Limitations and Biases in Facial Landmark Detection – An Empirical Study on Older Adults with Dementia

Azin Asgarian[1,2,3], Shun Zhao[3], Ahmed B. Ashraf[3,4], M. Erin Browne[5], Kenneth M. Prkachin[6], Alex Mihailidis[3,7,8], Thomas Hadjistavropoulos[5,9], and Babak Taati[3,2,7,10]

[1]Georgian Partners Inc, [2]Department of Computer Science, University of Toronto, [3]Toronto Rehabilitation Institute, University Health Network, [4]Department of Electrical and Computer Engineering, University of Manitoba, [5]Department of Psychology, University of Regina, [6]Department of Psychology, University of Northern British Columbia, [7]Institute of Biomaterials and Biomedical Engineering, University of Toronto, [8]Department of Occupational Science and Occupational Therapy, University of Toronto, [9]Centre on Aging and Health, University of Regina, [10]Vector Institute for Articial Intelligence

Abstract

Accurate facial expression analysis is an essential step in various clinical applications that involve physical and mental health assessments of older adults (e.g. diagnosis of pain or depression). Although remarkable progress has been achieved toward developing robust facial landmark detection methods, state-of-the-art methods still face many challenges when encountering uncontrolled environments, different ranges of facial expressions, and different demographics of population. A recent study has revealed that the health status of individuals can also affect the performance of facial landmark detection methods on front views of faces. In this work, we investigate this matter in a much greater context using seven facial landmark detection methods. We perform our evaluation not only on frontal faces but also on profile faces and in various regions of the face. Our results shed light on limitations of the existing methods and challenges of applying these methods in clinical settings by indicating: 1) a significant difference between the performance of state-of-the-art when tested on the profile or frontal faces of individuals with vs. without dementia; 2) insights on the existing bias for all regions of the face; and 3) the presence of this bias despite re-training/fine-tuning with various configurations of six datasets.

1. Introduction

Facial landmark detection is a prerequisite for many facial analysis applications. Example clinical use cases include detecting pain in non-communicative individuals, clinical assessment of depression, and orofacial and speech assessment in individuals with a neurological motor disorder [6, 30]. For a long time, active appearance models (AAM) were the method of choice for facial landmark detection [12]. In recent years, methods beyond AAM

have shown superior performance for landmark detection. Representative examples include Conditional Local Neural Fields [5], Coarse-to-Fine-Shape-Searching [35], Face Alignment Network [8], Mnemonic Descent Method [32], and Position Map Regression Network [14].

Despite the recent promising advances in this field, state-of-the-art methods still face many challenges when applied in realistic scenarios [9, 31]. To address these challenges, significant efforts have been made towards collecting images of faces in the wild (i.e. natural environment) and to cover the spectrum of age, gender, and ethnicity. However, recent studies have revealed that merely collecting more training data might not mitigate the effect of variables such as age, gender, and ethnicity [9, 31]. Hence, to develop algorithms that are fair with respect to potential biases, further research is required on the effect of different variables such as age, gender, and health conditions on the performance of facial landmark detection methods.

In a recent study, Taati et al. [31] have shown that cognitive ability (healthy vs. cognitive impairment) also affects the performance of facial landmark detection methods on frontal faces of older adults. In this paper, we experimentally examine the presence of such bias in a greater scope using seven facial landmark detection methods. Moreover, we perform our evaluation on profile faces as well as on frontal faces and for various regions of the face, i.e., jaw, brows, nose, eyes, and mouth. Additionally, to further evaluate the sources of bias, we assess the performance of these methods when re-trained/fine-tuned with various training configurations of six different datasets.

Our comprehensive evaluation shows that the performance of landmark detection methods drops on the frontal and profile faces of older people with dementia as compared to cognitively healthy older adults. It also indicates that the difference in the performance between the two groups is higher in some regions of the face, such as the mouth, the

eyes, and the nose, as compared to other regions such as the jaw and the brows. Moreover, our analysis shows that re-training/fine-tuning the methods improves the performance significantly on both groups, but the gap between the performance on individuals with and without dementia persists.

In the remainder of this paper, we first provide a brief overview of the datasets and landmark detection methods used in our evaluation in Sections 2 and 3 respectively. Sections 4 and 5 describe our experimental settings and results and Section 6 covers conclusions and future work.

2. Datasets

To conduct the experiments in this paper, we used the following six datasets: Helen [23], AFW [28], LFPW [7], MENPO Profile [34], UNBC-McMaster Pain Archive [25], and Pain Dataset for Dementia [16]. The MENPO Profile and Pain Dataset for Dementia include both frontal and profile faces, while the remaining four datasets only contain frontal and semi-frontal faces. The role of each dataset in each experiment (i.e. training or test) is described in §4. An overview of these datasets is provided below.

Helen: This dataset is constructed by crawling 2,330 face images from Flickr using keywords such as "family", "outdoor", "boy" etc. The faces were cropped and manually annotated using the PUT Face [20] 194 landmark points.

Annotated Faces in the Wild (AFW): This dataset is also collected from Flickr images and consists of 468 faces [28]. The images of this dataset come along with annotations for six landmark points.

Labeled Face Parts in the Wild (LFPW): This database consists of 3,000 faces downloaded from the web using search queries (Google, Yahoo, Flickr). Annotations include 29 facial landmark points.

Since the landmark annotation for the above three datasets did not use a consistent set of points, Sagonas et al. [29] later re-annotated a subset of examples from these datasets using a standard set of 68 landmark points[15] shown in Figure 1(a). From this consistently annotated subset we use 3,148 images (2,000 from Helen, 337 from AFW, and 811 from LFPW). The majority of images in these datasets are from young people and children with happy or neutral expressions. In the remainder of this paper we refer to the union of these three datasets as "Source 1" (S_1).

MENPO Profile: This dataset contains 2,300 profile images obtained from the FDDB [19] and AFLW [22] databases and re-annotated using 39 profile view landmark points (Figure 1(b). We denote this dataset with M_p.

