

The accuracy of automated facial landmarking - a comparative study between Cliniface software and patch-based Convolutional Neural Network algorithm

Bodore Al-baker¹, Xiangyang Ju², Peter Mossey³, Ashraf Ayoub^{4,*}

¹Orthodontic Department, Hamad Dental Centre, Hamad Medical Corporation, Onaiza Service Road, P.O. Box 3050, Doha, Qatar

²Medical Devices Unit, Department of Clinical Physics and Bioengineering, National Health Services of Greater Glasgow and Clyde, 378 Sauchiehall Street, Glasgow G2 3JZ, United Kingdom

³Dental Hospital and School, University of Dundee, Park Place, Dundee DD1 4HR, United Kingdom

⁴Scottish Craniofacial Research Group, Glasgow University Dental Hospital & School, School of Medicine, College of Medical, Veterinary and Life Sciences, University of Glasgow, 378 Sauchiehall Street, Glasgow G2 3JZ, United Kingdom

*Corresponding author. Glasgow University Dental School, 378 Sauchiehall Street, G2 3JZ, United Kingdom E-mail: ashraf.ayoub@glasgow.ac.uk

Abstract

Background: Automatic landmarking software packages simplify the analysis of the 3D facial images. Their main deficiency is the limited accuracy of detecting landmarks for routine clinical applications. Cliniface is readily available open-access software for automatic facial landmarking, its validity has not been fully investigated.

Objectives: Evaluate the accuracy of Cliniface software in comparison with the developed patch-based Convolutional Neural Network (CNN) algorithm in identifying facial landmarks.

Materials /Methods: The study was carried out on 30 3D photographic images; twenty anatomical facial landmarks were used for the analysis. The manual digitization of the landmarks was repeated twice by an expert operator, which considered the ground truth for the analysis. Each 3D facial image was imported into Cliniface software, and the landmarks were detected automatically. The same set of the facial landmarks were automatically detected using the developed patch-based CNN algorithm. The 3D image of the face was subdivided into multiple patches, the trained CNN algorithm detected the landmarks within each patch. Partial Procrustes Analysis was applied to assess the accuracy of automated landmarking. The method allowed the measurement of the Euclidean distances between the manually detected landmarks and the corresponding ones generated by each of the two automated methods. The significance level was set at 0.05 for the differences between the measured distances.

Results: The overall landmark localization error of Cliniface software was 3.66 ± 1.53 mm, Subalar exhibiting the largest discrepancy of more than 8 mm in comparison with the manual digitization. Stomion demonstrated the smallest error. The patch-based CNN algorithm was more accurate than Cliniface software in detecting the facial landmarks, it reached the same level of the manual precision in identifying the same points. The inaccuracy of Cliniface software in detecting the facial landmarks was significantly higher than the manual landmarking precision.

Limitations: The study was limited to one centre, one groups of 3D images, and one operator.

Conclusions: The patch-based CNN algorithm provided a satisfactory accuracy of automatic landmarks detection which is satisfactory for the clinical evaluation of the 3D facial images. Cliniface software is limited in its accuracy in detecting certain landmark which bounds its clinical application.

Keywords: facial landmarking; 3D; machine learning

Introduction

Over the past decade, the availability of non-invasive 3D facial imaging has facilitated the objective and reproducible analysis of the craniofacial morphology. In orthodontics and orthognathic surgery, 3D image analysis is essential for the diagnosis and management of craniofacial dysmorphology. The 3D facial images are essential for preoperative assessment, prediction planning, and the evaluation of post-operative changes [1]. Moreover, 3D technologies play a vital role as objective measurement tools in genetic and developmental studies. The integration of the 3D facial images and genomics

allowed the exploration of the genetic influences on morphological shape variation [2, 3].

However, the analysis of 3D facial images, ranging from simple linear measurements to the more comprehensive dense surface correspondence analysis, often requires the digitization of key landmarks which is challenging and prone to identification errors [1, 4, 5]. Hence, the development of robust and precise automated tools for 3D facial landmarking is crucial.

Cliniface, an open-source software, addresses this need. It provides the facility of automated landmarking of the face

for anthropometric and dysmorphological analysis. It was developed as an extension of the Meshmonk tool, which employed a non-rigid correspondence algorithm for landmark localization [6]. Cliniface offers a generic facial registration and landmarking process. This process involves the deformation of a symmetric anthropometric mask to an input target face which is followed by transferring the landmarks from the 'deformed' mask to the target 3D facial image.

Once landmarks are placed automatically, Cliniface extracts standardized facial measurements, including distances, angles, depth, and asymmetric differences which are used for the clinical assessment patient screening, treatment monitoring, and surgical planning.

Although Palmer et al., 2020, assessed Cliniface's accuracy using linear facial measurements, no previous studies have investigated its accuracy in detecting facial landmarks of 3D facial images.

In recent years, the Convolutional Neural Network (CNN) algorithms have emerged in computer vision as a promising tool for facial landmark detection that allows the analysis of complex patterns of the 3D facial images [7]. CNN is a powerful mathematical approach for deep learning. It is based on convolving local receptive fields over the image and apply the mathematical element-wise multiplication with learnable filters or kernels that allow CNN to extract valuable morphological features [8]. The interconnected mathematical layers of the CNN recognize the patterns across different image regions. Therefore, CNN is the ideal approach for facial landmark detection that requires a high level of accuracy. Over the last few years our team developed and validated automated facial landmarking based on the recent advances of the CNN algorithm [9]. Despite the extensive application of the CNN in computer vision, their application in clinical settings, particularly in orthodontics and orthognathic surgery, has been limited [10].

Recently, our team innovated the automatic landmark detection of 3D facial images in clinical settings using patch-based CNN algorithm. A patch is a surface around the centre of a landmark which is identified manually 'ground truth'. The patch is mathematically shifted along the x, y directions which create new augmented patches.

The augmented patches increase the sample size for the training of the Convolutional Neural Networks model and the machine learning process for the accurate automated landmarks detection. Data augmentation was carried out using translation cropping on 408 patches, resulting in a dataset of 10 200 PNG images (151×151 pixels) for each landmark. Initially, we built a high-quality, in-house ground truth dataset based on 408 3D facial images. This generated 408 patches for each of the 20 facial landmarks that were analysed in the study. These 408 patches were translated randomly to generate 10 200 images for each landmark, these were used for the deep learning algorithm. Full details can be seen in the previous article.

We demonstrated that the overall mean accuracy of facial landmarks detection using the patch-based CNN was 0.47 ± 0.52 mm. The lowest mean error of 0.41 ± 0.32 mm was along the y-axis, while the x-axis had a higher mean error of 0.45 ± 0.36 mm, and the z-axis had the highest mean error and standard deviation when compared to the other two axes (0.56 ± 0.89 mm). We concluded that the developed novel approach is accurate enough for the 3D analysis of facial morphology for clinical purposes [9].

This study aims to compare the accuracy of the well-established Cliniface software with the newly developed patch-based CNN in identifying anatomical facial landmarks of 3D facial images of a group of patients who require orthognathic surgical correction of their dento-facial deformities.

