Original Article

pISSN 2234-7518 • eISSN 2005-372X
https://doi.org/10.4041/kjod21.248
Korean J Orthod 2022;52(4):287-297

THE KOREAN JOURNAL of ORTHODONTICS
KJO

Accuracy of artificial intelligence-assisted landmark identification in serial lateral cephalograms of Class III patients who underwent orthodontic treatment and two-jaw orthognathic surgery

Mihee Hong[a,b]
Inhwan Kim[c]
Jin-Hyoung Cho[d]
Kyung-Hwa Kang[e]
Minji Kim[f]
Su-Jung Kim[g]
Yoon-Ji Kim[h]
Sang-Jin Sung[h]
Young Ho Kim[i]
Sung-Hoon Lim[j]
Namkug Kim[k]
Seung-Hak Baek[a]

[a]Department of Orthodontics, School of Dentistry, Dental Research Institute, Seoul National University, Seoul, Korea
[b]Department of Orthodontics, School of Dentistry, Kyungpook National University, Daegu, Korea
[c]Department of Convergence Medicine, Asan Medical Institute of Convergence Science and Technology, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Korea
[d]Department of Orthodontics, Chonnam National University School of Dentistry, Gwangju, Korea
[e]Department of Orthodontics, School of Dentistry, Wonkwang University, Iksan, Korea
[f]Department of Orthodontics, College of Medicine, Ewha Womans University, Seoul, Korea
[g]Department of Orthodontics, Kyung Hee University School of Dentistry, Seoul, Korea
[h]Department of Orthodontics, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Korea
[i]Department of Orthodontics, Institute of Oral Health Science, Ajou University School of Medicine, Suwon, Korea
[j]Department of Orthodontics, College of Dentistry, Chosun University, Gwangju, Korea
[k]Department of Convergence Medicine, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Korea

Objective: To investigate the pattern of accuracy change in artificial intelligence-assisted landmark identification (LI) using a convolutional neural network (CNN) algorithm in serial lateral cephalograms (Lat-cephs) of Class III (C-III) patients who underwent two-jaw orthognathic surgery. **Methods:** A total of 3,188 Lat-cephs of C-III patients were allocated into the training and validation sets (3,004 Lat-cephs of 751 patients) and test set (184 Lat-cephs of 46 patients; subdivided into the genioplasty and non-genioplasty groups, n = 23 per group) for LI. Each C-III patient in the test set had four Lat-cephs: initial (T0), pre-surgery (T1, presence of orthodontic brackets [OBs]), post-surgery (T2, presence of OBs and surgical plates and screws [S-PS]), and debonding (T3, presence of S-PS and fixed retainers [FR]). After mean errors of 20 landmarks between human gold standard and the CNN model were calculated, statistical analysis was performed. **Results:** The total mean error was 1.17 mm without significant difference among the four time-points (T0, 1.20 mm; T1, 1.14 mm; T2, 1.18 mm; T3, 1.15 mm). In comparison of two time-points ([T0, T1] vs. [T2, T3]), ANS, A point, and B point showed an increase in error ($p < 0.01$, 0.05, 0.01, respectively), while Mx6D and Md6D showeda decrease in error (all $p < 0.01$). No difference in errors existed at B point, Pogonion, Menton, Md1C, and Md1R between the genioplasty and non-genioplasty groups. **Conclusions:** The CNN model can be used for LI in serial Lat-cephs despite the presence of OB, S-PS, FR, genioplasty, and bone remodeling.

Key words: Convolutional neural network, Landmark identification, Two-jaw orthognathic surgery, Serial lateral encephalogram

Received September 24, 2021; Revised March 7, 2022; Accepted March 11, 2022.

Corresponding author: Seung-Hak Baek.
Professor, Department of Orthodontics, School of Dentistry, Dental Research Institute, Seoul National University, 101, Daehak-ro, Jongno-gu, Seoul 03080, Korea.
Tel +82-2-2072-3952 e-mail drwhite@unitel.co.kr
Corresponding author: Namkug Kim.
Professor, Department of Convergence Medicine, Asan Medical Center, University of Ulsan College of Medicine, 88, Olympic-ro 43-gil, Songpa-gu, Seoul 05505, Korea.
Tel +82-2-3010-6573 e-mail namkugkim@gmail.com

Mihee Hong and Inhwan Kim contributed equally to this work (as co-first authors).

