⁰
 21 set. 21

22 Model evaluation 22

23 A total of sixteen new models were created: Ten with the ground truth landmarks by²³
 24 dividing the patients' dataset into subgroups with different number of patients; three²⁴
 25 more were created by using the landmark information provided by each annotator;²⁵
 26 and three more were created by averaging the landmarks provided by each pair²⁶
 27 of annotators. The performance of these sixteen models as well as the original,²⁷
 28 pre-trained model is reported in this paper only for images that were not used for²⁸
 29 fine-tuning or validating the models, that is, the model testing subgroup of the first²⁹
 30 dataset, and the selected 15 photographs of the second dataset. 30

31 Accuracy in landmark localization was computed in terms of the Root-Mean-³¹
 32 Squared Error (RMSE) between manually annotated and FAN-predicted landmark³²
 33 33

positions normalized by the intercanthal distance (NRMSE) [28]. Statistical differences between models were evaluated using the non-parametric Wilcoxon signed-rank test. Statistical significance was considered at $p < 0.05$.

Results

Facial landmark detection with Pre-trained and fine-tuned FAN model

Figure 1 shows a box and whiskers plot representing the NRMSE obtained for the model testing group with the pre-trained FAN - corresponding to 0 on the x-axis. The mean and standard deviation of the NRMSE for the pre-trained model were $11.5 \pm 3.3\%$. The remaining results in Figure 1 show that fine-tuning the network with patients' photographs led to improved accuracy in facial landmark localization in this patient population. Moreover, results also show that using more than 40 patients (N=320 photographs) for fine-tuning did not produce further significant improvements in accuracy.

Specifically, fine-tuning the FAN model with 1 subject (N=8 photographs) produced a model with mean and standard deviation of the NRMSE of $10.1 \pm 3.4\%$, a significant reduction from the error produced by the pre-trained network ($p < 0.05$). Including more patients for fine-tuning improved the model accuracy, but including 2, 3, 4 or 5 patients for fine-tuning did not significantly improved the model performance, when compared with results yielded by the FAN model fine-tuned with 1 patient. Using 10 patients (N=80 photographs) resulted in a model with mean and standard deviation of the NRMSE of $8.4 \pm 2.6\%$ - a significant reduction from the error produced by the FAN model fine-tuned with 1-5 patients ($p < 0.05$). Increasing the number of patients to 40 (N=320 photographs) resulted in a model with mean and standard deviation of the NRMSE at $6.7 \pm 2.0\%$ - a significant reduction from the error produced by the FAN model fine-tuned with 10 patients ($p < 0.05$). Finally, increasing the number of patients used for fine-tuning to 80 or 160 (N= 640 or 1280 photographs respectively) did not result in significant improvement in accuracy when compared with results yielded by the FAN model fine-tuned with 40 patients ($p > 0.05$ in both cases).

Algorithmic bias

Figure 2 shows the mean and standard deviation of the NRMSE obtained for the photographs of near normal patients (blue line), patients with moderate paralysis

¹(green line), and patients with complete paralysis (red line) with the pre-trained¹
²model, corresponding to 0 in the x-axis. Results demonstrate a significant difference²
³in the model performance as a function of disease severity, NRMSE changed from³
⁴ $8.3 \pm 1.8\%$ for near normal patients, to $9.8 \pm 1.2\%$ for patients with moderate⁴
⁵paralysis, and to $12.3 \pm 2.1\%$ for patients with severe paralysis. Statistical analysis⁵
⁶showed a significant difference between near normal and severe paralysis group⁶
⁷($p < 0.05$), and non-significant difference among other groups.⁷

⁸ As observed in Figure 2, fine-tuning improved the model performance for all pa-⁸
⁹tients groups. In particular, after fine-tuning the model with 40 patients (N=320⁹
¹⁰photographs), the NRMSE decreased to $6.1 \pm 1.3\%$, $6.1 \pm 0.8\%$, and $7.1 \pm 1.6\%$ for the¹⁰
¹¹near normal, moderate paralysis, and complete paralysis groups respectively. Sta-¹¹
¹²tistical analysis showed no-significant difference between groups ($p=0.99$ between¹²
¹³near normal and moderate paralysis, $p=0.28$ between near normal and complete¹³
¹⁴paralysis, and $p=0.36$ between moderate paralysis and complete paralysis). Finally,¹⁴
¹⁵no further improvement in the NRMSE was observed after fine-tuning the model¹⁵
¹⁶with more than 40 patients.¹⁶