Material and methods

Ethical approvals for this study were obtained from the East of Scotland Research Ethics Committee (REC reference: 21/ES/0042) and NHS Greater Glasgow & Clyde Health Board NHS GG&C R&I reference: GN21OD153). All procedures including filing and storage of data were adhered to according to the guidelines and policies set forth by health authorities.

The study was carried out on thirty 3D stereophotographic facial images of adult orthognathic patients. The images were captured for the diagnosis and management of dentofacial deformities by the multidisciplinary orthognathic team. The patients gave their permission for their data to be used for research. All 3D facial images were captured under a controlled and strict 3D data collection protocol, using a passive stereophotogrammetry of the Di3D imaging system (Dimensional Imaging, Hillington, Glasgow, UK). The imaging system consisted of two-pod system, a stereo pair of cameras of each pod allowed the capture of each side of the face to build a photorealistic 3D image of the full face from ear to ear and from the hair line of the forehead to the hyoid bone. The accuracy of the system had been previously has been reported at 0.21 mm [11].

The sample size of thirty 3D facial images was based on previous studies which assessed the accuracy and reliability of automatic landmarking method [9, 12]. The 3D images were of the highest quality for analysis. Images which included missing regions, artefact or distortion were excluded from the study.

Twenty facial significant anatomical landmarks (Table 1) of the nose, eyes and the lips were used for the comparative evaluation of the detection accuracy between Cliniface and the patch-based CNN approach. Fig. 1 shows Cliniface's facial registration and landmark annotation. Fig. 2 shows the full set of landmarks detected by Cliniface and were used in the analysis of this study. The 20 landmarks in Fig. 1 and Table 1 can be detected by both the Cliniface software and the CNN algorithm. Therefore, therefore they were selected in this study for the comparative analysis.

Manual landmarking method was used as a ground truth of which the landmarks position was compared with both the Cliniface software and the patch-based CNN. Twenty landmarks were manually identified on each of the 3D facial images using the Di3D View software (Fig. 2b). The software allowed simultaneous viewing of the single image in three different windows, allowing rotation and magnification of the image. Accurate landmarks identification requires the operator to have full 3D control of the perspective and magnification of the images in order to correctly identify the landmarks on the 3D facial models. The coordinates of the soft tissue landmarks on the x, y, and z axis were extracted and recorded. The landmarks were identified by well-trained operator who went through a training process before landmarking the study sample. To assess the errors of the manual landmarking method, the whole set of landmarks were digitized twice, 2-week apart, by the same operator. The difference in landmarking was statistically analysed using

Table 1. Landmarks used in Cliniface software study.

	Landmark	Definition
1	Exocanthion (R)	Apex of the angle formed at the outer corner of the palpebral fissure where the upper and lower eyelids meet.
3	Exocanthion (L)	
2	Endocanthion (R)	Apex of the angle formed at the inner corner of the palpebral fissure where the upper and lower eyelids meet.
4	Endocanthion (L)	
5	Nasion	The midpoint on the soft tissue contour of the base of the nasal root where the frontal and nasal bones contact (nasofrontal suture).
6	Glabella	The most prominent midline point between the eyebrows, identical to bony glabella on the frontal bone.
7	Pronasale	Midline point marking the maximum protrusion of the nasal tip.
8	Subalare (R)	Point on the margin of the base of the nose where the ala disappears into the upper lip skin.
9	Subalare (L)	
10	Subnasale	Midpoint of angle at the columella base where the lower border of the nasal septum and the surface of the upper lip meet (the apex of the nasolabial angle).
11	Cheilion (R)	Point located at the corner of each labial commissure.
12	Cheilion (L)	
13	Crista philtre (R)	The peak of Cupid's bow.
14	Crista philtre (L)	
15	Labiale superius	The midpoint of the vermilion line of the upper lip.
16	Labiale inferius	The midpoint on the vermilion line of the lower lip.
17	Stomion	Midpoint of the labial fissure.
18	Sublabiale	Midpoint along the inferior margin of the cutaneous lower lip (labiomental sulcus).
19	Pogonion	The most anterior midpoint of the chin.
20	Gnathion	Midline point on the inferior border of the mandible.

R = Right, L = Left.

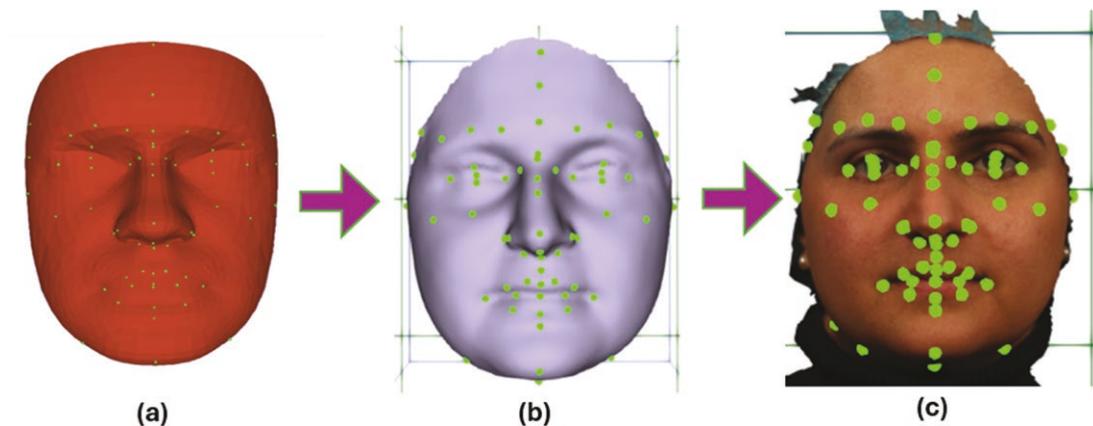


Figure 1. Cliniface's facial registration and landmark annotation showing anthropometric generic face mask (a) the non-rigid transformation 'deformation' of the template to the target face (b), landmarks transfer to the target face original surface (c).

paired Student t-test. Each 3D facial image, in OBJ format, was imported into Cliniface software and the landmarks were detected automatically (Fig. 2a).

The same set of the anatomical landmarks were automatically detected using the developed patch-based CNN. The 3D image of the face was subdivided into multiple patches, the trained network provided the automatic detection of the landmarks within each patch (Fig. 3). Partial Procrustes analysis allowed superimposition of landmarks configurations.

The assessment of that individual landmark accuracy by measuring the Euclidean distance between the corresponding individual landmarks of the two sets. We also analysed the landmarks detection accuracy of both methods, Cliniface and the patch-based CNN-based algorithm in relation to the intra-operator error of manual landmarking.

The accuracy of Cliniface software and the patch-based CNN for automatic landmark localization was measured in relation to the manually digitized landmarks (Ground truth).