How to cite this article: Hong M, Kim I, Cho JH, Kang KH, Kim M, Kim SJ, Kim YJ, Sung SJ, Kim YH, Lim SH, Kim N, Baek SH. Accuracy of artificial intelligence-assisted landmark identification in serial lateral cephalograms of Class III patients who underwent orthodontic treatment and two-jaw orthognathic surgery. Korean J Orthod 2022;52(4):287-297. https://doi.org/10.4041/kjod21.248

INTRODUCTION

Owing to the high prevalence of Class III malocclusion and negative social recognition of the prognathic appearance,[1,2] Korea has become one of the countries that performs two-jaw orthognathic surgery (TJ-OGS) extensively in patients with skeletal Class III malocclusion. To obtain a successful treatment outcome, the following four steps should be performed precisely: (1) diagnosis and gross treatment planning for pre-surgical orthodontic treatment and orthognathic surgery using initial cephalograms, (2) planning for the direction and amount of surgical movement using pre-surgical cephalograms, (3) assessment of surgical outcome and planning for post-surgical orthodontic treatment using post-surgical cephalograms, and (4) comprehensive assessment of orthodontic treatment and orthognathic surgery using debonding cephalograms.[3,4] Furthermore, superimposition of serial cephalograms taken at different time-points is also important to assess the outcomes of pre- and post-surgical orthodontic treatment and orthognathic surgery. Accurate detection of cephalometric landmarks is mandatory to perform these procedures.

An artificial intelligence (AI) algorithm including convolutional neural network (CNN) can help clinicians detect cephalometric landmarks with an accuracy that is close to that of human experts.[5-12] Previous AI studies have regarded the accuracy within a range of 2 mm as a clinically acceptable performance in landmark identification.[8,12-15] However, it appears to be a lenient standard for appropriate clinical use. Therefore, use of stricter criteria (i.e., range within at least 1.5 mm) is necessary in determining the accuracy of landmark identification for clinical relevance.

In addition, most AI studies on the accuracy of automated landmark identification[8,13-15] have trained and tested their models using initial lateral cephalograms only, which do not have orthodontic brackets (OB), surgical plates and screws (S-PS), fixed retainer (FR), and bone remodeling changes. To the best of our knowledge, no study has compared the accuracy of automated landmark identification in serial cephalograms at the four time-points covering from the initial, pre-surgery, post-surgery, to debonding stages in orthognathic surgery cases. Therefore, the purpose of the study was to investigate the pattern of accuracy change in AI-assisted landmark identification in serial lateral cephalograms of Class III patients who underwent pre- and post-surgical orthodontic treatment and TJ-OGS using a cascade CNN algorithm and strict criteria for determining the degree of accuracy.

MATERIALS AND METHODS

Data set

A total of 3,188 lateral cephalograms of 797 patients with Class III malocclusion were used for the training and validation sets and the test set for automated landmark identification using the CNN model. The inclusion criteria were as follows: (1) Class III patient who underwent pre- and post-surgical orthodontic treatment and TJ-OGS with/without genioplasty and (2) Class III patient whose serial lateral cephalograms were available. The exclusion criterion was Class III patient who had craniofacial deformities.

The training and validation sets for automated landmark identification by the CNN model included 3,004 lateral cephalograms of 751 Class III patients from 10 institutions (Table 1). Some of the patients who belonged to the training or validation set had more than

Table 1. Composition of the training, validation, and test dataset

	Institution	Training set	Validation set	Test set	Sum
	Seoul National University Dental Hospital	1,292	100	52	1,444
	Kyung Hee University Dental Hospital	607	100	48	755
	Kyungpook National University Dental Hospital	133	30	20	183
	Asan Medical Center	144	32	24	200
	Ewha University Medical Center	116	20	12	148
	Wonkwang University Dental Hospital	95	26	8	129
	Ajou University Dental Hospital	84	20	12	116
	Korea University Anam Hospital	62	25	0	87
	Chonnam National University Dental Hospital	48	16	8	72
	Chosun University Dental Hospital	45	9	0	54
Total	Lateral cephalograms	2,626	378	184	3,188
	Class III patients	751*		46	797

*Class III patients had various numbers of lateral cephalograms, which belonged to the training or validation set.

https://doi.org/10.4041/kjod21.248

four lateral cephalograms because additional progress lateral cephalograms were taken between time-points, while some of them had missing lateral cephalograms at specific timepoints.