¹⁷ ¹⁸Annotator bias¹⁸

¹⁹Table 1 presents the Euclidean distance (mean $\pm$ standard deviation) in pixels¹⁹
²⁰between landmarks provided by each annotator. The results show that annotators²⁰
²¹1 and 2 had a higher landmark localization agreement than annotators 1 and 3, and²¹
²²annotators 2 and 3.²²

²³ Table 2 presents the results of the annotator bias analysis. The table shows the²³
²⁴NRMSE (mean $\pm$ standard deviation) between the position of the landmarks yielded²⁴
²⁵by the six models trained with the databases created by combining different number²⁵
²⁶of annotators compared to the manually annotated landmarks provided by each²⁶
²⁷annotator.²⁷

²⁸ The first three rows of Table 2 present the results of fine-tuning the model with²⁸
²⁹data from one single annotator. These results indicated that a model trained with²⁹
³⁰data from one annotator provided significantly lower NRMSE for testing data from³⁰
³¹the same annotator as compared to testing data provided by other annotators.³¹
³²Specifically, the model fine-tuned with data from Annotator 1, yielded a NRMSE³²
³³of $7.9 \pm 2.0\%$ when compared against manual annotations provided by the same³³

¹annotator, the NRMSE significantly increased to $8.2 \pm 2.2\%$ ($p < 0.05$) when¹
²compared against manual annotations provided by Annotator 2, and to $8.7 \pm 2.4\%$ ²
³($p < 0.05$) when compared against manual annotations provided by Annotator 3.³
⁴Similar results were obtained for models fine-tuned with data from other annotators.⁴
⁵Finally, the last three rows of Table 2 present the results of fine-tuning the model⁵
⁶with data obtained by averaging the landmark positions provided by each pair⁶
⁷of annotators. Results indicated that averaging the landmark positions from two⁷
⁸annotators could help to remove the bias against other annotator not seen in the⁸
⁹training set; however, these results seems to depend on the agreement between⁹
¹⁰annotators. For instance, we observed that the model created by averaging the¹⁰
¹¹annotations provided by Annotators 1 and 3, yielded a NRMSE of $8.0 \pm 2.0\%$ and¹¹
¹² $7.9 \pm 2.2\%$ when compared against manual annotations yielded by Annotator 1 and 3¹²
¹³respectively; and the NRMSE did not increase significantly when compared against¹³
¹⁴manual annotations provided by Annotator 2. Similar results were obtained for the¹⁴
¹⁵model created by averaging the annotations provided by Annotators 2 and 3. In¹⁵
¹⁶contrast, the model created by averaging the annotations provided by Annotators¹⁶
¹⁷1 and 2, who have the highest agreement, yielded a NRMSE of $7.6 \pm 1.9\%$ and¹⁷
¹⁸ $7.5 \pm 2.0\%$ when compared against manual annotations yielded by Annotator 1 and¹⁸
¹⁹2 respectively, and the NRMSE significantly increased to $8.5 \pm 2.3\%$ ($p < 0.05$)¹⁹
²⁰when compared against manual annotations provided by Annotator 3.²⁰

²²Discussion²²

²³Automatic assessment of orofacial function has the potential to revolutionize the²³
²⁴diagnosis and monitoring of neurological diseases including Parkinson's disease,²⁴
²⁵stroke, ALS, and facial palsy. Translation from laboratory methods to clinically²⁵
²⁶useful tools requires accurate algorithms for automatic localization of facial land-²⁶
²⁷marks in photographs and videos of patients. However, advancements in FA tech-²⁷
²⁸nology have mostly focused on applications tailored for young healthy population,²⁸
²⁹such as identity recognition [12]. Thus, publicly available FA algorithms perform²⁹
³⁰poorly when applied to populations with orofacial abnormalities [9] or elderly sub-³⁰
³¹jects with neurological conditions [21]. Previously, addressing this challenge implied³¹
³²acquiring thousands of images with manually annotated landmarks and training FA³²
³³algorithms using these images [9, 21] However, recent advances in neural network³³