UNBC-McMaster Pain Archive: The publicly available part of this dataset consists of 48,398 face images from 25 participants [24]. Participants in this dataset had a shoulder injury in one of their shoulders. During data collection, par-

ticipants were asked to move their injured shoulder in one session, and their healthy shoulder in another session and their videos were recorded. Each image is annotated with the location of 68 facial landmarks, and also with the level of pain expressed in each image. Pain is coded using a 0 to 16 pain scale [27] based on the Facial Action Coding System (FACS) [24], where 0 indicates no pain and 16 indicates the highest level of pain observed.

Using the FACS-based pain ratings, we subsampled the UNBC-McMaster dataset to 2,951 images while preserving the same distribution of pain ratings as the full dataset. In the rest of this paper we denote this subset of the UNBC-McMaster archive as "Source 2" (S_2).

Pain Dataset for Dementia: This dataset contains data from 102 older adult participants [16] (mean age: 78.8) with and without dementia. From this dataset, Taati et al. [31] selected data from 86 participants based on the availability of high-quality images. Of these 86 older adults, 44 were cognitively healthy and 42 were living in long-term care facilities with various degrees of dementia. Each participant was video recorded during a baseline state when lying flat on a bed, and also an exam state in which a licensed physiotherapist assisted the participant to execute a sequence of movements to identify painful areas [18]. Each session was filmed with three cameras, one capturing the frontal view and two capturing the side views (right and left). The entire dataset was annotated manually for the level of pain by trained pain coders using a FACS-based pain rating [27] and a PACSLAC-II pain rating [10]; clinically validated methods to score pain in individuals with severe dementia [16].

We used two subsets of this data in our experiments which we denote by "Target:Frontal" (T_f) and "Target:Profile" (T_p). To construct "Target:Frontal", we subsampled 688 frontal view images from the 86 participants. To ensure the existence of expressions corresponding to various levels of pain for each person, images of the exam state were clustered into 7 groups based on the level of pain expressed and one image was chosen at random from each group. Also, to account for the existence of neutral expressions, one image from each participant was selected at random from the baseline state. All images were manually rotated when needed to place the face in an upright position and were manually annotated using the standard 68 landmark points. Similarly, to build "Target:Profile", 679 profile view images were sub-sampled and manually annotated using the 39 landmark points shown in Figure 1(b).

3. Landmark Detection Methods

The following methods (and models) were used in our analysis: Active Appearance Models (AAM) [11], Constrained Local Neural Field (CLNF) [5], Coarse-to-Fine Shape Searching (CFSS) [35], Face Alignment Network

(FAN) [8], Mnemonic Descent Method (MDM) [32], and Position Map Regression Network (PRNet) [14]. In the following we briefly review these methods.

Active Appearance Models (AAM): An AAM [11] is a generative model that captures variations of shape and appearance of a deformable object from a set of labeled images. The model thus has two components, one for shape, and another for appearance. To train the AAM model, first Procrustes analysis is applied on training data and then PCA is performed on the shape labels and image pixels, to build the shape and appearance models. During fitting, the AAM initializes from the mean shape and tries to find the best set of parameters that minimizes the difference between the input image and the reconstructed image (based on shape and appearance parameters).

Constrained Local Neural Field (CLNF): A CLNF by Baltruaitis et al. [5] is an instance of the Constrained Local Model (CLM) [13] that incorporates Local Neural Field patch experts. Local Neural Field patch experts are applied on the landmark areas to learn non-linear relationships of the pixels around the landmark. Similar to AAM, the CLNF also has a shape component that models the location of the landmark points as a combination of a mean shape and a set of transformations. During fitting, the CLNF model tries to find the best set of transformation parameters that optimizes the patch expert responses while taking the reliability of each patch expert into account.

Coarse-to-Fine Shape Searching (CFSS): Unlike many facial landmark detection methods that require an initial shape (usually the mean shape) to start the fitting process, Zhu et al. [35] proposed CFSS which initializes searching from a shape space. A CFSS builds a large space of candidate shapes and performs face alignment in a given number of stages. The model starts searching by sampling from a large region in the shape space and estimates a finer subregion to perform searches in the later stages. The adaptive stage-by-stage approach prevents the model from being trapped in local optima due to poor initialization.

Face Alignment Network (FAN): The FAN model, proposed by Bulat et al. [8], regresses landmark heatmaps directly. To regress the 2D landmarks, FAN-2D employs a stack of four hourglass (HG) networks [26] and trains them with RGB images as input, and 68 2D Gaussian heatmaps as target output, one for each of the 68 facial landmark points. A FAN-3D network is jointly trained with an additional 2D-to-3D FAN network, where FAN-3D predicts the 2D projection of the 3D landmark points and the 2D-to-3D FAN estimates the corresponding z coordinates for the 2D landmark points predicted by FAN-3D. In this work we fine-tuned FAN-2D with everything but the last hourglass network frozen, which we refer to as FFAN-HG.

Mnemonic Descent Method (MDM): Trigeorgis et al. proposed MDM [32], which is an end-to-end face alignment model; i.e., it predicts the landmark coordinates directly from raw image pixels. Instead of more traditional hand-crafted features such as HOG or SIFT, MDM learns a two layer Convolutional Neural Network (CNN) as the feature extractor. For fitting, the MDM model employs the idea of learning descent directions [33] with an additional RNN component that learns information about the past descent directions during training and then uses this information in the fitting process.

Position Map Regression Network (PRNet): The PRNet model proposed by Feng et al. [14] employs a Neural Network architecture that contains Residual Blocks [17] and convolutional layers to simultaneously reconstruct the 3D facial structure and perform facial landmark alignment. For training, ground truth 3D facial shapes are first projected into UV space (a 2D image representation of 3D coordinates) and then the obtained UV images are used to train the model. For fitting, the PRNet model first predicts the UV images from the input image and then 3D facial structure and aligned facial landmarks are derived from the predicted UV images.