For the test set, Class III patients with cephalograms obtained at the following timepoints were selected: initial (T0), pre-surgery (T1, taken at least 1 month before TJ-OGS; presence of OBs), post-surgery (T2, taken at least 2 months after TJ-OGS; presence of OBs and S-PS), and debonding (T3, presence of S-PS, FR, and bone remodeling change). As a result, the test set consisted of 184 cephalograms of 46 Class III patients from eight institutions (Table 1). It was subdivided into the genioplasty and non-genioplasty groups (n = 23 patients per group). Their characteristics are enumerated in Figure 1.

Ethical approval

This nationwide multicenter study was reviewed and approved by the Institutional Review Board (IRB) Committee of 10 institutions: Seoul National University Dental Hospital (ERI18002), Kyung Hee University Dental Hospital (KH-DT19006), Kyungpook National University Dental Hospital (KNUDH-2019-03-02-00), Asan Medical Center (2019-0408), Ewha University Medical Center (EUMC 2019-04-017-009), Wonkwang University Dental Hospital (WKDIRB201903-01), Ajou University Dental Hospital (AJIRB-MED-MDB-19-039), Korea University Anam Hospital (K2019-0543-010), Chonnam National University Dental Hospital (CNUDH-EXP-2021-001), and Chosun University Dental Hospital (CUDHIRB 1901 005 R01).

Cascade CNN

Data sets were obtained from 10 centers using anonymized Digital Imaging and Communications in Medicine (DICOM) file format. Since finding the exact location of landmarks in a large lateral cephalogram image is relatively difficult, a fully automated landmark prediction algorithm with the cascade network was developed.[12] Two steps were followed: 1) detection of the region of interest (256 × 256 and 512 × 512 pixels depending on the landmark) using the RetinaNet[16] and 2) prediction of the landmark using the U-Net[17] (Figure 2).

Cephalometric landmarks

Definitions of 12 skeletal and eight dental landmarks are presented in Figure 3 and Table 2. The landmarks were digitized by a single orthodontist who had 20 years of experience (human gold standard, HMH) and by the CNN model.

Measurement variables (Table 3)

The mean values of absolute errors for each landmark were calculated using the absolute distance between the human gold standard and AI-assisted detection. The degree of error was allocated into excellent (< 1.0 mm), good (1.0–1.5 mm), fair (1.5–2.0 mm), acceptable (2.0–2.5 mm), and unacceptable (> 2.5 mm) groups. Then, the accuracy percentage (AP) was calculated using

