Int. J. Oral Maxillofac. Surg. 2021; 50: 1443–1449
https://doi.org/10.1016/j.ijom.2021.01.002, available online at https://www.sciencedirect.com

International Journal of
Oral &
Maxillofacial
Surgery

Research Paper
Artificial Intelligence

Detection of facial landmarks by a convolutional neural network in patients with oral and maxillofacial disease

M. Ding[1], Y. Kang[1], Z. Yuan[2], X. Shan[1], Z. Cai[1]

[1]Department of Oral and Maxillofacial Surgery, Peking University School and Hospital of Stomatology, Beijing, China; [2]Peking University School of Electronics Engineering and Computer Science, Beijing, China

M. Ding, Y. Kang, Z. Yuan, X. Shan, Z. Cai: Detection of facial landmarks by a convolutional neural network in patients with oral and maxillofacial disease. Int. J. Oral Maxillofac. Surg. 2021; 50: 1443–1449. © 2021 International Association of Oral and Maxillofacial Surgeons. Published by Elsevier Inc. All rights reserved.

Abstract. Facial nerve dysfunction is common in patients with Bell's palsy, trauma, tumour, or iatrogenic injuries. Imaging assessment is the most convenient method for patients and their treating physician. With developments in artificial intelligence (AI), manual work will be replaced. In this study, a database of facial images of patients with oral and maxillofacial diseases was set up to develop a facial nerve functional assessment system based on AI. This database was then used to evaluate the accuracy of a state-of-the-art algorithm for facial landmark detection named 'HRNet'. Utilizing this database and with appropriate human intervention, HRNet was used in facial annotation. The accuracy of annotations was evaluated through the normalized mean error. A total of 912 images were collected from 300 people; 546 of these images had abnormal features including defects, swelling, scars, or facial paralysis. The accuracy for the abnormal group was lower than that for the normal group before and after training, but improvements in accuracy were identified in both groups post-training. In conclusion, this new database demonstrates the ability of HRNet to localize facial landmarks in patients with oral and maxillofacial diseases. More images for training should be added to this database to diversify it in the future.

Key words: maxillofacial injuries; facial nerve diseases; anatomic landmarks; neural networks; database.

Accepted for publication 5 January 2021
Available online 5 March 2021

Facial nerve dysfunction is common in patients with Bell's palsy, trauma, tumour, or iatrogenic injuries. For such patients and their treating physician, the assessment of facial nerve function has roles from diagnosis to treatment.

Compared with other modalities (e.g., electromyography), imaging assessment is the most intuitive and convenient method. With developments in artificial intelligence (AI), manual work will be replaced to make entire imaging procedures faster, easier, more precise, and more user-friendly. Hence, we sought to develop a functional assessment system for facial nerves based on facial landmark detection by a convolutional neural network (CNN). In this way, patients need only a mobile device with a camera (e.g., mobile phone, laptop) to conduct self-assessment rapidly, anytime and anywhere.

Several databases have been used to train neural networks, most of which have been collected from healthy people.

0901-5027/01101443 + 07

Table 1. Movements provided by participants and instructions given to participants.

Term	Movement	Instruction
MV0	Neutral	Looking at the camera at rest
MV1	Eyebrows raised	Raising the eyebrows as if surprised
MV2	Frown	Wrinkling the eyebrows as if angry
MV3	Eyes closed	Closing eyes
MV4	Nose wrinkled	Wrinkling nose
MV5	Pout	Pouting mouth as if whistling
MV6	Maximal smile	Smiling showing teeth and clenched teeth
MV7	Mouth opened	Opening mouth

However, the deformities often found in patients with oral and maxillofacial diseases may reduce the accuracy of localization.

The aims of this study were (1) to measure the accuracy of AI detection in patients with oral and maxillofacial diseases; and (2) to establish a facial image database for patients with oral and maxillofacial diseases in order to obtain further improvements in accuracy.