4. Experiments

For fair and comprehensive evaluation, we consider four different experiments. In the first three experiments we compare the performance of all methods on the cognitively healthy older adult subset ($44 \times 8 = 352$ images) vs. the dementia subset ($42 \times 8 = 336$ images) of the T_f dataset. In the last experiment, we use the healthy subset (338 images) and the dementia subset (325 images) of the T_p dataset for evaluation. The training set configurations explored in each experiment are described below. In any configuration that involved training examples from T_f and T_p, leave-one-participant-out cross-validation was employed to ensure images from the same person did not appear in both training and test data.

4.1. Experimental Settings

Experiment 1: In this experiment, we used the off-the-shelf versions of the following seven methods: CLNF [5], CFSS [35], AAM [1], FAN-2D and FAN-3D [8], MDM [32], and PRNet [14]. Many groups offer pre-trained AAM models which are usually trained on S_1. For consistency, we also trained the the AAM model on S_1.

Experiment 2: In the second experiment, we evaluated different models when re-trained/fine-tuned with T_f. Models AAM, CLNF, and CFSS were re-trained with T_f. However, since T_f was significantly smaller than the original dataset used to train model FAN-2D, a fine-tuned version of this model with T_f (which we call FFAN-HG) was included in

this experiment. Models MDM, FAN-3D, and PRNet were excluded due to unavailability of the training code and lack of 3D ground truth landmark annotations for images in T_f.

Experiment 3: In our third experiment, we evaluated various methods when re-trained with the following configurations: $S_1, S_2, S_1 \cup S_2, T_f \cup S_1, T_f \cup S_2, T_f \cup S_1 \cup S_2$. In addition to methods MDM, FAN-3D, and PRNet, method FAN-2D was also excluded from this experiment as it is originally trained on a super set of S_1.

Experiment 4: The off-the-shelf versions of methods FAN-2D, FAN-3D, and PRNet work on profile faces; therefore, they were included in this experiment. However, the rest of the methods only work on frontal and semi-frontal faces and need re-training to handle profile faces. We re-trained model AAM with configurations $T_p, M_p, T_p \cup M_p$. But considering the size of these training configurations, re-training was not an option for the rest of the methods and thus they were excluded from this experiment.

4.2. Evaluation

We compare the performance of different methods on the cognitively healthy older adult subset (H) versus the dementia subset (D) of T_f and T_p in terms of the convergence rate. To measure the convergence rate, we use its standard definition in the literature [2, 8, 32] as the percentage of test examples that converge to the ground truth landmark points given a tolerance in the root mean squared (RMS) fitting error (here, 5% of the face diagonal).

We also show convergence curves that plot the percentage of test examples converged to the ground truth as a function of tolerance in RMS fitting error (normalized by the face diagonal). A typical comparison point is the point on the curve corresponding to 5% tolerance. We perform this evaluation for the landmark points spanning the whole face and also for points that lie in specific regions i.e., jaw, brows, nose, eyes, and mouth. To evaluate statistical significance, we use the non-parametric Wilcoxon rank-sum test (on RMS errors) and consider three standard significance levels 0.05, 0.01 and 0.001.

To ensure a fair comparison, for each image in T_f and T_p, the same face bounding box (detected by the Dlib face detector [21]) was used to initialize all the models. Results in Experiments 1-3 are evaluated using the 68 landmark points. In Experiment 4, the AAM model gives 39 landmark points while the rest of the methods output the standard set of 68 landmark points [15]. The two sets of landmark points are shown in Figures 1(b) and 1(a). Since there is not a one to one correspondence between all the points in these markups, results in Table 3 and Figure 5 are evaluated on the 25 points that are common between the two mark-ups from all regions of the face except the jaw line. These 25 points are shown in blue in Figure 1(b).
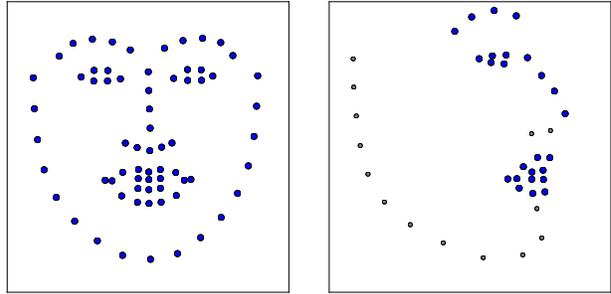
(a) 68 landmark points (frontal)   (b) 39 landmark points (profile)

Figure 1. Different sets of landmark points used in the evaluations.

5. Results

The convergence rates obtained for all regions of the face with the methods explored in experimental settings 1, 2, and 4 on healthy (H) and dementia (D) subsets of T_f and T_p are shown respectively in Tables 1, 2, and 3. The results of Wilcoxon rank-sum tests that evaluate the statistical significance of difference between the performance on healthy (H) and dementia (D) subsets of T_f and T_p are also reported for all methods and regions of the face.

Figures 2, 3, and 5 show the average convergence curves obtained on healthy (H) and dementia (D) subsets of T_f and T_p using different methods from experiments 1, 2 and 4 respectively. In these figures, the x-axis shows the RMS fitting error normalized by the face size (diagonal), while the y-axis shows the percentage of cases with a fitting error less than the corresponding x-axis value averaged over all methods included in the evaluation. Figure 4 shows the convergence rates obtained on the 68 landmark points (whole face) for healthy (H) and dementia (D) subsets of T_f using the re-trained versions of methods AAM, CFSS, and CLNF on the training configurations of experiment 3.

The general trend in Experiments 1-3 show that the relationship between convergence rates of all evaluated methods and dementia is significant on frontal faces (T_f). Although increasing the variation in the training data by including images from various datasets (T_f, S_1, and S_2) improves the performance on both healthy and dementia subsets of T_f, the difference between the convergence rates for these two subsets remains large and statistically significant (Experiments 2-3). Results of Experiment 4 show a similar trend on profile faces (T_p) with less difference between the convergence rates obtained for healthy and dementia subsets when compared to frontal faces (T_f).