models have provided new alternatives whereby CNN models pre-trained with a¹
 large corpus of data are fine-tuned with a relatively small dataset to improve their²
 performance for a specific task or population [22, 29]. In this study, we explored³
 the utility of this technique to improve the accuracy of FA algorithms in patient⁴
 population, and demonstrated that only a small corpus of representative data –⁵
 40 patients with a total of 320 photographs – is needed to remove the algorithmic⁶
 bias of a well-known CNN model for facial alignment. Next, we discuss the different⁷
 contributions of this study and its clinical implications. 8

⁹ Fine-tuning FA models 9

¹⁰ The pre-trained FAN, trained with 230,000 images of healthy subjects [16], per-¹¹
¹²formed poorly when applied to images of patients with facial palsy, as measured¹²
¹³by a large NRMSE and large error variability. The poor performance might be ex-¹³
¹⁴plained by missing information in the training dataset related to 1) patients' facial¹⁴
¹⁵abnormalities, 2) patients' pose and expression, 3) photographs illumination, and¹⁵
¹⁶4) differences in manual annotations between databases used to train and evalu-¹⁶
¹⁷ate the model. The results indicated that by fine-tuning the FAN with a handful¹⁷
¹⁸of patient-specific photographs the network learned additional information that¹⁸
¹⁹boosted its performance. Even fine-tuning with only 8 photographs, produced a¹⁹
²⁰significant improvement in performance (NRMSE reduced from 11.5% to 10.1%). 20

²¹ Including more patients' photographs in the training dataset further improved²¹
²²the model performance. However, the improvement was not linearly related to the²²
²³number of images used for fine-tuning, and there was a limit to what the network²³
²⁴could learn from these images. Our results showed that using 40 patients (N=320²⁴
²⁵photographs) decreased the NRMSE to $6.7 \pm 2.0\%$, but adding more patients for fine-²⁵
²⁶tuning resulted in non-significant improvements in model accuracy. These results²⁶
²⁷can be explained by the fact that background, illumination, and pose are carefully²⁷
²⁸controlled in these clinical photographs; the main sources of variability are identity,²⁸
²⁹expressions, and orofacial abnormalities. Our results showed that the network was²⁹
³⁰able to explain these sources of variability with a relatively small corpus of data. 30

³¹ Algorithmic bias 31

³² In this study, we demonstrated that the well-known FAN model for FA is biased 32
³³against patients with facial palsy. The model performed significantly worse when 33

¹applied to patients with complete paralysis than when applied to near normal pa-¹
²tients. This result is in line with our recent results on elderly subjects with dementia²
³[17], [20], and stroke survivors [18]. Furthermore, our results showed that fine-tuning³
⁴CNN models with representative data significantly improved the model performance⁴
⁵when applied to subject with severe facial deformities, and eliminated the algorithm⁵
⁶bias against this population. We observed that after fine-tuning the model with 40⁶
⁷patients, there was only a small, non-significant, difference in the error obtained⁷
⁸when applied to patients with complete paralysis and to near normal patients. We⁸
⁹expect that these results will be transferable to other clinical populations. ⁹

10

10

¹¹Annotator bias ¹¹

¹²To our knowledge, our results are the first to quantitatively demonstrate that FA¹²
¹³CNN models trained with data provided by a single annotator are biased against¹³
¹⁴other annotators; these results agreed with observations made in other areas of¹⁴
¹⁵machine learning [23]. ¹⁵

¹⁶These results supported our hypotheses that was possible to mitigate the effects¹⁶
¹⁷of annotator bias by averaging the landmark positions provided by multiple annota-¹⁷
¹⁸tors; and that this effect was influenced by the level of agreement among annotators.¹⁸
¹⁹We observed that averaging annotators 1 and 3, and 2 and 3 produced models that¹⁹
²⁰were not biased against another annotator. However, averaging annotators 1 and 2,²⁰
²¹who had the highest agreement, did not have the same effect. Current training ap-²¹
²²proaches for FA algorithms involve combining multiple datasets manually annotated²²
²³by different groups without any standards, aside from the number of landmarks, and²³
²⁴their general locations. Our results indicated that this training approach might be²⁴
²⁵beneficial, as the CNN model will have to adapt to the multiple annotation styles,²⁵
²⁶potentially leading to more flexible models. ²⁶