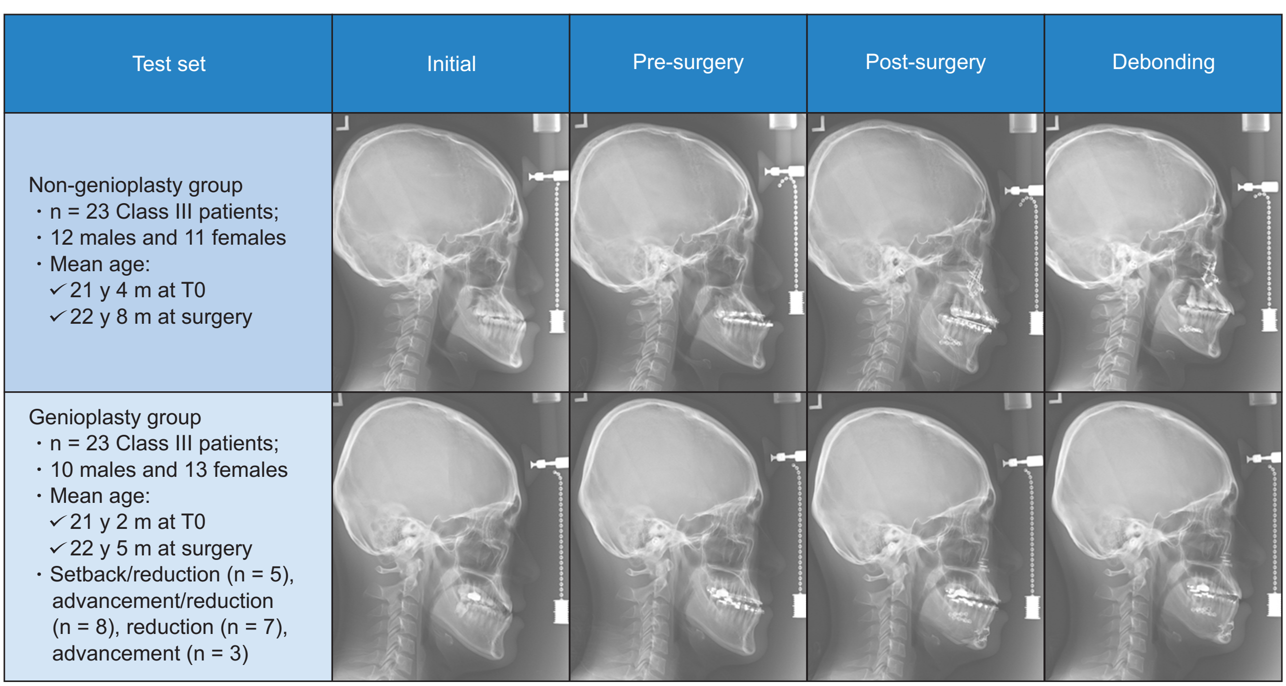

Figure 1. Composition of the test set.
T0, initial.

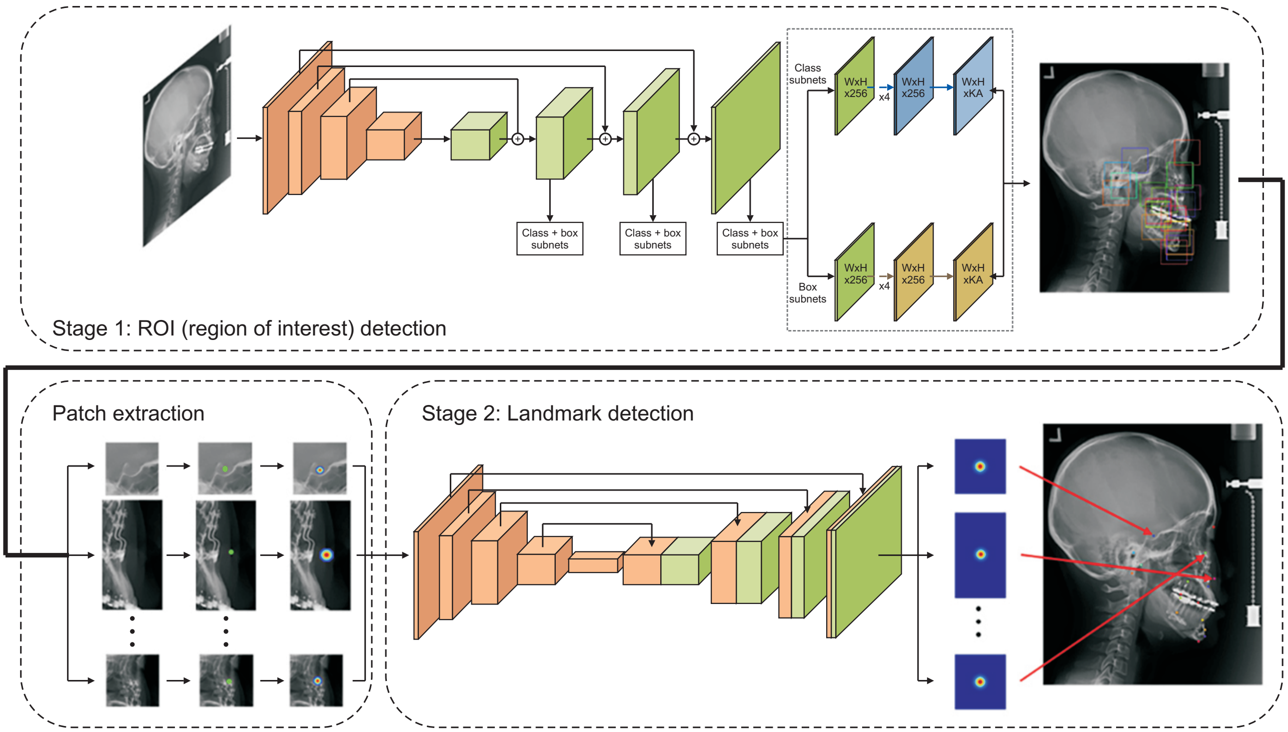

Figure 2. General schematic of the cascade convolution neural network algorithm for artificial intelligence-assisted landmark identification.