Methods

Ethical approval of the study protocol

The study protocol was approved by the Ethics Committee of Peking University School and Hospital of Stomatology (PKUSSIRB-201949128) in Beijing, China. Patients provided written informed consent to participate in this study.

Subject recruitment

Hospital inpatients suffering from tumours, trauma, or palsy of the face who consented to participate in this study were assessed. They could have a normal appearance or show abnormal facial features, including swelling, defects, scars, or dynamic asymmetry of the face. The essential information of the participants was recorded, including their medical record number, sex, and age. People who had lost soft tissue in the face were not imaged due to the loss of landmarks.

Image collection

Frontal images were collected from two sources. The first was a camera (EOS 7D; Canon, Tokyo, Japan) with a ring flash, which captured images of 2304 × 3456 pixels in size. The second was a high-definition camcorder (FDR-AX100E; Sony, Tokyo, Japan), which was used to record videos from beginning to end, from which manual screenshots of key frames 1080 × 1920 pixels in size were obtained. The latter is similar to the patient taking a photograph using a front-facing camera.

The capture was performed under reasonable illumination indoors. The participants were required to sit in front of a solid-coloured background and look at the camera. Makeup, glasses, gastric tubes, and dressings were removed. The camera was adjusted to head height, and the whole head was in the centre of the camera screen.

The study participants were instructed to display different expressions (Table 1, Fig. 1) prompted by verbal and behavioural instructions from the researchers. Due to restrictions (time, illness), not all of the eight movements listed in Table 1 could be collected for each participant.

Datasets from individual patients were divided into 10 groups to execute 10-fold

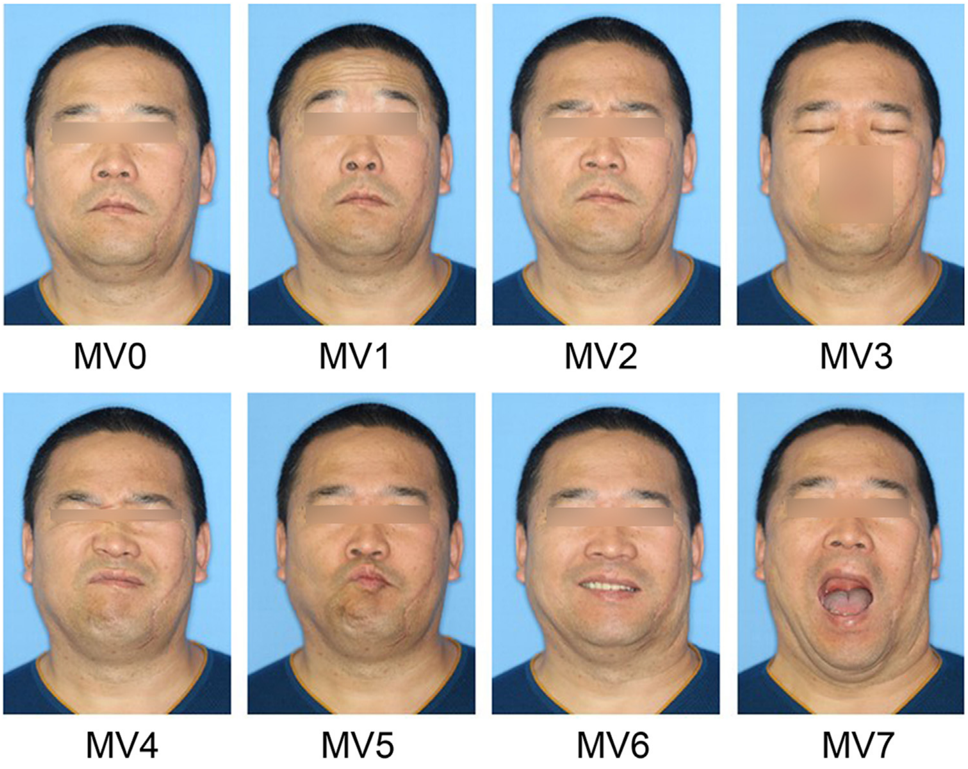

Fig. 1. Images of one study participant. All eight images were defined as 'abnormal'. The first row was labelled as 'scar' and the second row as 'facial paralysis and scar'.