27

27

²⁸Clinical implications ²⁸

²⁹Our results have important implications for future clinical applications of FA. We²⁹
³⁰showed that only a small manually annotated dataset significantly improved the³⁰
³¹landmark localization accuracy in a clinical population, and that there was a limit³¹
³²to what the model can learn from similar photographs. Based on these results, we³²
³³suggest that models should be fine-tuned incrementally as data become available,³³
³⁴because using even a very small corpus of representative data could significantly

¹improve the model performance with that population. This approach would help¹
²detect when the CNN model stops improving from the addition of more data, which²
³would help to guide the data collection procedure. Finally, we demonstrated that³
⁴CNN models learn annotators' styles and might be biased against other annotators.⁴
⁵Thus, it is advisable to include multiple annotators in the training and validation⁵
⁶data to help mitigate annotator bias. 6

⁷ 7

⁸**Conclusions** 8

⁹We demonstrated that fine-tuning CNN models with representative data can help⁹
¹⁰to eliminate the model bias against clinical populations. We produced an ML algo-¹⁰
¹¹rithm for automatic facial landmark localization in patients with facial palsy that is¹¹
¹²highly accurate, as compared to pre-trained models, using a only 320 manually an-¹²
¹³notated photographs for fine-tuning. Our results, open-access tools, and algorithms¹³
¹⁴pave the way to future clinical application of FA, as they facilitate and provide¹⁴
¹⁵useful guidelines for the application of these techniques in clinical populations. 15

¹⁶ 16

¹⁷Ethics Approval and Consent to Participate 17
¹⁸Not applicable 18

¹⁹Consent for publication 19
²⁰Not applicable 20

²¹Availability of supporting data 21

²²All the models discussed here and Python-based Jupyter notebooks exemplifying the use of the models to obtain 22
²³landmark position from images are available online (github.com/dguari1/Improving_facial_alignment). The first 23
²⁴annotated database cannot be shared online due to patient privacy concerns. Request from research institutes to 24
²⁵share the data are reviewed on a case-by-case basis by the MEE Institutional Review Board. The second annotated 25
²⁶database is available online for research purposes. For more information on how to obtain the images please refer 26
²⁷to [28]. 27

²⁸**Competing interests** 28

²⁹The authors declare that the research was conducted in the absence of any commercial or financial relationships 29
³⁰that could be construed as a potential conflict of interest. 30

³¹**Funding** 31

³²DLG is supported by the Michael J. Fox Foundation for Parkinson's Research. 32

³³**Author's contributions** 33

³⁴DLG designed the study, participated in data collection, wrote the code, performed the experiments, analyzed the 34
³⁵data, and wrote the manuscripts. BT provided technical guidance on the application of ML models for FA, and 35
³⁶helped to write the manuscript. TH directed the data collection, and provided clinical scores. YY provided support 36
³⁷with data analysis, and helped to write the manuscript. 37

³⁸**Acknowledgements** 38

³⁹Not applicable 39

¹ Author details

²¹KITE — Toronto Rehabilitation Institute, 550 University Av., Toronto, Canada. ²Department of Otolaryngology/Head and Neck Surgery, Massachusetts Eye and Ear Infirmary and Harvard Medical School, 243 Charles St, Boston, USA. ³Vector Institute for Artificial Intelligence, 661 University Ave Suite 710, Toronto, Canada. ⁴Department of Computer Science, University of Toronto, 214 College St, Toronto, Canada. ⁵Institute of Biomaterials and Biomedical Engineering, University of Toronto, Rosebrugh Bldg, 164 College St Room 407, Toronto, Canada. ⁶Department of Speech Language Pathology, University of Toronto, 500 University Av., Toronto, Canada. ⁷Hurvitz Brain Sciences Program, Sunnybrook Research Institute, 2075 Bayview Ave., Toronto, Canada.