From Experiment 1 (Table 1 and Figure 2), we can see that the difference in convergence rates obtained on the whole face between healthy and dementia subsets of T_f is large and statistically significant for every one of the seven methods evaluated. Figure 2 shows that for all regions of the face the convergence curves for the healthy subset lie
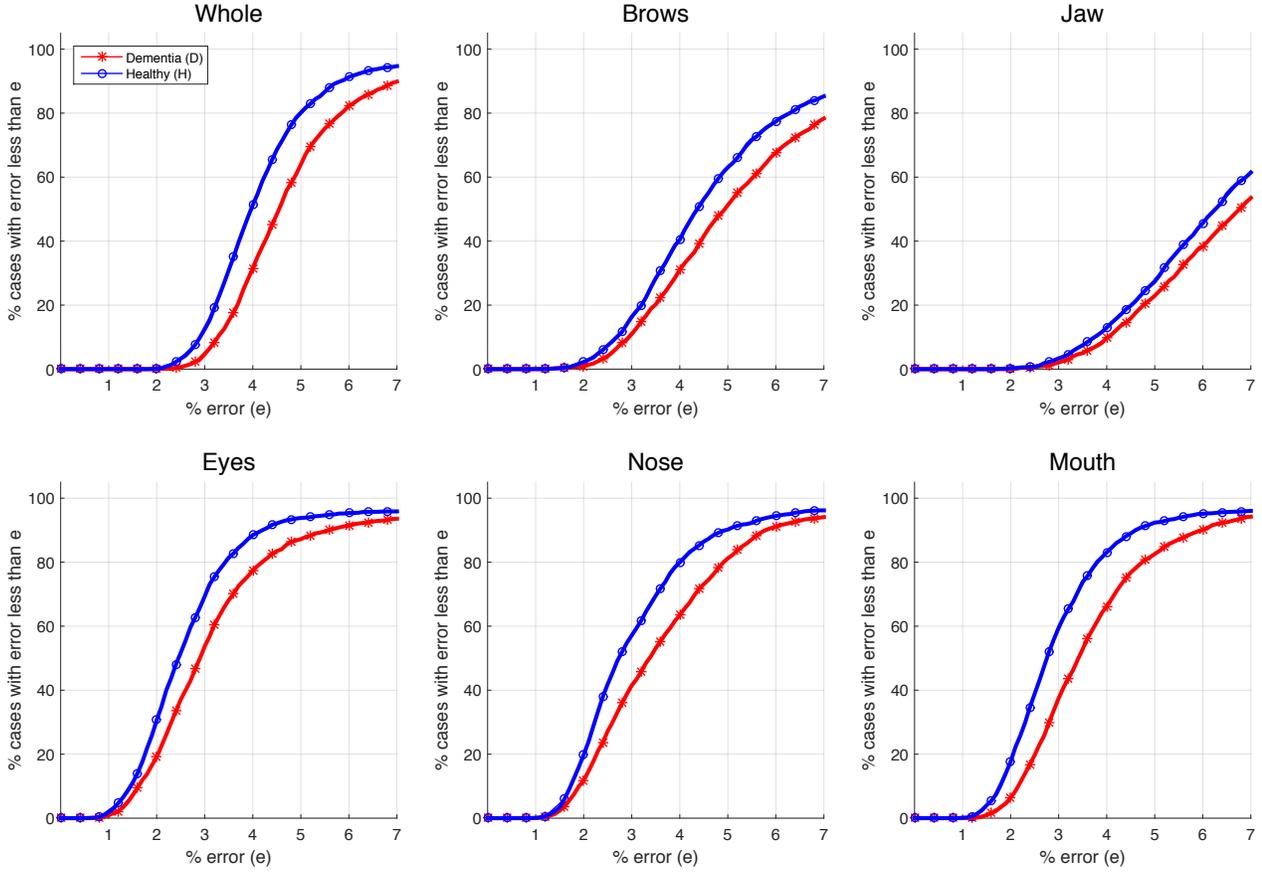
Figure 2. Experiment 1: Comparison of the average convergence curves obtained on healthy subset (H) and dementia subset (D) of T_f. The values on y-axis are averaged over seven methods: CLNF, CFSS, AAM, FAN-2D, FAN-3D, MDM, and PRNet.

Table 1. Experiment 1: Comparison of convergence percentage within 5% tolerance of RMS fitting error obtained on healthy subset (H) and dementia subset (D) of T_f. p-values are color coded with respect to three standard significant levels 0.05, 0.01 and 0.001.