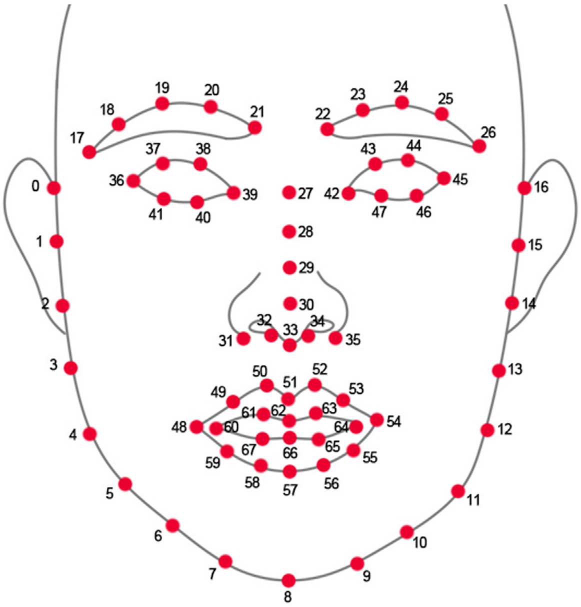

Fig. 2. Annotation scheme that accompanies HRNet.

Table 2. Details of the study participants.

	Number
Subjects	300
Sex	
Male	172
Female	128
Age (years)	
3–17	20
18–44	115
45–59	96
60–81	69

Table 3. Details of the subsets.

	Number
Images	912
Appearance	
Normal images	366 (40.1%)
Abnormal images	546 (59.9%)
Movements	
Neutral	294 (32.2%)
Eyebrows raised	91 (10.0%)
Frown	69 (7.6%)
Eyes closed	103 (11.3%)
Nose wrinkled	77 (8.4%)
Pout	102 (11.2%)
Maximal smile	98 (10.7%)
Mouth opened	78 (8.6%)

cross-validation. Each group in turn was used as a testing set and the others as the training sets. Each image was defined as ‘normal’ or ‘abnormal’ by an expert, on the basis of the demonstrated facial features. If the image showed a deformity such as swelling, defects, scars, or facial paralysis, it was labelled as ‘abnormal’.

Manual correction and CNN training

A ‘high-resolution network’ (HRNet) was selected as the CNN for the detection of the facial landmarks. This algorithm was developed jointly by the University of Science and Technology of China and Microsoft Research Asia (both of which are based in Beijing, China). This algorithm has shown excellent performance in facial landmark detection[1]. In tests of benchmark databases, including WFLW (Wider Facial Landmarks in-the-wild)[2], AFLW (Annotated Facial Landmarks in the Wild)[3], COFW (Caltech Occluded Faces in the Wild)[4], and 300 W (300 Faces In-the-Wild)[5], the ability of HRNet to detect key points surpassed that of all of its predecessors[6]. The annotation scheme used has 68 points (Fig. 2)[7]. This originates from the Carnegie Mellon University Multi-Pose Illumination Expression (CMU Multi-PIE) database, which does not define anatomical features for the position of each point. Hence, 68 points were defined according to anatomical structures to unify the correction criterion (**Supplementary Material** Table S1).

HRNet was pre-trained by the iBUG (Intelligent Behaviour Understanding Group) database, which was released as part of the first version of the 300 W challenge[5]. Then, all images were annotated by HRNet, the unsatisfactory output from which was corrected manually. The whole procedure was processed by a single researcher to avoid intra-group differences. The results of manual correction were approximated as ‘ground truth annotations’ (GTAs), which referred to the standard output for supervised learning. GTAs of training sets were used for training, and GTAs of testing sets were used to evaluate the accuracy of the output from HRNet.

