2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

Facial Landmark Tracking in Videos of Individuals with Neurological Impairments: Is There a Trade-off Between Smoothness and Accuracy?

Leif E. R. Simmatis and Yana Yunusova

Abstract— **Orofacial kinematics are valuable markers of function and progression in a variety of neurological disorders. Recent advances in facial landmark detection have been used to improve landmark tracking in video, for example by accounting for interframe optical flow. It has been demonstrated that finetuning (a type of transfer learning) can improve the performance of some facial landmark detection systems. Here, we asked whether a neural network model that is pretrained using video data (supervision by registration, SBR) can be finetuned to improve landmark detection and tracking, using data from the Toronto Neuroface Dataset (n=36), which comprises 3 different clinical populations. We finetuned the supervision by registration (SBR) model using data from 3 individuals from each of 3 clinical populations (n=9), with or without neurological impairments. The remaining individuals from our dataset (n=27) were used for evaluation. Finetuning SBR moderately improved the model's accuracy but substantially increased the smoothness of tracked landmarks. This suggests that finetuning on video-trained models, like SBR, could improve the estimation of orofacial kinematics in individuals with neurological impairments. This could be used to improve the detection and characterization of neurological diseases using video data.**

Clinical Relevance— **This work demonstrated that transfer learning applied to video-trained facial landmark detectors could improve the measurement of orofacial kinematics in individuals with neurological impairments.**

I. Introduction

Orofacial kinematics provide an important window into the state of muscle function in neurological diseases like amyotrophic lateral sclerosis (ALS), Parkinson's disease, Bell's Palsy, or stroke [1], [2] and are used for diagnostic purposes, patient stratification, and disease tracking and prediction. Lab-based objective instrumental assessment techniques used for kinematic assessments employ reflective markers or sensor coils to track facial landmarks [3], [4]. While precise, these systems are complex and expensive, which has limited their clinical adoption.

Computer vision and deep learning have made it possible to quantify orofacial movements from video without complex hardware; however, presently, these systems have some substantial limitations related to video-based facial landmark tracking in clinical populations. Two of these problems are landmark jitter and algorithmic bias. Landmark jitter is unwanted noise in the time series of facial landmark positions arising from random error of landmark estimation, as opposed to true physiological movement. Recent work has sought to correct this by developing models that are trained using videos [5]; however, these techniques have not been applied to clinical data previously to the best of our knowledge. Smoother landmark tracking would require less filtering, which can adversely impact the estimation of kinematics [6], and could improve the measurement of articulatory acceleration and jerk, which can provide an advantage in distinguishing healthy individuals from those with a disease [7].

Algorithmic bias is the observation that computer vision facial landmark detectors have less precision when identifying facial landmarks in populations on which these models have not been trained, specifically older individuals or those with neurological impairments. Previous work has identified that various facial landmark detection systems have significantly higher error on individuals with neurological disorders, compared to healthy older individuals [8]. Finetuning (a type of transfer learning) has been explored as a means to reduce the extent of this bias [8], [9].

Facial landmark tracking using deep neural networks can be done accurately using popular models such as the facial alignment network (FAN) [10]. More recently, models such as supervision-by-registration (SBR) have been developed that incorporate video-based training signals such as optical flow [5], which may improve smoothness in landmark tracking in video. Previous work has demonstrated that finetuning using relevant clinical data can dramatically improve FAN accuracy [9]; to our knowledge, the impact of finetuning on accuracy and smoothness using SBR has not been evaluated, nor how it compares to finetuning a non-video optimized (i.e., FAN) model.

In the present study, we finetuned SBR using data from individuals from clinical populations and compared its accuracy and smoothness of tracking to FAN and its finetuned version. This builds upon a previous analysis of the SBR algorithm's tracking performance [11]. We hypothesized that finetuning SBR would lead to improvements in video facial landmark tracking compared to both 1) the non-finetuned SBR model, and 2) a finetuned facial landmark detector [9] that is not optimized for video-based landmark tracking.

*Research supported in part by each of: AGE-WELL NCE, Canadian Partnership for Stroke Recovery, National Institutes of health, Natural Sciences and Engineering Research Council, and Michael J. Fox Foundation.

Leif E. R. Simmatis is with Toronto Rehabilitation Institute, Toronto, ON, Canada (e-mail: leif.simmatis@uhn.ca).