Methods	Whole		Jaw		Brows		Nose		Eyes		Mouth	
	H	D	H	D	H	D	H	D	H	D	H	D
CLNF	71.88	63.39	24.15	26.49	50.00	42.26	84.94	77.38	85.51	72.32	84.66	75.60
	$p < 0.001$		$p = 0.367$		$p = 0.078$		$p < 0.001$		$p < 0.001$		$p < 0.001$	
CFSS	80.11	65.77	27.27	27.98	61.08	50.00	90.63	88.69	85.51	77.38	90.34	79.17
	$p < 0.001$		$p = 0.848$		$p < 0.001$		$p < 0.001$		$p < 0.001$		$p < 0.001$	
AAM	87.78	71.73	44.89	29.76	62.50	44.94	95.45	95.24	94.60	87.80	93.75	86.31
	$p < 0.001$		$p < 0.001$		$p < 0.001$		$p < 0.001$		$p < 0.001$		$p < 0.001$	
FAN-2D	81.53	61.90	19.89	19.94	68.18	56.25	83.52	63.39	99.43	96.43	97.44	85.12
	$p < 0.001$		$p = 0.095$		$p = 0.065$		$p < 0.001$		$p < 0.001$		$p < 0.001$	
FAN-3D	71.88	50.60	18.18	12.20	61.93	52.98	85.80	65.48	98.86	94.64	98.58	89.88
	$p < 0.001$		$p < 0.001$		$p = 0.013$		$p < 0.001$		$p < 0.001$		$p < 0.001$	
MDM	91.76	70.24	36.36	25.30	65.06	47.02	97.16	90.18	97.44	91.96	95.74	85.42
	$p < 0.001$		$p < 0.001$		$p < 0.001$		$p < 0.001$		$p < 0.001$		$p < 0.001$	
PRNet	76.14	64.29	20.74	18.75	72.73	65.18	93.75	88.10	95.45	89.88	86.08	76.49
	$p < 0.001$		$p = 0.008$		$p = 0.004$		$p < 0.001$		$p < 0.001$		$p < 0.001$	
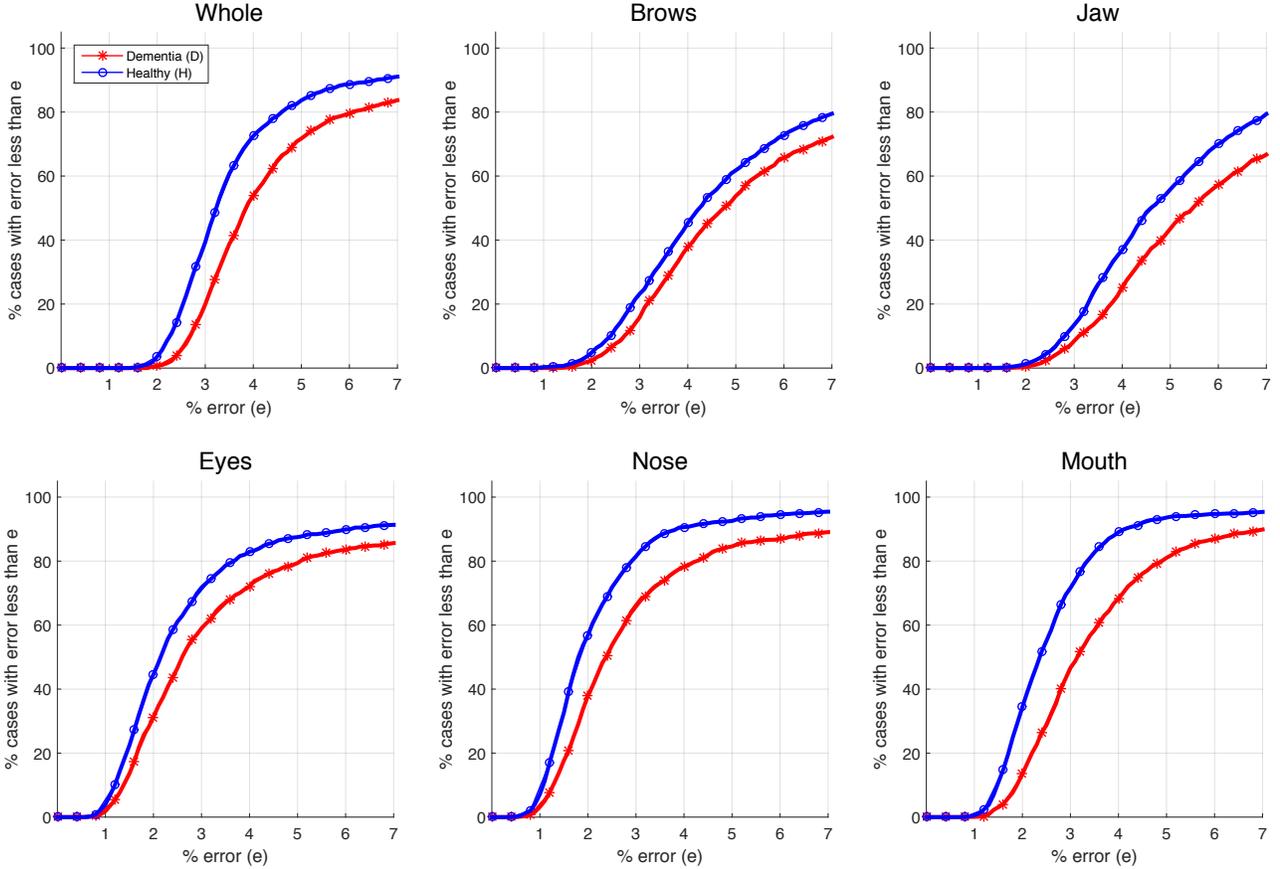
Figure 3. Experiment 2: Comparison of the average convergence curves obtained on healthy subset (H) and dementia subset (D) of T_f. The values on y-axis are averaged over four methods: CLNF, CFSS, AAM, and FFAN-HG.

Table 2. Experiment 2: Comparison of convergence percentage within 5% tolerance of RMS fitting error obtained on healthy subset (H) and dementia subset (D) of T_f. p-values are color coded with respect to three standard significant levels 0.05, 0.01 and 0.001.

Methods	Whole		Jaw		Brows		Nose		Eyes		Mouth	
	H	D	H	D	H	D	H	D	H	D	H	D
CLNF	87.78	75.89	65.91	48.51	75.57	63.39	90.63	77.68	91.19	81.55	91.19	80.65
	$p < 0.001$		$p < 0.001$		$p < 0.001$		$p < 0.001$		$p < 0.001$		$p < 0.001$	
CFSS	56.53	36.9	24.43	13.1	34.09	28.87	82.67	70.83	65.91	57.14	87.78	65.48
	$p < 0.001$		$p < 0.001$		$p = 0.018$		$p < 0.001$		$p < 0.001$		$p < 0.001$	
AAM	94.32	83.93	62.22	48.51	69.89	59.82	98.86	92.86	98.01	91.37	97.16	84.52
	$p < 0.001$		$p < 0.001$		$p = 0.004$		$p < 0.001$		$p < 0.001$		$p < 0.001$	
FFAN-HG	96.02	90.18	70.74	63.99	67.61	63.1	98.01	97.02	94.89	87.5	98.3	93.45
	$p < 0.001$		$p = 0.003$		$p = 0.003$		$p < 0.001$		$p < 0.001$		$p < 0.001$	

above the convergence curves for the dementia subset. We also notice that the difference between convergence curves is larger in the mouth, eyes, and nose regions of the faces as compared to the brows and the jaw. This has implications in applications where the tracking of the mouth, eyes, or nose regions is important, e.g., in the detection of pain [27].