HRNet was fine-tuned using the GTAs of training sets. The theory is based on an algorithm called ‘error back propagation’[8]. For each training sample, the algorithm first generates weights and parameters in the neural network randomly. Hence, the sample generates an output value after passing through the neural network and calculates the error between the output value and the ground-truth result. Then, the error is propagated back to the previous layers to adjust the weights and parameters. This step continues until the cumulative error of the whole training set reaches the minimum.

Accuracy evaluation

A coordinate system was set up in each image, and the coordinates of each point were obtained. The accuracy was measured by the point-to-point normalized mean error (NME) between the output results of the testing sets and their GTAs, and normalized by inter-ocular distance (the Euclidean distance between point 36 and point 45). Points on GTAs are denoted as (x_i, y_i). Points on output results are denoted as (x_i', y_i'). Greater accuracy is denoted by a small NME.

$$NME = \frac{\sum_{i=1}^{68}\sqrt{(x_i - x_i')^2 + (y_i - y_i')^2}}{68\sqrt{(x_{36} - x_{45})^2 + (y_{36} - y_{45})^2}}$$

Results

A total of 912 images were collected from 300 people. The proportions of male (57.3%) and female participants (42.7%) were similar (Table 2).

Overall, 366 of the 912 images were normal in appearance. The other 546 images showed defects, swelling, scars, or paralysis of the face. The number and proportion of each subset is shown in Table 3.

Before training, the NME of annotation on abnormal images was significantly higher than that on normal images ($P = 0.04$). This finding confirmed the conjecture proposed: the deformity features in patients with oral and maxillofa-

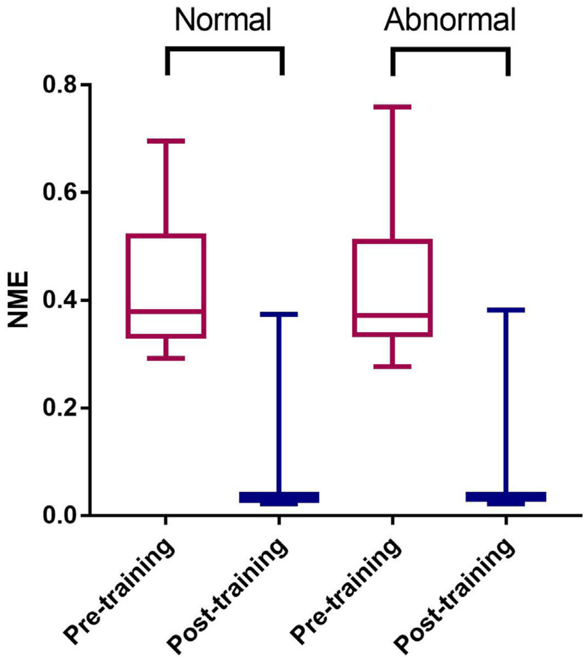

Fig. 3. Normalized mean error (NME) of the different subsets. The NME of normal images was higher than that of abnormal images. After training, the NME of both subsets decreased significantly.

cial diseases can reduce the accuracy of localization.

The NME of the testing set decreased markedly after training ($P < 0.001$), with the abnormal group continuing to show a significant difference compared with the normal group ($P < 0.001$), and both groups showed improvements ($P < 0.001$ and $P < 0.001$). These data are shown in Fig. 3. It was demonstrated that the oral and maxillofacial diseases database improved the accuracy of annotation on normal and abnormal images, but the sample size was not large enough for them to reach the same level.

Fig. 4 shows two cases of annotation: Fig. 4A and C are the pre-training images and Fig. 4B and D are the post-training images. In the first case, HRNet was unaffected by the scar on the nasolabial sulcus, but showed poor performance on the lower lip and nasal floor. In the second case, HRNet was confused by a defect in the pre-auricular area. After fine-tuning, HRNet performed better in both cases.