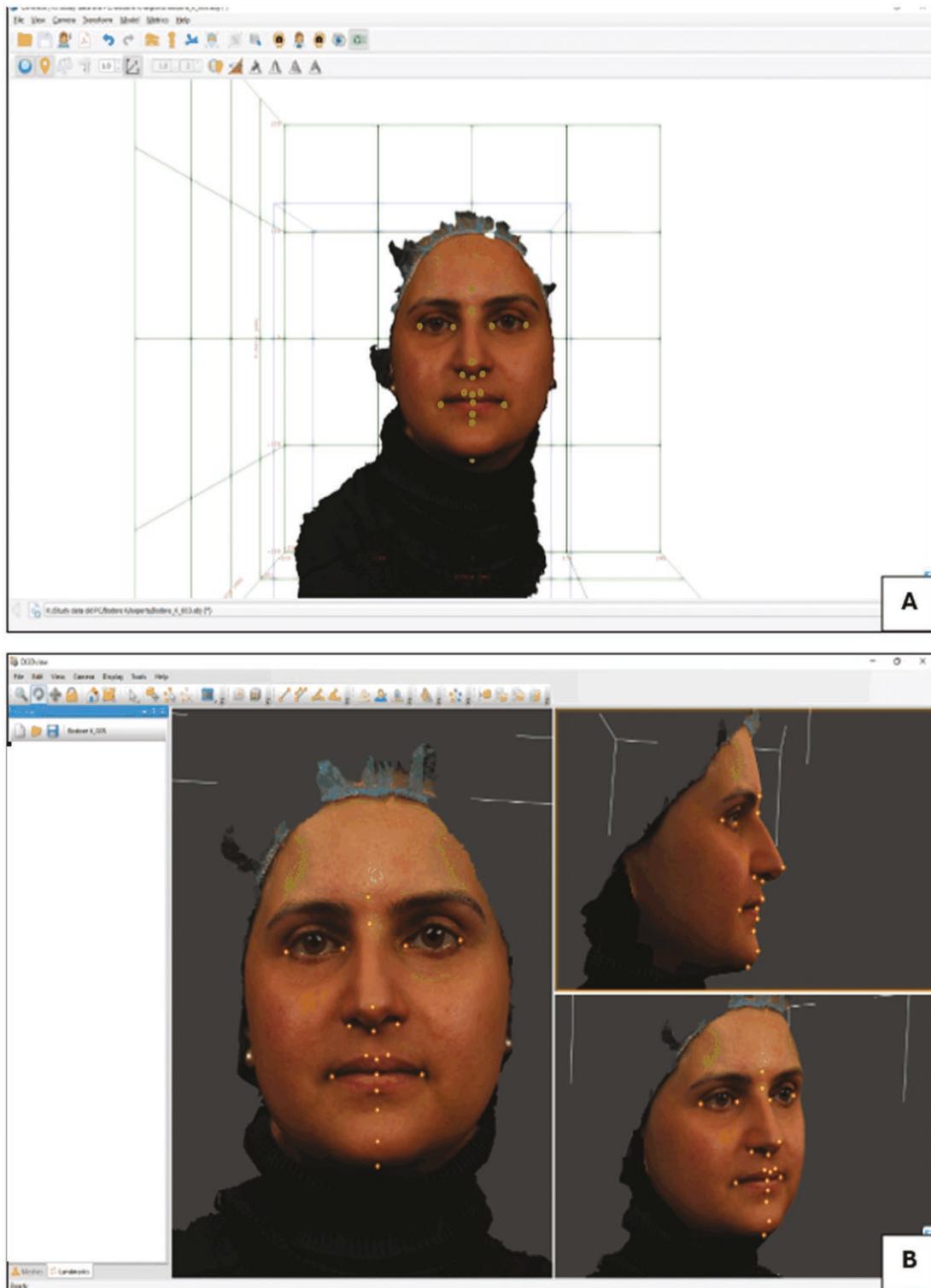


Figure 2. Automated detection of the landmarks' set in Table 1 using Cliniface software (Figure a) and Di3D software manual landmarking software (Figure b).

This was achieved by comparing the mean absolute distance in the x, y and z coordinates between manually digitized and automatically detected landmarks. The distances of each landmark on each image between manual and automated detection methods were calculated with a 3D Euclidian distance formula,

$$\text{Distance} = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2}$$

where x_1, y_1, z_1 are coordinates for manual detection and x_2, y_2, z_2 are coordinates for automated detection.

Statistical analysis

A one-sample t-test was used to assess the statistical significance of the mean difference in each landmark's position between the manual digitization and the automatic approaches

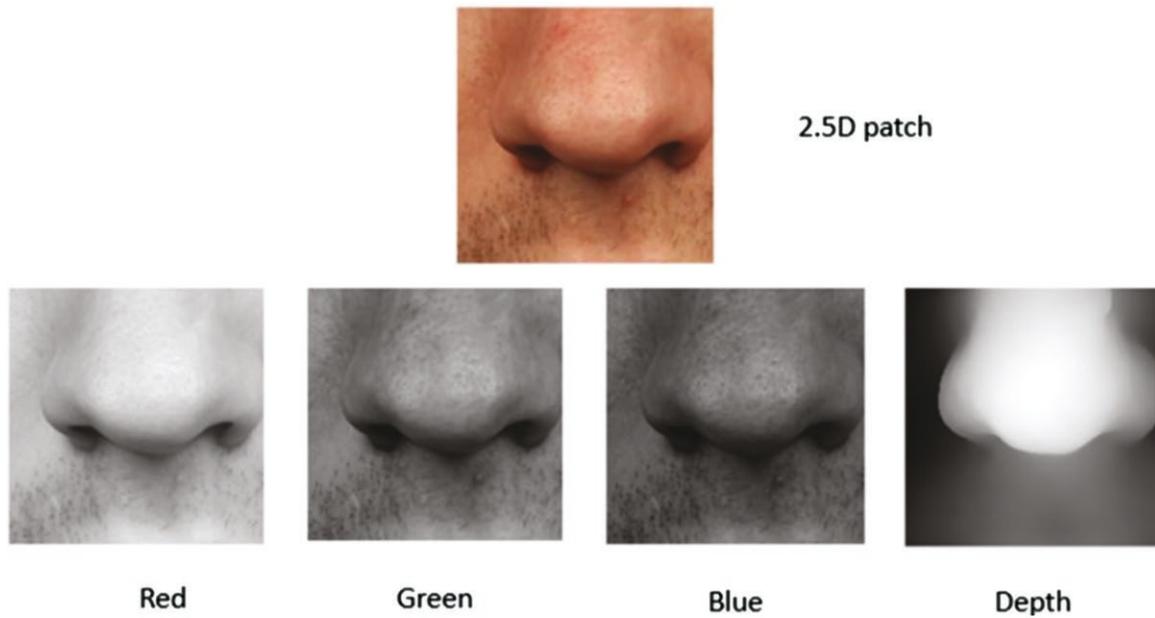


Figure 3. The mathematical algorithm of the 2.5D patch-based identification of the Pronasale, it shows the display of different colour shades that were used to maximize the accuracy of landmarks detection.

of the Cliniface software and the patch-based CNN. The significance level was set as 0.05 for the study outcomes.