Table 2 and Figure 3 show the performance of re-trained/fine-tuned versions of methods CLNF, CFSS, AAM and FFAN-HG with images from T_f on various regions of

the face. Comparing the results reported in Tables 2 and 1, we see that the performance for all methods except CFSS has largely increased on both healthy and dementia subsets after including images from T_f in the training data. This is possibly because of the searching mechanism used in the CFSS model and the significant difference between the size of T_f and the data used originally to train it.

Although we see a boost in the convergence curves for most regions when comparing Figure 3 to Figure 2, the con-
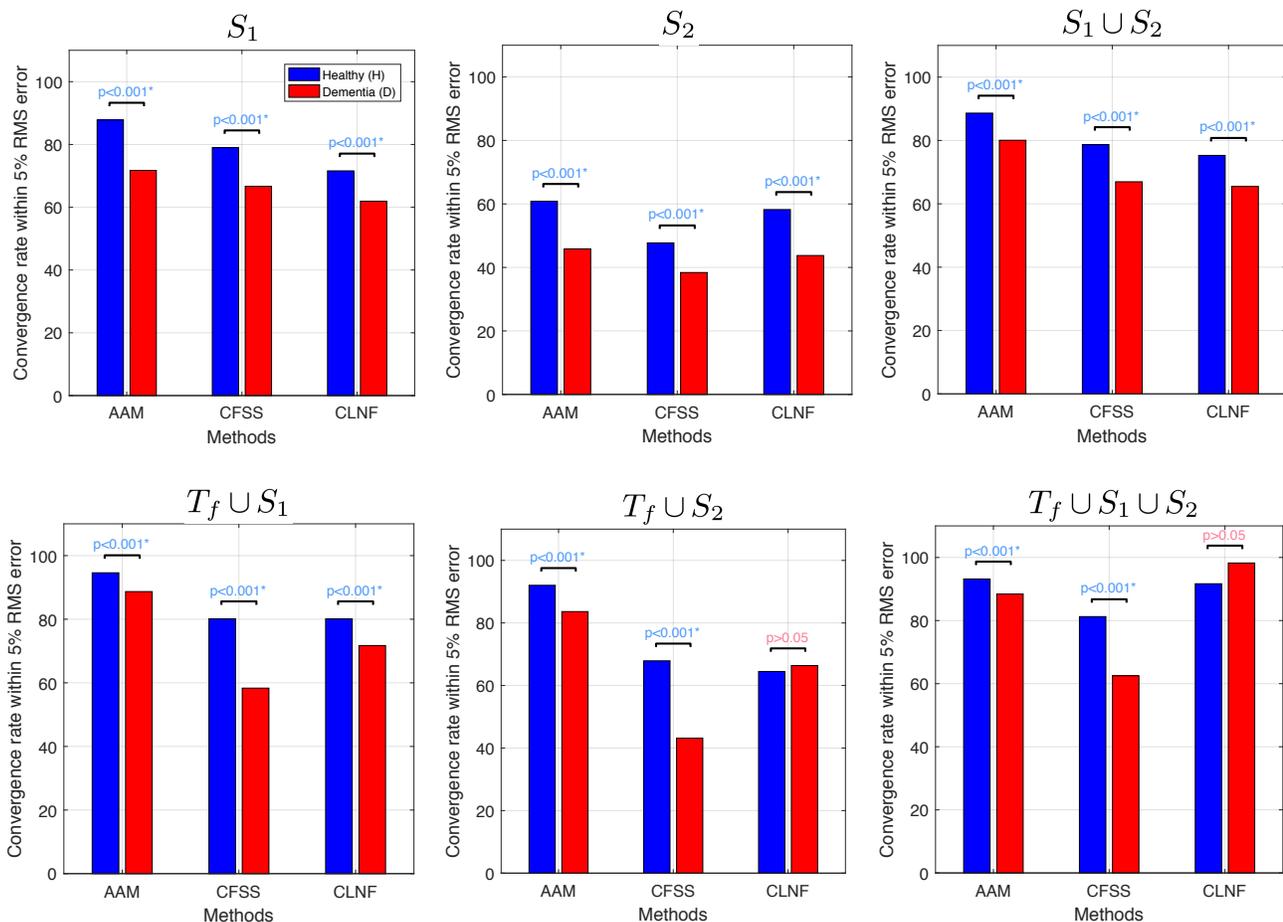
Figure 4. Experiment 3: Comparison of convergence percentage within 5% tolerance of RMS fitting error obtained on healthy subset (H) and dementia subset (D) of T_f using various versions of three methods AAM, CFSS, and CLNF trained on configurations $S_1, S_2, S_1 \cup S_2, T_f \cup S_1,\ T_f \cup S_2,\ T_f \cup S_1 \cup S_2$. RMS fitting errors are computed over the standard 68 landmark points (whole face).

vergence rates for the dementia subset are still lower compared to those for the healthy subset of T_f and the difference is significant for all regions of the face (Table 2). This trend is particularly noticeable in the jaw and in the eyes.

Figure 4 shows the convergence rates obtained on healthy (H) and dementia (D) subsets of T_f using the retrained versions of methods AAM, CFSS and CLNF on the following training configurations: $S_1, S_2, S_1 \cup S_2, T_f \cup S_1,\ T_f \cup S_2,\ T_f \cup S_1 \cup S_2$. We see that the convergence rates for healthy and dementia subsets vary largely by configuration; however, the difference between them remains significant for all configurations and methods (except for method CLNF when trained on $T_f \cup S_2$ and $T_f \cup S_1 \cup S_2$). A similar trend was also observed in the convergence rates for all regions of the face (included in the supplementary materials). Comparing the results of Experiments 1, 2 and 3, we notice that the inclusion of additional variation in the training data can help to improve the performance in general, but it does not help with mitigating the gap between

the performance on healthy and dementia subsets.