Yana Yunusova is with the University of Toronto Department of Speech Language Pathology, Toronto, ON, Canada (e-mail: yana.yunusova@utoronto.ca); University Health Network – KITE; Sunnybrook Research Institute

II. METHODS

A. Participants and clinical assessment

Individuals diagnosed with amyotrophic lateral sclerosis (ALS) (n=11, median age=62), individuals who had strokes (n=14, median age=64), and healthy older adults (n=11, median age=65) were recruited as part of the Toronto Neuroface Dataset, which has been described in detail previously [12]. See Table I for a brief summary of this dataset. The dataset was split into two parts: a finetuning set, and an evaluation set. The finetuning set was used here to finetune the pretrained SBR model (see below), and consisted of 9 individuals – 3 from each group – with the highest facial asymmetry from each group (as judged from the combined ratings of 2 speech-language pathologists [SLPs]). The remaining 27 individuals were used for evaluation of all models in the present study, so that data would be consistent across models. The study was approved by the Research Ethics Boards at the Sunnybrook Research Institute and University Health Network: Toronto Rehabilitation Institute.

TABLE I. BRIEF DEMOGRAPHIC SUMMARY

	Set	***ALS***	***Stroke***	***Control***
N	(FT, E)	3, 8	3, 11	3, 8
Age*	FT	[45, 55, 57]	[21, 62, 64]	[63, 65, 76]
	E	63 [58-75]	67 [60-89]	64.5 [33-78]
Sex (num. female)	FT	2/3	0/3	0/3
	E	5/8	4/11	4/8
Disease duration *. **	FT	34 [22-109]	125 [3-3262]	-
	E	[28, 41, 99]	[17, 49, NA]	-
ALSFRS-R*	FT	36.5 [26-40]	-	-
	E	[26, 35, 39]	-	-

FT = 'finetuning' dataset, E = 'Evaluation' dataset. *presented as median [min-max] for the E set, and as raw values for the FT set. **Disease duration is since resolution of symptoms in stroke (days), or since reported symptom onset in ALS (months). ALSFRS-R = ALS Functional Rating Scale – Revised. NA = not available.

B. Recording and tasks

Videos of participants performing oromotor tasks were recorded in a controlled laboratory setting using an Intel RealSense SR300 camera mounted on a tripod and placed 30-60 cm from the participants, who were seated. Frontal lighting was provided to illuminate participants' faces. Participants completed 3 oromotor tasks that we used to derive estimates of landmark tracking during steady-state (REST), vertical lower lip movement (OPEN), and horizontal mouth corner movement (SPREAD), respectively. During REST, participants had to be still and look straight ahead into the camera. OPEN involved participants maximally lowering their jaw, while looking straight ahead at the camera. Finally, during SPREAD participants had to spread the corners of their mouth as wide as possible, similar to a smile but without vertical elevation of the corners of the mouth. REST was not included in the Toronto Neuroface dataset; however, REST was collected at the same time as OPEN and SPREAD, under identical conditions with the same individuals.

C. Facial landmark tracking

Facial landmarks were tracked using 3 different algorithms: 1) a baseline method, 2) the facial alignment network (FAN) [10], and 3) the supervision by registration (SBR) algorithm [5]. The latter 2 models were further finetuned.

1) Baseline

The baseline facial landmark detector is a well-known regression tree-based model [13]. This system is implemented as a pretrained model in the Dlib machine learning library [14], and hence we refer to it as DLIB. The model was originally trained on the 300W dataset.

2) FAN

FAN is a neural network-based facial landmark detector that uses stacked hourglass networks to create heatmaps for each of 68 facial landmarks [10]. The pretrained FAN model can be run in either a 2-dimensional (2D) or 3-dimensional (3D) landmark prediction configuration; here, we used the 2D variant. Our lab has successfully used FAN for facial landmark detection in either its standard or finetuned form (see below) on data from different clinical populations [9], [15].

3) SBR

SBR is an approach to facial landmark detection that incorporates temporal consistency from video data into the landmark estimation process. It does this via a differentiable Lucas-Kanade (LK) module. The LK module refines a convolutional pose machine-based backbone [5] and penalizes not only landmark prediction error, but also predictions based on LK optical flow, thus providing a second training signal. Pretrained SBR model[1] – trained on a mixture of static (e.g., 300W and video (e.g., 300VW) data – is available online.

D. Finetuning neural networks

Finetuning was performed to improve the accuracy of landmark estimation in individuals that might be underrepresented in the pretrained models, i.e., older adults and those with neurological impairments. It is acknowledged that facial landmark detection error tends to be higher in older adults and those with neurological impairments [8], and so finetuning is a way to overcome this problem.