Results

The overall mean of the intra-observer error calculated across subjects along all axes for all landmarks was 0.56 ± 0.69 mm. Values range between 0.20 mm and 2.23 mm. Most of the landmark's coordinates did not exhibit any statistically significant error based on Paired Student t-tests. The ICC was >0.90 , which had high rate of reproducibility in intra-examiner repetitive identification (Al Baker et al. [9])

Table 2 shows the accuracy of the automated Cliniface software in automatic detecting of the facial landmarks, comparing the errors between the automated landmarking (Cliniface) and manual landmarking.

The overall localization error, as determined by the Euclidean distance, was 3.66 ± 1.53 mm. Notably, the subalare landmarks exhibited the largest error, with an average discrepancy exceeding 8 mm. Conversely, the stomion landmark demonstrated the smallest error, with mean and standard deviation (SD) values of 1.49 ± 0.60 mm.

Among the three axes, the y-axis exhibited the highest mean error, while the x-axis showed the least error. Furthermore, statistical analysis revealed significant differences between the automated and manual methods of identifying most landmarks, except for the subnasale, crista philtra (R, L), labiale superius, and Labiale inferius. These differences were observed in at least one axis (with a P-value less than 0.0008 after Bonferroni correction).

The automated location of Nasion had the smallest mean error of 0.34 mm in the x-axis, whereas gnathion had the largest mean error of 6.56 mm in the y-axis. Other landmarks that demonstrated discrepancies of over 3 mm included the right and left subalare (x- and z- axes), right Cheilion (x-axis), and pogonion (y-axis). However, certain landmarks exhibited highly accurate identification, with discrepancies of less than 0.5 mm in their respective locations. These landmarks include

glabella (x-axis), pronasale (z-axis), subnasale (x-axis), labiale superius (x-axis), labiale inferius (x-axis), stomion (x-axis), and sublabiale (x-axis).

Table 3 displays landmark location errors between the Cliniface software and the manual method for all 20 landmarks across each axis. The results indicate that when using a threshold of 1 mm, 11 landmarks were within range for both the x- and z-axes, whereas only 4 landmarks were within 1 mm range for the y-axis.

Table 4 shows the error of the automatic detecting of the facial landmarks, using the patch-based CNN algorithm in relation to the manual landmarking (Gold standard). The overall localization errors measured by the Euclidean distances between the manual and automated landmarking was less than 1, in most of them it was within 0.5 mm except for Gn point which reached 1.16 mm.

Figure 4 shows the comparative accuracy between Cliniface software and the patch-based CNN for the automatic detection of the 20 facial landmarks. The results indicate that CNN model was more accurate and outperformed the Cliniface in the automatic detection of the facial landmarks.

Table 5 shows the mean differences 'Euclidean distances' between the manual intra-operator landmarks digitization errors and the automatic detection of the same set of the landmarks using the Cliniface software and the patch-based CNN algorithm. The patch-based CNN-based method was as accurate as the repeated manual digitization of the trained operator which was considered the ground truth in this study. In contrast, the Cliniface software showed a mean automatic detection error of the landmarks that is significantly higher than the intra-operator error.

The main reason for the inaccuracies encountered may be attributed to discrepancies in the registration stage initiated by Cliniface software before a common coordinate system was applied to measure the errors of automatic identification of the landmarks. This step is completely eliminated with the patch-based detection of landmarks using CNN approach.

Table 2. Mean errors of automatic landmarks identification of Cliniface software.

LM	Mean Absolute error ± SD			P-values of t-test*			Euclidian distance	
	(mm)			X	Y	Z	Mean error ± SD	95% Confidence interval of Mean
	X	Y	Z					
EX-R	1.55 ± 1.28	1.59 ± 0.93	2.39 ± 1.55	.063	.0024	<.001	3.51 ± 1.78	2.88; 4.15
EX-L	1.98 ± 1.56	1.49 ± 1.17	2.51 ± 1.67	.1206	<.001	<.001	3.92 ± 1.90	3.24; 4.6
EN-R	1.37 ± 0.88	0.85 ± 0.59	0.74 ± 0.61	<.001	.0477	.4532	1.97 ± 0.85	1.66; 2.27
EN-L	1.13 ± 0.92	0.90 ± 0.70	0.75 ± 0.48	<.001	.0052	.099	1.85 ± 0.86	1.55; 2.16
N	0.34 ± 0.29	1.86 ± 1.70	0.95 ± 0.45	.7131	<.001	<.001	2.31 ± 1.51	1.77; 2.85
GI	0.49 ± 0.36	5.26 ± 2.14	1.18 ± 0.74	.289	<.001	<.001	5.5 ± 2.09	4.75; 6.24
PRN	0.63 ± 0.41	2.30 ± 1.51	0.37 ± 0.31	.1201	<.001	.1356	2.55 ± 1.35	2.06; 3.03
SA (R)	6.32 ± 1.85	1.04 ± 0.77	5.65 ± 1.62	<.001	.389	<.001	8.75 ± 1.70	8.14; 9.36
SA (L)	6.21 ± 2.08	0.99 ± 0.83	5.46 ± 1.53	<.001	.5221	<.001	8.56 ± 1.82	7.91; 9.22
SN	0.49 ± 0.30	1.31 ± 0.86	0.94 ± 0.83	.0106	.0344	.3891	1.85 ± 0.96	1.5; 2.19
CH-R	3.44 ± 1.80	1.23 ± 0.80	0.83 ± 0.85	<.001	<.001	.4536	4.00 ± 1.61	3.42; 4.58
CH-L	2.91 ± 1.74	1.50 ± 1.01	0.72 ± 0.63	<.001	<.001	.4933	3.56 ± 1.73	2.94; 4.17
CP-R	1.07 ± 0.89	1.08 ± 0.75	0.79 ± 0.46	.8417	.4328	.0044	1.92 ± 0.90	1.6; 2.24
CP-L	0.94 ± 0.86	1.08 ± 0.76	0.69 ± 0.49	.5457	.269	.042	1.83 ± 0.90	1.53; 2.13
Lab-Sup	0.36 ± 0.24	1.20 ± 1.01	0.77 ± 0.50	.0254	.0149	.0584	1.63 ± 0.90	1.31; 1.95
Lab-Inf	0.45 ± 0.36	1.51 ± 1.52	1.09 ± 0.94	.025	.6857	.0191	2.13 ± 1.56	1.57; 2.68
STO	0.37 ± 0.33	1.04 ± 0.72	0.71 ± 0.51	.9067	<.001	.5779	1.49 ± 0.60	1.27; 1.71
SL	0.36 ± 0.34	2.05 ± 1.38	1.67 ± 1.57	.0977	<.001	<.001	2.84 ± 1.88	2.16; 3.51
PO	0.57 ± 0.45	5.61 ± 2.39	1.13 ± 1.03	.0866	.001	<.001	5.90 ± 2.27	5.09; 6.71
GN	0.90 ± 0.59	2.56 ± 1.79	6.56 ± 3.10	<.001	<.001	<.001	7.20 ± 3.41	5.98; 8.42
	1.59 ± 0.85	1.82 ± 1.14	1.79 ± 0.97				3.66 ± 1.53	

EX-R: Exocanthion (right), EN-R: Endocanthion (right), EN-L: Endocanthion (left), EX-L: Exocanthion (left), N: Nasion, PRN: Pronasale, SA-R: Subalare (right), SA-L: Subalare (left), SN: Subnasale, CH-R: Chelion (right), CH-L: Chelion (left), CP-R: Crista philtre (right), CP-L: Crista philtre (left), Lab-Sup: Labiale superius, Lab-Inf: Labiale inferius, STO: Stomion, SL: Sublabiale, PO: Pogonion, GN: Gnathion, GL: Glabella.