Table 3 and Figure 5 show the performance of AAM, FAN-2D, FAN-3D, and PRNet when evaluated on the profile face of T_p. Performance is poor as compared to performance on T_f; but, similar to the previous experiments, we see that the average convergence curves for the healthy subset lie above the curves for the dementia subset in all regions of the face. The difference between the performance on healthy and dementia subsets of T_p is smaller compared to the ones for frontal faces in T_f, yet it is significant on some regions of the face such as the nose and the mouth.

6. Conclusions

Accurate detection of facial landmark points is an important requirement for a wide range of clinical applications involving older adults and/or individuals with a cognitive or a physical disability. In this paper, we provide a comprehensive evaluation of state-of-the-art facial landmark detection on faces of older adults with and without dementia.
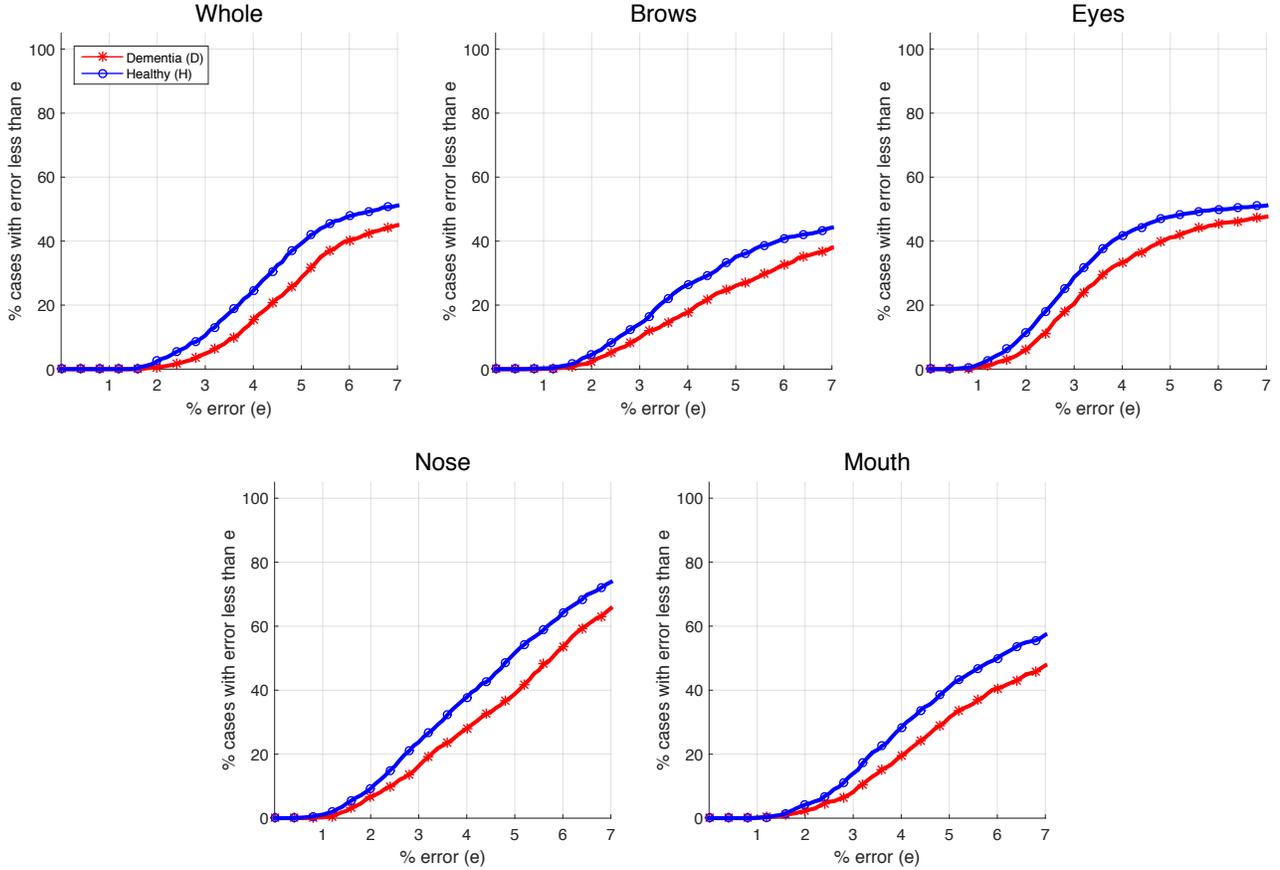
Figure 5. Experiment 4: Comparison of the average convergence curves obtained on healthy subset (H) and dementia subset (D) of T_p. The values on y-axis are averaged over four methods: AAM, FAN-2D, FAN-3D, and PRNet.

Table 3. Experiment 4: Comparison of convergence percentage within 5% tolerance of RMS fitting error obtained on healthy subset (H) and dementia subset (D) of T_p. p-values are color coded with respect to three standard significant levels 0.05, 0.01 and 0.001.

Methods	Whole		Brows		Nose		Eyes		Mouth	
	H	D	H	D	H	D	H	D	H	D
AAM	44.67	37.23	32.84	21.54	51.18	46.15	47.93	44	52.66	37.54
	$p < 0.001$		$p = 0.015$		$p = 0.034$		$p = 0.010$		$p < 0.001$	
FAN-2D	35.21	26.15	39.64	32.31	35.21	21.85	49.41	44	35.5	29.23
	$p = 0.193$		$p = 0.817$		$p < 0.001$		$p = 0.268$		$p = 0.173$	
FAN-3D	36.09	22.15	36.39	28.31	43.2	23.08	49.41	41.23	38.76	29.23
	$p = 0.043$		$p = 0.709$		$p < 0.001$		$p = 0.143$		$p = 0.030$	
PRNet	41.12	28.92	31.66	22.15	76.92	64.31	43.79	35.38	36.98	29.85
	$p = 0.192$		$p = 0.951$		$p < 0.001$		$p = 0.553$		$p = 0.082$	

Our evaluation demonstrates an algorithmic bias in state-of-the-art facial landmark detection methods, which affects the performance of these methods for older adults with dementia. Furthermore, our empirical analysis shows that techniques such as fine-tuning and re-training can improve the performance for both groups; however, these methods cannot reduce the gap between the performance for adults with and without dementia. As interest in employing facial analysis methods in clinical applications grows [3, 4], our study sheds light on the limitations of existing facial landmark detection methods and the challenges of applying these methods to clinical populations. In future work, we plan to investigate potential solutions to overcome these biases in facial landmark detection methods.