1) Finetuning FAN

FAN finetuning has been successful in previous studies in order to improve landmark detection accuracy in clinical/older healthy populations. The process is described in detail elsewhere [9], [12]. Briefly, FAN was finetuned by freezing all weights in the model except for those of the last hourglass. These were left unfrozen and were updated using a leave-one-subject-out cross validation on data from the Toronto Neuroface Dataset. Finetuned FAN is referred to from hereon as FAN-FT.

[1] SBR: https://github.com/D-X-Y/landmark-detection

2) Finetuning SBR

SBR was finetuned to explore the impact of finetuning on a video-optimized model. We used a total of 910 images from 9 individuals from the Neuroface dataset (3 from each of the 3 Neuroface cohorts; see Table I) to finetune the pretrained SBR model. Individuals from each group were chosen based on who had the highest average facial asymmetry rating as judged by 2 SLPs. These individuals used for finetuning were held out from further analyses, to see how well the finetuning process generalized to the 27 unseen individuals. Finetuned SBR is referred to from hereon as SBR-FT.

E. Measuring accuracy and smoothness

1) Accuracy

Accuracy of 20 predicted landmarks in the mouth region was evaluated to compare the models. We computed the normalized room mean square error (NRMSE), where the normalization factor is the diagonal length of the ground truth bounding box area.

2) Smoothness

Smooth landmark tracking is important for accurate estimation of position derivatives such as acceleration and jerk. We compared smoothness of the unfiltered position signals across all 5 models using 2 measures that have been recently proposed and/or evaluated with kinematic signals of differing lengths and complexities [16]. The first was the modified spectral arc length (SPARC) [16], which is estimated from the arc length of the Fourier spectrum of the movement. For brevity, we refer the reader to the full calculation of this metric in [16]. The second smoothness metric was the log dimensionless jerk (LDJ) [16], [17]. LDJ calculates the cumulative jerk, normalized by the peak velocity of a movement. For brevity, we refer the interested reader to [16], [17]for detailed calculation of this metric. We calculated smoothness measures using the first 5 seconds of each movement (other windows between 2 and 5 seconds were tested with comparable outcomes). For both smoothness measures, less-negative values indicate smoother tracking. For example, a value of -11 would indicate more smoothness than -12.

F. Statistical analysis

Statistical analyses were undertaken to compare the NRMSE, SPARC, and LDJ measures across the 5 landmark detection systems. Pairwise Wilcoxon paired-sample tests were conducted to evaluate differences between metrics for each pair of models, with correction for multiple comparisons via the Holm-Bonferroni method.

III. Results

A. Accuracy

Accuracy varied substantially across the 5 models, but the finetuned version of FAN and SBR typically had the lowest NRMSEs. SBR-FT had the lowest NRMSE for the ALS group, whereas FAN-FT had the lowest NRMSE for stroke and control groups. SBR-FT had significantly better ($p<1.66e-3$) accuracy than all models across all populations, except for the stroke group, in which FAN-FT had significantly better accuracy, and in the control group where FAN-FT had trending-towards significantly better accuracy ($p<0.05$). See Table II for a complete summary of NRMSE values.

TABLE II. AVERAGE NRMSE (%) FOR EACH POPULATION AND MODEL

		DLIB	FAN	FAN-FT	SBR	SBR-FT
ALS	NRMSE Mean ± SD	2.84 ± 1.19†	2.18 ± 0.39†	1.67 ± 0.30†	1.83 ± 0.45†	**1.59 ± 0.36**
Str.	NRMSE Mean ± SD	4.96 ± 5.59†	2.29 ± 0.37†	**1.50 ± 0.25†**	1.96 ± 0.31†	1.54 ± 0.24
HC	NRMSE Mean ± SD	2.37 ± 0.78†	2.21 ± 0.35†	**1.41 ± 0.26***	1.89 ± 0.36†	1.44 ± 0.35

FT = "fine-tuned". HC = "healthy control". Str. = "Stroke". Italicized values represent the lowest average NRMSE for each group. **Bolded** values reflect the lowest NRMSE for each group. †$p < 0.00166$ (significant after multiple comparisons correction). *$0.00166 < p < 0.05$ (not significant after multiple comparisons correction). Statistical significance is determined by comparison to SBR-FT only.