*One sample t-test. The level of significance was set at $p < .0008$ after Bonferroni correction (red colour).

Orange highlight indicates highest error. Bold indicates least mean error.

Table 3. Landmark location discrepancy between automated Cliniface software and manual method in x-, y- and z- axes for 20 landmarks.

Degree of discrepancy (mm)	X			Y			z			Total number	Total %
	LM	No	%	LM	No	%	LM	No	%		
≤0.5	N,GI,SN,Lab-Sup, Lab-Inf, STO,SL	7	35%		0	0%	PRN	1	5%	8	13%
0.5 < x ≤ 1	PRN,CP-L, PO, GN	4	20%	EN-R, EN-L, SA-L, STO	4	20%	EN-R, EN-L, N, SN, CH-R, CH-L, CP-R, CP-L, Lab-Sup, STO	10	50%	18	30%
1 < x ≤ 1.5	EN-R,EN-L, CP-R	3	15%	EX-L, SA-R, SN, CH-R, CH-L, CP-R, CP-L, Lab-Sup	8	40%	GI, Lab-Inf, PO	3	15%	14	24%
1.5 < x < 2	EX-R, EX-L	2	10%	EX-R, N, Lab-Inf	3	15%	SL	1	5%	6	10%
2 ≤ x < 3	CH-L	1	5%	PRN, SL, GN	3	15%	EX-R, EX-L	2	10%	6	10%
≥3	SA-R,SA-L,CH-R	3	15%	GL, PO	2	10%	SA-R, SA-L, GN	3	15%	8	13%

No: number, EX-R: Exocanthion (right), EN-R: Endocanthion (right), EN-L: Endocanthion (left), EX-L: Exocanthion (left), N: Nasion, PRN: Pronasale, SA-R: Subalare (right), SA-L: Subalare (left), SN: Subnasale, CH-R: Cheilion (right), CH-L: Cheilion (left), CP-R: Crista philtre (right), CP-L: Crista philtre (left), Lab-Sup: Labiale superius, Lab-Inf: Labiale inferius, STO: Stomion, SL: Sublabiale, PO: Pogonion, GN: Gnathion, GL: Glabella.

Table 4. Mean errors of automatic landmarks identification of the patch-based CNN algorithm.

LM	Mean Absolute error \pm SD			p-values of t-test*			Euclidian distance
	(mm)			X	Y	Z	Mean error \pm SD
	X	Y	Z				
EX-R	0.29 \pm 0.23	0.35 \pm 0.28	0.40 \pm 0.70	.0500	.3825	.0888	0.73 \pm 0.67
EX-L	0.29 \pm 0.20	0.26 \pm 0.23	0.27 \pm 0.23	.4543	.8300	.1093	0.53 \pm 0.28
EN-R	0.30 \pm 0.29	0.32 \pm 0.22	0.22 \pm 0.22	.1208	.2756	.6802	0.56 \pm 0.34
EN-L	0.31 \pm 0.21	0.23 \pm 0.18	0.22 \pm 0.23	.0167	.8523	.1182	0.50 \pm 0.27
N	0.27 \pm 0.26	0.40 \pm 0.30	0.10 \pm 0.15	.3164	.2755	.6126	0.54 \pm 0.37
GI	0.29 \pm 0.22	0.77 \pm 0.54	0.07 \pm 0.12	.2654	.1411	.1091	0.86 \pm 0.55
PRN	0.26 \pm 0.18	0.30 \pm 0.21	0.04 \pm 0.06	.8855	.2533	.2090	0.44 \pm 0.21
SA (R)	0.41 \pm 0.22	0.24 \pm 0.21	0.43 \pm 0.38	.3737	.1027	.4362	0.72 \pm 0.37
SA (L)	0.29 \pm 0.29	0.23 \pm 0.17	0.23 \pm 0.23	.8487	.1877	.5173	0.48 \pm 0.34
SN	0.30 \pm 0.24	0.23 \pm 0.18	0.15 \pm 0.17	.9386	.7565	.6195	0.47 \pm 0.28
CH-R	0.36 \pm 0.27	0.20 \pm 0.19	0.14 \pm 0.15	.1862	.5945	.6355	0.49 \pm 0.30
CH-L	0.32 \pm 0.22	0.20 \pm 0.15	0.16 \pm 0.13	.0773	.2201	.5885	0.47 \pm 0.20
CP-R	0.38 \pm 0.32	0.26 \pm 0.20	0.09 \pm 0.13	.0924	.1485	.8149	0.51 \pm 0.33
CP-L	0.38 \pm 0.31	0.25 \pm 0.22	0.10 \pm 0.13	.1081	.2969	.3155	0.52 \pm 0.1
Lab-Sup	0.40 \pm 0.37	0.24 \pm 0.17	0.08 \pm 0.07	.3952	.3379	.8568	0.53 \pm 0.34
Lab-Inf	0.32 \pm 0.33	0.34 \pm 0.29	0.20 \pm 0.21	.8668	.0041	.1077	0.59 \pm 0.39
STO	0.37 \pm 0.39	0.31 \pm 0.22	0.19 \pm 0.18	.1783	.8125	.4127	0.59 \pm 0.33
SL	0.54 \pm 0.45	0.35 \pm 0.24	0.09 \pm 0.08	.4085	.8954	.6545	0.73 \pm 0.33
PO	0.67 \pm 0.45	0.50 \pm 0.51	0.12 \pm 0.18	.4195	.4185	.0760	0.96 \pm 0.54
GN	0.76 \pm 0.62	0.34 \pm 0.31	0.57 \pm 0.60	.6631	.6201	.2051	1.16 \pm 0.72
	0.37 \pm 0.13	0.31 \pm 0.13	0.19 \pm 0.14				0.62 \pm 0.19

Discussion

Review of the literature identified the lack of information regarding the threshold value of acceptable landmarking error for clinical and biological use. The threshold of manual landmarking varies in the literature. Some consider errors larger than 0.5 mm to be significant [13], while others consider 1mm as the clinical threshold of landmarking errors [14]. Others reported a threshold of landmarking error of 2 mm [15]. Only 5% of the reviewed studies agreed that a threshold of 1 mm landmarking errors is considered acceptable. Based on the reported reproducibility of 3D facial manual landmarking the 0.5 mm remains the gold standard for clinical applications [16]. Facial landmarking is time consuming and requires a comprehensive training for the operator to achieve this level of accuracy. Therefore, the automated facial landmarking has always been the holy grail of facial analysis. The lack of a reliable automated landmarking with a satisfactory accuracy has inspired this study.