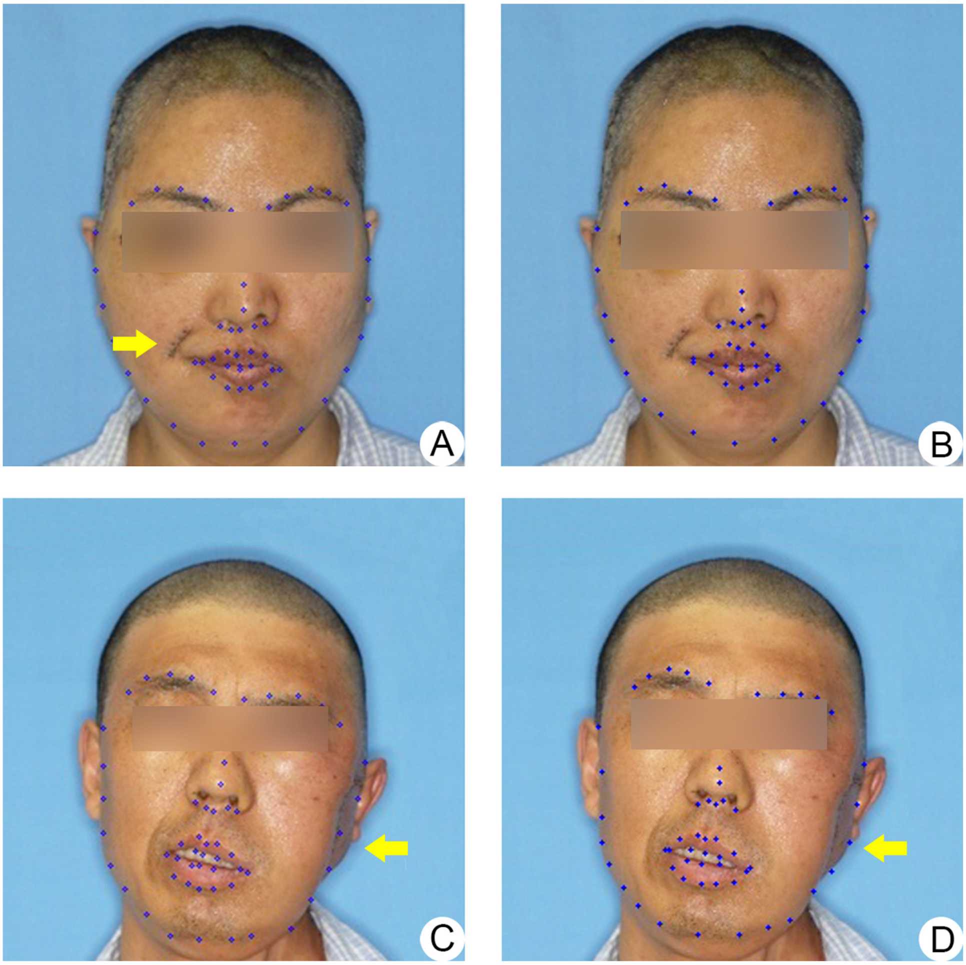

Fig. 4. Annotation by HRNet of two study participants. Images A and C are pre-training; images B and D are post-training.