B. Smoothness

The smoothness of facial landmarks tracked using SBR-FT was compared to other models via SPARC and LDJ. Figure 1 depicts kinematic traces of 3 individuals, 1 from each cohort, demonstrating subjectively smoother landmark tracking by SBR-FT and SBR than other models. Objective measures of smoothness are summarized in Table III, and they agree with the subjective impressions. Overall, using SPARC, SBR-FT demonstrated the smoothest tracking in 4/9 population/task combinations, followed by FAN-FT (3/9). Overall, using LDJ, SBR-FT and SBR had the smoothest tracking in 8/9 and 1/9 cases, respectively. Finetuning appears to have improved tracking in SBR, but not FAN: SBR-FT had smoother tracking than SBR in 9/9 (SPARC) and 8/9 cases (LDJ). Interestingly, FAN-FT had smoother tracking than FAN in 6/9 using SPARC but only in 2/9 cases using LDJ. Smoothness comparisons trended towards significance in many cases, but only one of these survived the Bonferroni correction for multiple comparisons (significant p = 7.0e-4). In this sole exception, SBR-FT had significantly smoother tracking than FAN-FT by the LDJ metric for controls in the REST task.

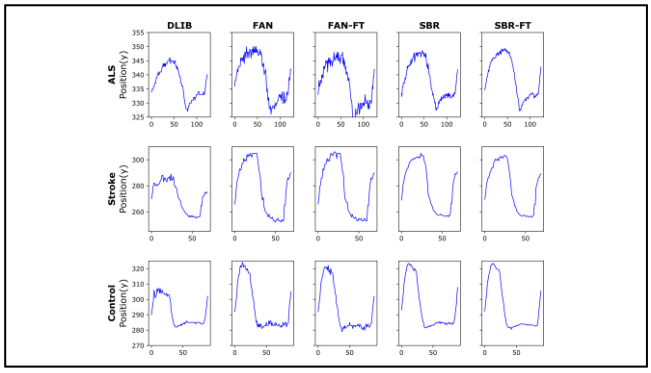

Figure 1. Examples of kinematics extracted from videos of a single individual from each cohort performing the OPEN task, processed by each of the 5 facial landmark detection systems.

TABLE III. SUMMARY OF SMOOTHNESS MEASURES

		Task	DLIB	FAN	FAN-FT	SBR	SBR-FT
SPARC	ALS	O	-19.34*	-13.40	**-13.25**	-13.56	-13.38
		R	-20.57*	-15.95	-15.72	-15.47	**-14.99**
		S	-16.90	-20.35	-19.22	-20.46	**-16.87**
	Stroke	O	-14.55*	-14.37*	-15.91*	-11.79*	**-11.17**
		R	-15.49	-16.78	**-15.42**	-16.65	-16.43
		S	**-15.42**	-17.62	-18.28	-17.00	-15.71
	HC	O	-14.62*	-11.43*	-10.36*	-10.68*	**-9.56**
		R	-24.69*	-19.92	**-18.88**	-23.26	-21.26
		S	-19.46	**-19.22**	-20.68	-20.70	-19.23
LDJ	ALS	O	-19.00*	-18.55*	-18.37*	-17.43	**-16.96**
		R	-19.71	-19.72	-20.08	-18.97	**-18.69**
		S	-19.20	-20.12	-20.26*	-19.29	**-18.55**
	Stroke	O	-19.43*	-18.37	-18.90*	-18.13*	**-17.93**
		R	-20.64*	-20.92*	-21.11*	-20.01*	**-19.71**
		S	-20.03*	-20.35	-20.42*	-19.88	**-19.34**
	HC	O	-18.91*	-18.14*	-18.33*	-17.48*	**-17.15**
		R	-17.01	-18.07*	-18.80†	-17.20	**-16.82**
		S	-19.85	-20.31	-20.05	**-19.83**	-19.83
	Total (SPARC)		1/9	1/9	3/9	0/9	**4/9**
	Total (LDJ)		0/9	0/9	0/9	1/9	**8/9**

O = OPEN, R = REST, S = SPREAD. FT = "fine-tuned". HC = "healthy control". *p < 7.0e-4, *7.0e-4 < p < 0.05 for comparison to SBR-FT only. Values in **bold** represent the best smoothness estimate for the given row. "Total" is the sum of cases where each model had the smoothest tracking.

IV. DISCUSSION AND CONCLUSION

In this study, we used finetuned SBR and FAN models to improve their performance detecting facial landmarks in older individuals with or without neurological impairments. We observed good accuracy and smoothness using the finetuned models, with FAN-FT was generally more accurate and SBR-FT tending to track more smoothly. This has interesting implications for landmark tracking in clinical populations.

We observed that finetuning generally improved accuracy, although we noted some trade-offs between accuracy and smoothness. With SBR, no such trade-off seemed to exist: smoothness and accuracy were both improved via finetuning. In contrast, FAN-FT was not uniformly improved in terms of smoothness of tracking, despite substantially and consistently improved accuracy. This is likely a consequence of the different architectures used by FAN and SBR – the latter has a built-in module to account for interframe consistency [5] whereas the former does not. This could, on the one hand, impose restrictions on how accurate SBR can become with finetuning, with the tradeoff of improving interframe consistency.