The emergence of deep learning, particularly the CNN, has led to significant advancements in landmark detection [17]. This is well documented in computer vision applications [18]. CNN have shown promising results in identifying specific visual patterns that correspond to landmarks' locations [19]. This accuracy is dependent on the training of the CNNs using a large dataset of annotated images where the landmarks of interest are labelled. These advancements have been supported by the availability of publicly accessible large-scale annotated databases [20]. However, caution is necessary when applying these databases to develop landmark models for clinical purposes. Special attention must be considered to the quality of

the ground truth landmarking in such cases [19]. There are two common CNN approaches for facial landmark detection; the heatmap regression and the dense regression method. Heatmap regression generates a heatmap of the face, where each point corresponds to the likelihood of a facial landmark being located at that point [21]. In contrast, dense regression directly outputs the coordinates of each facial landmark [22]. For our dataset, which comprises standardized 3D facial images, we opted for dense regression over heatmap method due to its speed and efficiency. The development of the innovative CNN patch based approach and its robust validation of our team [9] provided the basis of the comparative validation of the readily available and commonly used software (Cliniface) in this study.

Cliniface is an open-source software that can automatically detect craniofacial landmarks and provide linear and angular measurements of the face, making it a valuable tool for clinical and research professionals. Although Palmer et al [23] have assessed the Cliniface's accuracy using linear facial measurements, no studies have investigated its effectiveness in the detection facial landmark of 3D facial images. Therefore, we assessed the accuracy of Cliniface software in identifying facial landmarks in our population sample. The rationale of the study was the evaluation of this software in clinical applications.

In this study, we evaluated the accuracy of automatic identification for 20 facial landmarks using Cliniface software in comparison to the newly developed and validated patch-based CNN algorithm. Our findings revealed that while some landmarks were identified accurately by Cliniface software,

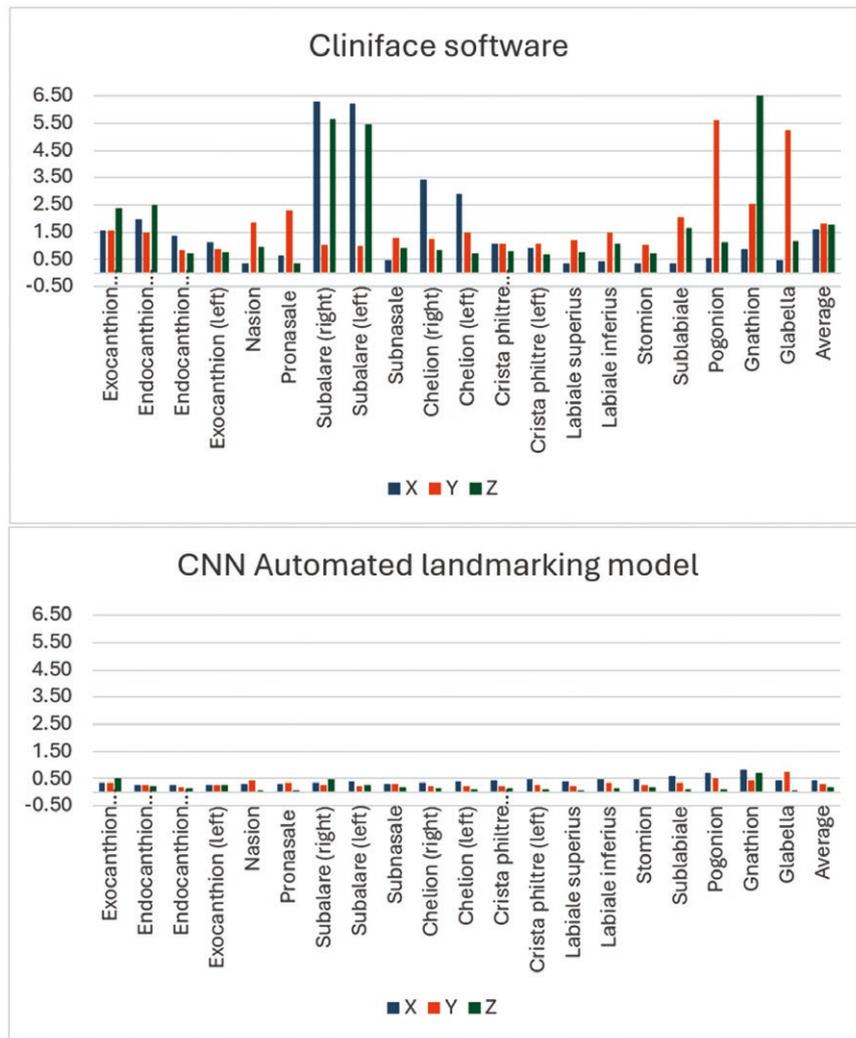


Figure 4. The difference, in x, y, z axis, between manual digitization and the automatic detection of the landmarks using Cliniface software (left) and CNN patch-based approach (right).

notable discrepancies were observed in others. The automated location of nasion showed the smallest mean error of 0.34 mm in the x-axis, whereas gnathion exhibited the largest mean error of 6.56 mm in the y-axis. Additionally, several other landmarks, such as the right and left subalare (x and z-axes), right cheilion (x-axis), and pogonion (y-axis), displayed discrepancies exceeding 3 mm.

These discrepancies were particularly prominent in peripheral landmarks, consistent with findings from previous studies. Torres et al. [24] reported limitations in their automated model for detecting non-featured and flat regions. Wen et al. [25] also found higher identification accuracy in central landmarks compared to peripheral ones. This pattern of errors was similar to that found with manual landmarking in previous studies [13, 14, 26].

Despite being considered reliable in manual landmarking, we encountered difficulty in identifying exocanthion landmarks using the Cliniface software. This aligns with previous automated landmarking studies, which found exocanthion among the most challenging points to locate automatically [27–29]. The inaccuracies encountered may be attributed to discrepancies in the registration stage initiated by Cliniface software, possibly influenced by facial variations.

It is widely recognized that an open access software for automatic facial landmarking, and soft tissue analysis would be valuable for researchers and clinicians. Although Cliniface software offers a user-friendly interface allowing users to adjust the detected landmark positions, this task demands high expertise and familiarity with accurate landmark placement. Consequently, clinicians should not solely rely on the software. The findings of the study emphasize the potential value and the limitations of Cliniface software as a tool for clinical and research studies. It is important to exercise caution when using the software as it is not a reliable substitute to the manual landmarking.

Guarin et al. [30] conducted a study revealing a bias in the CNN automated approach used for facial landmark localization, specifically in patients with facial palsy. The model showed significantly poorer performance when applied to individuals with complete paralysis compared to those with near-normal cases. This finding aligns with previous research on elderly individuals with dementia [31] where they concluded that training an automated model for facial landmark localization using facial images of a clinical database comprising elderly subjects with dementia and patients with Bell's palsy resulted in improved accuracy when dealing with

Table 5. The mean difference between the manual landmarking errors and the automatic detection of the same landmarks using the Cliniface software and the patch-based CNN method.