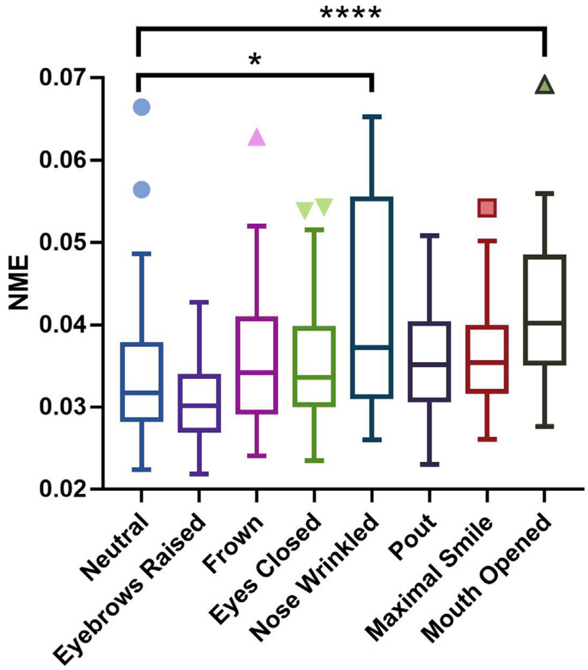

Fig. 5. Normalized mean error (NME) of the different movements after training. The NME of mouth opening was significantly higher than that of the other movements. Outliers >0.07 are not shown (*$P < 0.1$, ****$P < 0.0001$).

With regard to the different movements in the abnormal images, the NME of mouth opening was significantly higher than that for the other movements (Fig. 5). The landmarks around the mouth did not deviate significantly, but there was deviation on the facial margin. A possible reason is that soft tissue was squeezed when the mouth opened, which made the margin of the jaw unrecognizable (Fig. 6). Further research on regional NMEs is needed to confirm this hypothesis. In addition, although mouth-opening and eye-closing are orbicular movements, their NME showed differences. Landmarks apart from those of the eyes were similar to the neutral position when the eyes were closed, whereas almost all landmarks in the lower face were involved in mouth opening. This resulted in a higher sum of deviations of the individual points and a larger NME. With regard to nose-wrinkling, the relatively small sample size ($n = 57$) might have caused overfitting, leading to an inferior performance of HRNet.

With regard to the recognition of different abnormal features, HRNet had the strongest ability to annotate correctly with the impact of scars, followed by facial defects and swelling. The NME of facial paralysis was significantly higher than that of normal images (Fig. 7). This observation demonstrated that asymmetry of the facial organs had a greater influence than deformity of the facial margin.

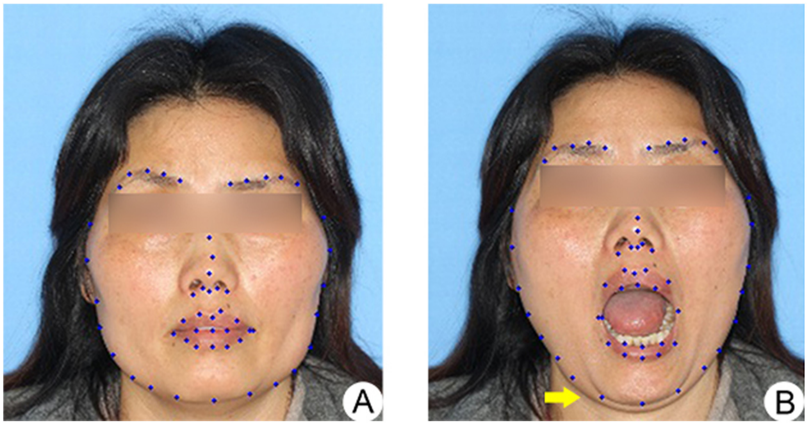

Fig. 6. Annotation by HRNet of a person with the mouth not opened (A) and opened (B). The soft tissue around the mandible is squeezed when the mouth is opened, which makes the margin of the jaw unrecognizable.

Discussion

Related works

The most widely used assessment method for facial nerve function is the House–Brackmann scale[9]. Clinicians evaluate disease conditions based on certain criteria empirically. Such a subjective method of assessment may lead to differences between clinicians. Therefore, objective methods, such as labelling specific landmarks on the face to make measurements and calculations, have been developed.

As far back as 1986, Burres[10] marked key points on patient faces with a pencil and measured the point-to-point distance manually using a ruler, at rest and during expression, to quantify facial asymmetry. This process was improved by using a reflective marker and image-editing software[11]. However, capturing and labelling images during busy sessions is difficult for clinicians. Objective methods are time-consuming compared with subjective methods, which hampers their application.