⁷ References

1. Yunusova, Y., Plowman, E.K., Green, J.R., Barnett, C., Bede, P.: Clinical measures of bulbar dysfunction in als. *Frontiers in neurology* **10** (2019)
2. Richards, M., Marder, K., Cote, L., Mayeux, R.: Interrater reliability of the unified parkinson's disease rating scale motor examination. *Movement Disorders* **9**(1), 89–91 (1994)
3. Thevenot, J., López, M.B., Hadid, A.: A survey on computer vision for assistive medical diagnosis from faces. *IEEE journal of biomedical and health informatics* **22**(5), 1497–1511 (2017)
4. Greene, J.J., Tavares, J., Guarin, D.L., Jowett, N., Hadlock, T.: Surgical refinement following free gracilis transfer for smile reanimation. *Annals of plastic surgery* **81**(3), 329–334 (2018)
5. Greene, J.J., Tavares, J., Guarin, D.L., Hadlock, T.: Clinician and automated assessments of facial function following eyelid weight placement. *JAMA Facial Plastic Surgery* **21**(5), 387–392 (2019)
6. Bandini, A., Green, J.R., Taati, B., Orlandi, S., Zinman, L., Yunusova, Y.: 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), pp. 150–157 (2018). IEEE
7. Bandini, A., Green, J.R., Richburg, B., Yunusova, Y.: Automatic detection of orofacial impairment in stroke. In: *Interspeech*, pp. 1711–1715 (2018)
8. Guarin, D.L., Dusseldorp, J., Hadlock, T.A., Jowett, N.: A machine learning approach for automated facial measurements in facial palsy. *JAMA facial plastic surgery* **20**(4), 335–337 (2018)
9. Guarin, D.L., Yunusova, Y., Taati, B., Dusseldorp, J.R., Mohan, S., Tavares, J., van Veen, M.M., Fortier, E., Hadlock, T.A., Jowett, N.: Toward an automatic system for computer-aided assessment in facial palsy. *Facial Plastic Surgery & Aesthetic Medicine* **22**(1), 42–49 (2020)
10. Bandini, A., Orlandi, S., Giovannelli, F., Felici, A., Cincotta, M., Clemente, D., Vanni, P., Zaccara, G., Manfredi, C.: Markerless analysis of articulatory movements in patients with parkinson's disease. *Journal of Voice* **30**(6), 766–1 (2016)
11. Li, M.H., Mestre, T.A., Fox, S.H., Taati, B.: Vision-based assessment of parkinsonism and levodopa-induced dyskinesia with pose estimation. *Journal of neuroengineering and rehabilitation* **15**(1), 97 (2018)
12. Jin, X., Tan, X.: Face alignment in-the-wild: A survey. *Computer Vision and Image Understanding* **162**, 1–22 (2017)
13. Yang, J., Liu, Q., Zhang, K.: Stacked hourglass network for robust facial landmark localisation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 79–87 (2017)
14. Kowalski, M., Naruniec, J., Trzcinski, T.: Deep alignment network: A convolutional neural network for robust face alignment. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 88–97 (2017)
15. Fan, H., Zhou, E.: Approaching human level facial landmark localization by deep learning. *Image and Vision Computing* **47**, 27–35 (2016)
16. Bulat, A., Tzimiropoulos, G.: 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*, pp. 1021–1030 (2017)
17. Bulat, A., Tzimiropoulos, G.: Two-stage convolutional part heatmap regression for the 1st 3d face alignment in the wild (3dfaw) challenge. In: *European Conference on Computer Vision*, pp. 616–624 (2016). Springer
18. Asgarian, A., Zhao, S., Ashraf, A.B., Erin Browne, M., Prkachin, K.M., Mihailidis, A., Hadjistavropoulos, T., Taati, B.: 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, pp. 28–36 (2019)
19. Bandini, A., Rezaei, S., Guarin, D.L., Kulkarni, M., Derrick Lim, M., Boulos, M.I., Zinman, L., Yunusova, Y., Taati, B.: A new dataset for facial motion analysis in individuals with neurological disorders. *Journal of Biomedical and Health Informatics* (submitted)
20. Buolamwini, J., Gebru, T.: Gender shades: Intersectional accuracy disparities in commercial gender classification. In: Conference on Fairness, Accountability and Transparency, pp. 77–91 (2018)
21. Taati, B., Zhao, S., Ashraf, A.B., Asgarian, A., Browne, M.E., Prkachin, K.M., Mihailidis, A., Hadjistavropoulos, T.: Algorithmic bias in clinical populations—evaluating and improving facial analysis technology in older adults with dementia. *IEEE Access* **7**, 25527–25534 (2019)
22. Zhou, Z., Shin, J., Zhang, L., Gurudu, S., Gotway, M., Liang, J.: Fine-tuning convolutional neural networks for biomedical image analysis: Actively and incrementally. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
23. Beigman, E., Klebanov, B.B.: Learning with annotation noise. In: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, pp. 280–287 (2009). Association for Computational Linguistics
24. Paun, S., Carpenter, B., Chamberlain, J., Hovy, D., Kruschwitz, U., Poesio, M.: Comparing bayesian models of annotation. *Transactions of the Association for Computational Linguistics* **6**, 571–585 (2018)
25. Tsuchiya, M.: Performance impact caused by hidden bias of training data for recognizing textual entailment. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018) (2018)
26. Gururangan, S., Swayamdipta, S., Levy, O., Schwartz, R., Bowman, S., Smith, N.A.: Annotation artifacts in natural language inference data. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 107–112 (2018)
27. Santosa, K.B., Fattah, A., Gavilán, J., Hadlock, T.A., Snyder-Warwick, A.K.: Photographic standards for patients with facial palsy and recommendations by members of the sir charles bell society. *JAMA facial plastic surgery* **19**(4), 275–281 (2017)
28. Greene, J.J., Guarin, D.L., Tavares, J., Fortier, E., Robinson, M., Dusseldorp, J., Quatela, O., Jowett, N., Hadlock, T.: The spectrum of facial palsy: The meei facial palsy photo and video standard set. *The Laryngoscope* **130**(1), 32–37 (2020)
29. Weiss, K., Khoshgoftaar, T.M., Wang, D.: A survey of transfer learning. *Journal of Big data* **3**(1), 9 (2016)