Landmark	Cliniface software		Patch-based CNN		Intra-operator error of manual method	
	Mean error	SD	Mean error	SD	Mean error	SD
EX-R	3.51	1.78	0.68	0.39	0.68	0.37
EX-L	3.92	1.9	0.53	0.32	0.64	0.37
EN-R	1.97	0.85	0.53	0.31	0.51	0.26
EN-L	1.85	0.86	0.45	0.3	0.43	0.24
N	2.31	1.51	0.75	0.45	0.81	0.48
Gl	5.5	2.09	1.22	0.73	1.05	0.66
PRN	2.55	1.35	0.64	0.37	0.97	0.57
SA (R)	8.75	1.7	0.59	0.34	0.68	0.49
SA (L)	8.56	1.82	0.56	0.32	0.78	0.48
SN	1.85	0.96	0.62	0.35	0.68	0.37
CH-R	4.00	1.61	0.54	0.34	0.64	0.34
CH-L	3.56	1.73	0.55	0.32	0.58	0.39
CP-R	1.92	0.9	0.64	0.36	0.60	0.36
CP-L	1.83	0.83	0.68	0.44	0.67	0.25
Lab-Sup	1.63	0.9	0.61	0.38	0.46	0.26
Lab-Inf	2.13	1.56	0.79	0.42	0.77	0.44
STO	1.49	0.61	0.66	0.43	0.58	0.28
SL	2.84	1.88	0.89	0.44	0.84	0.49
PO	5.90	2.27	1.18	0.68	0.95	0.44
GN	7.20	3.41	1.17	0.69	0.73	0.40
Overall	3.66	1.53	0.71	0.42	0.70	0.40

patients affected by the same condition. This improvement was observed by comparing the performance of the model trained on the disease-specific database with that of a model trained on a much larger database consisting of non-patient population. This indicates an algorithmic bias, suggesting that even large training datasets fail to capture the diversity present in the wide range of patient populations. Therefore, the use of publicly available automated model or datasets for developing automatic algorithm to detect facial landmarks in clinical settings is restricted in its applicability.

It is important to highlight the limitations of existing databases of 3D facial images which includes landmarks that are non-anatomical and ill-defined as seen in the Headspace repository which contains a set of 3D images of the human head that is available for university-based non-commercial research [32]. Likewise, the 3D Facial Norms Database, comprising 2454 3D images and ground truth marking of 24 landmarks, lacked the colour and texture surface features, focussing only on surface geometry [33]. While this simplifies facial analysis in some respects, it also limits the range of methods for automated landmark detection. Information about colour and texture can play a crucial role in machine learning and deep learning which we utilized in our study.

It is important to emphasize the importance of the quality and the accuracy of the manual digitalization of the landmarks to be used as the ground truth for comparative studies. We followed a strict protocol of manual landmarking to overcome the limitations highlighted in our systematic review [10] by improving the quality of the reference standards,

population selection, and study design for the reliable evaluation of the accuracy of the automated landmarking.

Limitations of the study include the involvement of only one centre and one annotator. Future research should incorporate datasets from multiple centres and diverse ethnic backgrounds. It is essential to be mindful of this limitation while utilizing the Cliniface software for clinical facial analysis. Caution should be exercised when dealing with landmarks located around the chin, side of the face, and gonial angle, as they are more prone to inaccuracy. Hence, it is advisable to conduct a visual inspection before proceeding with generating further soft tissue anthropometric measurements.

Conclusion

The patch-based CNN provided a satisfactory accuracy of automatic landmarks detection for the clinical evaluation of the 3D facial images. While Cliniface is readily available as an open-access tool for automatic facial landmarking, our study reveals notable discrepancies in certain landmark identifications which limits its reliability in facial landmark detection.

Author contributions

Bodore Albaker (Conceptualization [supporting], Data curation [lead], Formal analysis [lead], Writing—original draft [supporting]), Xiangyang Ju (Conceptualization [supporting], Methodology [supporting], Software [lead], Writing—review & editing [supporting]), Peter Mossey (Conceptualization

[supporting], Investigation [supporting], Supervision [lead], Writing—review & editing [supporting], and Ashraf Ayoub (Conceptualization [equal], Writing—review & editing [lead])

Conflict of interest

The authors of the manuscript have NO Conflict of interest regarding position, activities, or relationships of an individual, whether direct or indirect, financial or non-financial, could influence or be seen to influence the opinions or activities of the individual.

Funding

None declared.

Data availability

The data underlying this article cannot be shared publicly due to the privacy of the patients that participated in the study. The data will be shared on reasonable request to the corresponding author.