Smoother landmark tracking has potential clinical value. Filtering is common practice for kinematic timeseries, but can impact the estimation of kinematics [6], so starting with a smoother signal that requires less filtering is favorable. Smoother landmark tracking could also enable detailed assessment of between-trial articulatory variability, which can distinguish people with ALS from healthy individuals [18].

The present results illustrate the potential tradeoff between smoothness and accuracy from finetuning deep facial landmark tracking models. Smoother landmark estimation will improve the measurement of clinically-relevant kinematics, and thus may improve the detection and characterization of oromotor impairments.

V. ACKNOWLEDGEMENT

We sincerely thank Dr. Madhura Kulkarni her management support as well as Dr. Babak Taati for his feedback on an early draft of this manuscript.

REFERENCES

[1] P. Rong, Y. Yunusova, B. Richburg, and J. R. Green, "Automatic extraction of abnormal lip movement features from the alternating motion rate task in amyotrophic lateral sclerosis," Int. J. Speech. Lang. Pathol., vol. 20, no. 6, pp. 610–623, 2018.

[2] A. Bandini, J. R. Green, B. D. Richburg, and Y. Yunusova, "Automatic detection of orofacial impairment in stroke," in Interspeech, 2018, no. September, pp. 1711–1715.

[3] J. Schmidt, D. R. Berg, H. L. Ploeg, and L. Ploeg, "Precision, repeatability and accuracy of Optotrak optical motion tracking systems," Int. J. Exp. Comput. Biomech., vol. 1, no. 1, p. 114, 2009.

[4] J. J. Berry, "Accuracy of the NDI wave speech research system," J. Speech, Lang. Hear. Res., vol. 54, no. 5, pp. 1295–1301, 2011.

[5] X. Dong, S. I. Yu, X. Weng, S. E. Wei, Y. Yang, and Y. Sheikh, "Supervision-by-registration: An unsupervised approach to improve the precision of facial landmark detectors," arXiv, no. 1, pp. 360–368, 2018.

[6] J. Sinclair, P. John Taylor, and S. Jane Hobbs, "Digital filtering of three-dimensional lower extremity kinematics: An assessment," J. Hum. Kinet., vol. 39, no. 1, pp. 25–36, 2013, doi: 10.2478/hukin-2013-0065.

[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," J. Speech, Lang. Hear. Res., vol. 61, no. 5, pp. 1118–1129, 2018.

[8] B. Taati et al., "Algorithmic bias in clinical populations - Evaluating and improving facial analysis technology in older adults with dementia," IEEE Access, vol. 7, no. April, pp. 25527–25534, 2019.

[9] D. L. Guarin et al., "The Effect of Improving Facial Alignment Accuracy on the Video-based Detection of Neurological Diseases," J. Biomed. Heal. Informatics, pp. 1–9, 2020.

[10] P. Kopp, "Analysis and Improvement of Facial Landmark Detection," no. May, 2019, doi: 10.13140/RG.2.2.10980.42886.

[11] A. Bandini et al., "A New Dataset for Facial Motion Analysis in Individuals With Neurological Disorders," IEEE J. Biomed. Heal. Informatics, vol. 25, no. 4, pp. 1111–1119, 2020.

[12] A. Bulat and G. Tzimiropoulos, "How far are we from solving the 2D and 3D Face Alignment problem? (and a dataset of 230,000 3D facial landmarks)," arXiv, pp. 1021–1030, 2017.

[13] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., pp. 1867–1874, 2014.

[14] D. E. King, "Dlib-ml: A machine learning toolkit," J. Mach. Learn. Res., vol. 10, pp. 1755–1758, 2009.

[15] D. L. 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," arXiv, pp. 3–6, 2020.

[16] S. Balasubramanian, A. Melendez-Calderon, A. Roby-Brami, and E. Burdet, "On the analysis of movement smoothness," J. Neuroeng. Rehabil., vol. 12, no. 1, pp. 1–11, 2015.

[17] N. Hogan and D. Sternad, "Sensitivity of Smoothness Measures to Movement Duration, Amplitude and Arrests," J. Mot. Behav., vol. 41, no. 6, pp. 529–534, 2009.

[18] M. Kuruvilla-Dugdale and A. Mefferd, "Spatiotemporal movement variability in ALS: Speaking rate effects on tongue, lower lip, and jaw motor control," J. Commun. Disord., vol. 67, no. November 2016, pp. 22–34, 2017.