The application of AI has accelerated the process of locating facial landmarks. Several studies have used AI to detect facial landmarks, grading them by measuring certain indicators. Dong et al.[12,13] combined various algorithms to detect 14 landmark points on the face and calculate the distance between certain points. Song et al.[14] used AI trained by open-source databases to detect landmarks. Among these studies, the variety of AI for detecting landmarks was rich, and different training databases were used to identify facial landmarks, but the accuracy of annotation was not evaluated. Therefore, we believe that it is necessary to measure the accuracy of annotation in patients with oral and maxillofacial diseases.

Comparison with other databases

Existing databases can be divided into two categories: those in which images were captured under controlled conditions and those in which images were captured under unconstrained conditions[15]. 'Controlled conditions' refers to an indoor environment, with a unified background, stable illumination, and stable position. The CMU Multi-PIE database[7] is one of the largest databases with

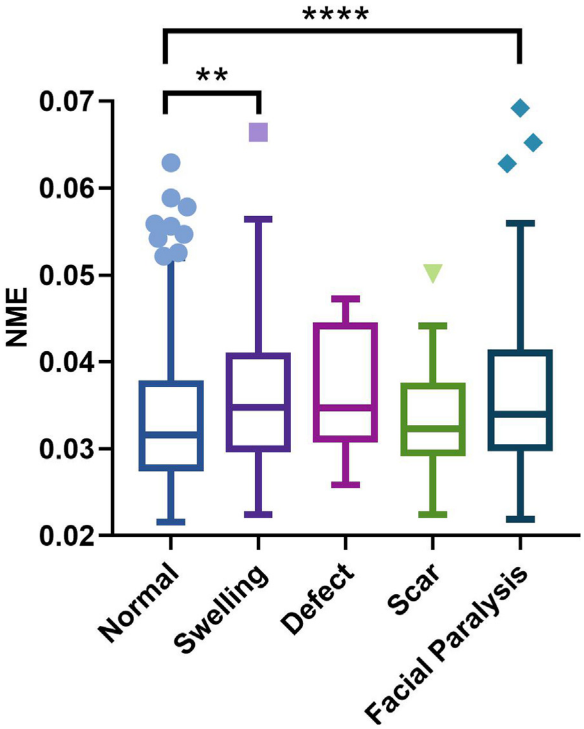

Fig. 7. Normalized mean error (NME) of the different features after training. The NME of facial paralysis was significantly higher than that of the normal condition. Outliers >0.07 are not shown (*$P <$ 0.1, ****$P < 0.0001$).

controlled conditions. Images captured under unconstrained conditions are sourced from a network that covers more types of situations, but the environment in which the images were taken is not known. Such databases, i.e., LFPW (Labelled Face Parts in the Wild)[16] and iBUG[5], contain even larger variations in expression, pose, and illumination than databases with controlled conditions.

Databases focused on patients with oral and maxillofacial diseases are distinct from the two categories mentioned above. The purpose of the oral and maxillofacial diseases database presented here is to improve the accuracy of annotation by CNN on patients with oral and maxillofacial diseases so that an assessment system of facial nerve function based on facial landmarks can be developed for patients and clinicians to make evaluations anytime and anywhere. Hence, the photographs used for annotation are taken in various environments. The background and pose are relatively easy to adjust, but illumination and the focal length of the camera are not. Therefore, the images used for training in the database should be as close as possible to the application environment so that the neural network can acquire the ability to discern interference factors.

Deep learning typically requires over 10,000 images for training. Hence, the number of samples in the oral and maxillofacial diseases database presented here is not sufficient. Illumination, background, and sessions need to be diversified further.

Competing interests

There are no conflicts of interest related to this study.

Funding

National Key Research and Development Program of China (grant number 2016YFC1102902) and Beijing Municipal Development Foundation of China (grant numbers Z161100000516016 and Z171100001017129).

Ethical approval

This study was approved by the Institutional Ethics Committee of Peking University School and Hospital of Stomatology (PKUSSIRB-201949128).

Patient consent

Written patient consent was obtained.

Statement to confirm

All authors have viewed and agreed to submission.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.ijom.2021.01.002.