References

[1] 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 *ACM International Conference on Multimedia*, 2014. 3

[2] A. Asgarian, A. B. Ashraf, D. Fleet, and B. Taati. Subspace selection to suppress confounding source domain information in aam transfer learning. In *IJCB*, 2017. 4

[3] A. Asgarian, P. Sobhani, J. C. Zhang, M. Mihailescu, A. Sibilia, A. B. Ashraf, and B. Taati. A hybrid instance-based transfer learning method. *arXiv preprint arXiv:1812.01063*, 2018. 8

[4] A. Ashraf and B. Taati. Automated video analysis of hand-washing behavior as a potential marker of cognitive health in older adults. *IEEE JBHI*, 2016. 8

[5] T. Baltrusaitis, P. Robinson, and L.-P. Morency. Constrained local neural fields for robust facial landmark detection in the wild. In *ICCV Workshops*, pages 354–361, 2013. 1, 2, 3

[6] 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*, 281(Supplement C):7 – 20, 2017. 1

[7] P. N. Belhumeur, D. W. Jacobs, D. J. Kriegman, and N. Kumar. Localizing parts of faces using a consensus of exemplars. *PAMI*, 35(12):2930–2940, 2013. 2

[8] A. Bulat and G. Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In *ICCV*, 2017. 1, 3, 4

[9] J. Buolamwini and T. Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *IEEE FAT*, 2018. 1

[10] S. Chan, T. Hadjistavropoulos, J. Williams, and A. Lints-Martindale. Evidence-based development and initial validation of the pain assessment checklist for seniors with limited ability to communicate-ii (pacslac-ii). *The Clinical journal of pain*, 30(9):816–824, 2014. 2

[11] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. In *ECCV*, 1998. 2, 3

[12] T. F. Cootes, G. J. Edwards, C. J. Taylor, et al. Active appearance models. *IEEE Transactions on PAMI*, 2001. 1

[13] D. Cristinacce and T. F. Cootes. Feature detection and tracking with constrained local models. In *Bmvc*, volume 1, page 3. Citeseer, 2006. 3

[14] Y. Feng, F. Wu, X. Shao, Y. Wang, and X. Zhou. Joint 3d face reconstruction and dense alignment with position map regression network. In *ECCV*, 2018. 1, 3

[15] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker. Multi-pie. *Image Vision Comput.*, 2010. 2, 4

[16] T. Hadjistavropoulos, M. Browne, K. Prkachin, B. Taati, A. Ashraf, and A. Mihailidis. Pain in severe dementia: A comparison of a fine-grained assessment approach to an observational checklist designed for clinical settings. *European Journal of Pain*, 22(5):915–925, 2018. 2

[17] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 3

[18] B. S. Husebo, L. I. Strand, R. Moe-Nilssen, S. B. Husebo, A. L. Snow, and A. E. Ljunggren. Mobilization-observation-behavior-intensity-dementia pain scale (mobid): development and validation of a nurse-administered pain assessment tool for use in dementia. *Journal of pain and symptom management*, 34(1):67–80, 2007. 2

[19] V. Jain and E. Learned-Miller. Fddb: A benchmark for face detection in unconstrained settings. Technical report, University of Massachusetts, Amherst, 2010. 2

[20] A. Kasinski, A. Florek, and A. Schmidt. The put face database. *Image Processing and Communications*, 2008. 2

[21] D. E. King. Dlib-ml: A machine learning toolkit. *JMLR*, 10(Jul):1755–1758, 2009. 4

[22] 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 ICCV Workshops)*, pages 2144–2151. IEEE, 2011. 2

[23] V. Le, J. Brandt, Z. Lin, L. Bourdev, and T. S. Huang. Interactive facial feature localization. In *ECCV*, 2012. 2

[24] P. Lucey, J. F. Cohn, K. M. Prkachin, P. E. Solomon, S. Chew, and I. Matthews. Painful monitoring: Automatic pain monitoring using the unbc-mcmaster shoulder pain expression archive database. *Image and Vision Computing*, 2012. 2

[25] 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 *Automatic Face and Gesture Recognition and Workshops*, pages 57–64. IEEE, 2011. 2

[26] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, 2016. 3

[27] K. M. Prkachin and P. E. Solomon. The structure, reliability and validity of pain expression: Evidence from patients with shoulder pain. *Pain*, 139(2):267–274, 2008. 2, 6

[28] D. Ramanan and X. Zhu. Face detection, pose estimation, and landmark localization in the wild. In *CVPR*, 2012. 2

[29] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. A semi-automatic methodology for facial landmark annotation. In *CVPR Workshops*, pages 896–903, 2013. 2

[30] A. Stuhrmann, T. Suslow, and U. Dannlowski. Facial emotion processing in major depression: a systematic review of neuroimaging findings. *Biology of Mood & Anxiety Disorders*, 1(1):10, Nov 2011. 1

[31] 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*, 2019. 1, 2

[32] G. Trigeorgis, P. Snape, M. A. Nicolaou, E. Antonakos, and S. Zafeiriou. Mnemonic descent method: A recurrent process applied for end-to-end face alignment. In *CVPR*, pages 4177–4187, 2016. 1, 3, 4

[33] X. Xiong and F. De la Torre. Supervised descent method and its applications to face alignment. In *CVPR*, 2013. 3

[34] S. Zafeiriou, G. Trigeorgis, G. Chrysos, J. Deng, and J. Shen. The menpo facial landmark localisation challenge: A step towards the solution. In *CVPR Workshops*, 2017. 2

[35] S. Zhu, C. Li, C. C. Loy, and X. Tang. Face alignment by coarse-to-fine shape searching. In *CVPR*, pages 4998–5006. IEEE Computer Society, 2015. 1, 2, 3