Figures

Tables

Table 1 Mean $\pm$ standard deviation of the agreement between annotators as measured by the Euclidean distance between landmarks (in pixels) for each photograph. Smaller numbers indicate better agreement.

	Annotator 1	Annotator 3	Annotator 3
Annotator 1	0	—	—
Annotator 2	58.6 ± 22.5	0	—
Annotator 3	72.9 ± 27.3	69.0 ± 23.6	0

1					1
2					2
3					3
4					4
5					5
6					6
7					7
8					8
9					9
10					10
11					11
12					12
13	Table 2 Mean ± standard deviation of the NRMSE yielded by models fine-tuned with data provided				
14	by a single annotator (rows 1 to 3) and with data obtained by averaging each pair of annotators				
15	(rows 4 to 6) when compared to manual annotations provided by each annotator. Bolded results				
16	indicate that errors are significantly different ($p < 0.05$) than others presented in the same row.				
17					17
18					18
19					19
20					20
21					21
22					22
23					23
24					24
25					25
26					26
27					27
28					28
29					29
30					30
31					31
32					32
33					33

Model \ Data	Annotator 1	Annotator 2	Annotator 3
Annotator 1	7.9 ± 2.0	8.2 ± 2.2	8.7 ± 2.4
Annotator 2	8.2 ± 2.1	7.3 ± 1.9	8.6 ± 2.4
Annotator 3	9.0 ± 2.5	8.7 ± 2.6	8.3 ± 2.6
Annotator 1&2	7.6 ± 1.9	7.5 ± 2.0	8.5 ± 2.3
Annotator 1&3	8.0 ± 2.0	8.0 ± 2.3	7.9 ± 2.2
Annotator 2&3	8.2 ± 2.2	8.1 ± 2.2	8.3 ± 2.5

Figures

Figure 1

Normalized Root Mean Square Error (NRMSE) of facial landmark localization for the FAN model as a function of the number of patients used for fine-tuning the model when applied to photographs of patients with facial palsy.

Figure 2

Normalized Root Mean Square Error (NRMSE) of facial landmark localization for the FAN model as a function of the number of patients used for ne-tuning the model when photographs of near normal patients (blue), patients with moderate facial paralysis (green), and patients with complete facial paralysis (red).