References

1. Sun K., Xiao B., Liu D., Wang J. Deep high-resolution representation learning for human pose estimation. Computer Vision and Pattern Recognition. arXiv.org, February 25, 2019. https://arxiv.org/abs/1902.09212 [March 1, 2020].
2. Wu W., Qian C., Yang S., Wang Q., Cai Y., Zhou Q. Look at boundary: a boundary-aware face alignment algorithm. Computer Vision and Pattern Recognition. arXiv.org, May 26, 2018. https://arxiv.org/abs/1805.10483 [March 1, 2020].
3. Köstinger M, Wohlhart P, Roth PM, Bischof H. Annotated facial landmarks in the wild: a large-scale, real-world database for facial landmark localization. *Proceedings of the IEEE International Conference on Computer Vision* 2011:2144–51.
4. Burgos-Artizzu XP, Perona P, Dollár P. Robust face landmark estimation under occlusion. *Proceedings of the IEEE International Conference on Computer Vision* 2013:1513–20.
5. Sagonas C, Tzimiropoulos G, Zafeiriou S, Pantic M. 300 Faces in-the-wild challenge: the first facial landmark localization challenge. *Proceedings of the IEEE International Conference on Computer Vision* 2013:397–403.
6. Sun K., Zhao Y., Jiang B., Cheng T., Xiao B., Liu D., Mu Y., Wang X., Liu W., Wang J. High-resolution representations for labeling pixels and regions. arXiv.org, April 9, 2019. https://arxiv.org/abs/1904.04514 [March 1, 2020].
7. Gross R., Matthews I., Cohn J., Kanade T., Baker S. Multi-PIE. Proc Int Conf Autom Face Gesture Recognit 2010: 28: 807–813.
8. Rumelhart DE, Hinton GE, Williams RJ. Learning internal representations by error-propagation. *Readings in Cognitive Science* 1988;**323**:399–421.
9. Fattah AY, Gurusinghe AD, Gavilan J, Hadlock TA, Marcus JR, Marres H, et al. Facial nerve grading instruments: systematic review of the literature and suggestion for uniformity. *Plast Reconstr Surg* 2015;**135**:569–79.
10. Burres SA. Objective grading of facial paralysis. *Ann Otol Rhinol Laryngol* 1986;**95**(3 Pt 1):238–41.
11. Lou J, Yu H, Wang FY. A review on automated facial nerve function assessment from visual face capture. *IEEE Trans Neural Syst Rehabil Eng* 2020;**28**:488–97.

12. Dong J., Ma L., Li Q., Wang S., Liu L., Lin Y., Jian M. An approach for quantitative evaluation of the degree of facial paralysis based on salient point detection. 2008 International Symposium on Intelligent Information Technology Application Workshops, Shanghai, 2008: 483–486.
13. Dong J, Lin Y, Liu L, Ma L, Wang S. An approach to evaluation of degree of facial paralysis based on image processing and pattern recognition. *Journal of Information and Computational Science* 2008;**5**:639–46.
14. Song A, Xu G, Ding X, Song J, Xu G, Zhang W. Assessment for facial nerve paralysis based on facial asymmetry. *Australas Phys Eng Sci Med* 2017;**40**:851–60.
15. Sagonas C, Antonakos E, Tzimiropoulos G, Zafeiriou S, Pantic M. 300 Faces in-the-wild challenge: database and results. *Image Vis Comput* 2016;**47**:3–18.
16. Belhumeur PN, Jacobs DW, Kriegman DJ, Kumar N. Localizing parts of faces using a consensus of exemplars. *IEEE Trans Pattern Anal Mach Intell* 2013;**35**:2930–40.

Address:
Zhigang Cai
Department of Oral and Maxillofacial Surgery
Peking University School and Hospital of Stomatology
No. 22 South Avenue
Zhongguancun
Haidian District
Beijing 100081
China. Fax: +86 1062173402
E-mail: c2013xs@163.com