References

- Al Mukkhtar A, Khamabay B, Ju X, et al. Comprehensive analysis of soft tissue changes in response to orthognathic surgery: mandibular versus bimaxillary advancement. *Int J Oral Maxillofac Surg* 2018;47:732–7. <https://doi.org/10.1016/j.ijom.2017.11.014>
- Lee MK, Shaffer JR, Leslie EJ, et al. Genome-wide association study of facial morphology reveals novel associations with *FREM1* and *PARK2*. *PLoS One* 2017;12:e0176566. <https://doi.org/10.1371/journal.pone.0176566>
- Hsokens H, Liu D, Naqvi S, et al. 3D facial phenotyping by biometric sibling matching used in contemporary genomic methodologies. *PLoS One Genet* 2021;7:e1009528. <https://doi.org/10.1371/journal.pgen.1009528>
- Patel A, Islam SM, Murray K, et al. Facial asymmetry assessment in adults using three-dimensional surface imaging. *Prog Orthod* 2015;16:36–42. <https://doi.org/10.1186/s40510-015-0106-9>
- Ozdemir SA, Esenlik E. Three-dimensional soft-tissue evaluation in patients with cleft lip and palate. *Int Med J Exp Clin Res* 2018;24:8608. <https://doi.org/10.12659/MSM.912305>
- White JD, Ortega-Castrillon A, Mathews H, et al. Mesh Monk: open-source large-scale intensive 3D phenotyping. *Sci Rep* 2019;5:91–9. <https://doi.org/10.1038/s41598-019-42533-y>
- Olivetti EC, Ferretti J, Cirrincione G, et al. Deep CNN for 3D face recognition. In: Rizzi C, Andrisano AO, Leali F, Gherardini F, Pini F, Vergnano A (eds), *Design Tools and Methods in Industrial Engineering*. ADM 2019. *Lecture Notes in Mechanical Engineering*. Cham: Springer, 2020. https://doi.org/10.1007/978-3-030-31154-4_56
- Sahu M, Dash R. A survey on deep learning: Convolution Neural Network (CNN). In: Mishra D, Buyya R, Mohapatra P, Patnaik S (eds), *Intelligent and Cloud Computing*. *Smart Innovation, Systems and Technologies*. Vol. 153. Singapore: Springer, 2021. https://doi.org/10.1007/978-981-15-6202-0_32
- Al Baker B, Ayoub A, Ju X, et al. Patch-based convolutional neural networks for automatic landmark detection of 3D facial images in clinical settings. *Eur J Orthod* 2024;46:12. <https://doi.org/10.1093/ejo/cjae056>
- Al-Baker B, Alkalaly A, Ayoub A, et al. Accuracy and reliability of automated three-dimensional facial landmarking in medical and biological studies. A systematic review. *Eur J Orthod* 2023;45:382–95. <https://doi.org/10.1093/ejo/cjac077>
- Khambay B, Nairn N, Bell A, et al. Validation and reproducibility of a high-resolution three-dimensional facial imaging system. *Br J Oral Maxillofac Surg* 2008;46:27–32. <https://doi.org/10.1016/j.bjoms.2007.04.017>
- Baksi S, Frezer S, Matsumoto T, et al. Accuracy of an automated method of 3D soft tissue landmark detection. *Eur J Orthod* 2021;43:622–30. <https://doi.org/10.1093/ejo/cjaa069>
- Hajeer MY, Ayoub AF, Millett DT, et al. Three-dimensional imaging in orthognathic surgery: the clinical application of a new method. *Int J Adult Orthod Orthognathic Surg* 2002;17:318–30.
- Gwilliam JR, Cunningham SJ, Hutton T. Reproducibility of soft tissue landmarks on three-dimensional facial scans. *Eur J Orthod* 2006;28:408–15. <https://doi.org/10.1093/ejo/cjl024>
- Aynechi N, Larson B, Leon V, et al. Accuracy and precision of a 3D anthropometric facial analysis with and without landmark labeling before image analysis. *Angle Orthod* 2011;82:245–54. <https://doi.org/10.2319/041810-210.1>
- Toma AM, Zhurov A, Playle R, et al. Reproducibility of facial tissue landmarks on 3D laser-scanned facial images. *J Orthod Craniofac Res* 2019;12:32–42. <https://doi.org/10.1111/j.1601-6343.2008.01435.x>
- Terada T, Chen Y, Kimura R. 3D facial landmark detection using deep convolutional neural networks. In: The 14th International Conference on Natural Computation Fuzzy Systems and Knowledge Discovery (ICNC-FSKD), Huangshan, China, 2018, 390–3. <https://doi.org/10.1109/fskd.2018.8687254>
- Johnston B, Chazal P. A review of image-based automatic facial landmark identification techniques. *EURASIP J Imag Video Proces* 2018;21:86. <https://doi.org/10.1186/s13640-018-0324-4>
- Yamashita R, Nishio M, Do RK, et al. Convolutional neural network: an overview and application in radiology. *Insights Imaging* 2018;9:611–29. <https://doi.org/10.1007/s13244-018-0639-9>
- Savran A, Alyz N, Dibeklioglu H, et al. Bosphorus database for 3D face analysis. *Lect Note Comput Sci* 2008;5373:47–56. https://doi.org/10.1007/978-3-540-89991-4_6
- Zhang Q, Xiao J, Tian C, et al. A robust deformed convolutional neural network (CNN) for image denoising. *CAAI Trans Intell Technol* 2022;8:331–40. <https://doi.org/10.1016/j.ijleo.2021.166977>
- Jin H, Che H, Chin H. Unsupervised domain adaptation for anatomical landmark detection. *Int J Comput Vision* 2021;129:3174–94. https://doi.org/10.1007/978-3-031-43907-0_66
- Palmer R, Helmholz P, Baynam G. Cliniface: phenotypic visualisation and analysis using non-rigid registration of 3d facial images. *Int Arch Photogramm Remote Sens Spat Inf Sci* 2020;43:301–8. <https://doi.org/10.5194/isprs-archives-XLIII-B2-2020-301-2020>
- Torres HR, Morais P, Fritze A, et al. Anthropometric landmark detection in 3D head surfaces using a deep learning approach. *IEEE J Biomed Health Inform* 2020;25:2643–54. <https://doi.org/10.1109/jbhi.2020.3035888>
- Wen A, Zhu Y, Xiao N, et al. Comparison study of extraction accuracy of 3D facial anatomical landmarks based on non-rigid registration of face template. *Diagnostics (Basel, Switzerland)* 2023;13:1086. <https://doi.org/10.3390/diagnostics13061086>
- Plooi J, Swennen G, Rangel F, et al. Evaluation of reproducibility and reliability of 3D soft tissue analysis using 3D stereophotogrammetry. *Int J Oral Maxillofac Surg* 2009;38:273. <https://doi.org/10.1016/j.ijom.2008.12.009>
- Liang S, Wu J, Weinberg SM, et al. Improved detection of landmarks on 3D human face data. *Annu Int Conf IEEE Eng Med Biol Soc* 2013;3:6482–5. <https://doi.org/10.1371/journal.pone.0176566>
- Sunko FM, Waddington JL, Whelan PF. 3-D facial landmark localization with asymmetry patterns and shape regression from incomplete local features. *IEEE Trans Cybernetics* 2014;45:1717–30. <https://doi.org/10.1109/TCYB.2014.2359056>
- Bannister J, Crites S, Aponte JD, et al. Fully automatic landmarking of syndromic 3D facial surface scans using 2D images. *Sensors* 2020;11:3171–4. <https://doi.org/10.3390/s20113171>

30. Guarin DL, Taati B, Hadlock T, *et al.* Automatic facial landmark localization in clinical populations - improving model performance with a small dataset. *Res Sq* 2020. <https://doi.org/10.21203/rs.3.rs-19405/v1>
31. Asgarian A, Zhao S, Ashraf AB, *et al.* Limitations and biases in facial landmark detection, an emperical study on older adults with dementia. *CVPR workshops*, 2019. <https://doi.org/10.48550/arXiv.1905.07446>
32. Berends B, Bielevelt F, Schreurs R, *et al.* Fully automated landmarking and facial segmentation on 3D photographs. *Sci Rep* 2024;**14**:6463. <https://doi.org/10.1038/s41598-024-56956-9>
33. Weinberg S, Raffensperger Z, Marazita M. The 3D facial norms database: Part 1. A web-based craniofacial anthropometric and image repository for the clinical and research community. *Cleft Palate Craniofac J* 2016;**53**:e185–97. <https://doi.org/10.1597/15-199>



E 100 S

100th European Orthodontic Society Congress 2-6 June 2025, Krakow, Poland

Register at early-bird discounted rates by 6 April 2025.

KEYNOTE SPEAKERS:



Visit the website to learn more:
congress.eoseurope.org

With thanks to our sponsors

Platinum Sponsors

align | invisalign | iTero | exocad

DW Lingual Systems GmbH

solventum

Follow us



#EOScongress25

Enquiries: EOScongress@tfiledestar.com

Gold Sponsor



Silver Sponsor

Ormco



Bronze Sponsor

Leone ORTHERO

DentalMonitoring ortoplus

Smartee HuFriedyGroup