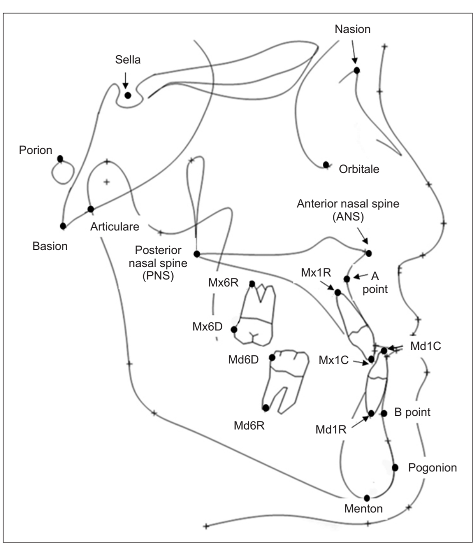

Figure 3. The skeletal and dental landmarks.
See Table 2 for definitions of the other landmarks.

a formula (percentage of the excellent and good groups among the total degree of error groups), which means that the error range within 1.5 mm was considered accurate. The degree of accuracy was defined as “very high” (AP > 90%), “high” (AP, 70–90%), “medium” (AP, 50–70%), and “low” (AP < 50%).

Intra-examiner reliability

Twenty randomly selected lateral cephalogram images were re-digitized with an interval of 2 weeks by the same operator (HMH). Since no significant difference was found in the values of the x- and y-coordinates between the first and second measurements in the Wilcoxon signed rank test ($p > 0.05$), the first set of measurements was used for further analysis.

Statistics

Repeated measures analysis of variance (ANOVA), and post-hoc test for within-subject by Tukey’s adjustment for multiple comparisons were performed to find out the difference between T0, T1, T2, and T3 stages. Repeated measures multivariate analysis of variance (MANOVA) was performed to compare between ‘before-surgery group’, including T0 and T1, and ‘after-surgery group’, including T2 and T3. Statistical analysis was done using SPSS ver. 23.0 (IBM Corp., Armonk, NY, USA) and SAS 9.4 (SAS Institute Inc., Cary, NC, USA.) and p-values of < 0.05 were considered statistically significant.

RESULTS

Evaluation of total landmarks (Table 3)

The total landmarks showed a good mean error value (1.17 mm), and the total AP had a high degree of accu-

Table 2. The definition of cephalometric landmarks

Compartment			Landmark	Description
Skeletal landmark	Cranial base		Nasion (N)	The most anterior point on the frontonasal suture in the midsagittal plane
			Sella (S)	Center of the Sella Turcica
			Porion (Por)	The most superior point of the external auditory meatus
			Orbitale (Or)	The most inferior point of the orbital cavity contour
			Basion (Ba)	The most posterior and inferior point of the occipital bone
	Maxilla	Anterior	ANS	The tip of anterior nasal spine
			A point	The deepest point between ANS and the upper incisal alveolus
		Posterior	PNS	The most posterior point of the hard palate
	Mandible	Anterior	B point	The deepest point between Pogonion and the lower incisal alveolus
			Pogonion (Pog)	The most anterior point on the symphysis
		Posterior	Articulare (Ar)	Intersection between the inferior cranial base surface and the posterior surface of condyle
		Bottom	Menton (Me)	The most inferior point on the symphysis
Dental landmark	Maxillary dentition	Anterior	Mx1C	Crown tip of the maxillary central incisor
			Mx1R	Root apex of the maxillary central incisor
		Posterior	Mx6D	Distal contact point of the maxillary first molar
			Mx6R	Distobuccal root apex of the maxillary first molar
	Mandibular dentition	Anterior	Md1C	Crown tip of the mandibular central incisor
			Md1R	Root apex of the mandibular central incisor
		Posterior	Md6D	Distal contact point of the mandibular first molar
			Md6R	Distal root apex of the mandibular first molar

racy (74.2%).

Evaluation of skeletal landmarks (Table 3)

Nasion and Sella showed an excellent mean error value and a very high degree of accuracy (0.59 mm and 95.1%; 0.46 mm and 100%, respectively), while Porion and Orbitale showed a good mean error value and a high degree of accuracy (1.07 mm and 76.1%; 1.21 mm and 73.9%, respectively). On the other hand, Basion showed a fair mean error value (1.64 mm) and a medium degree of accuracy (63.1%).

ANS and A point showed a good mean error value and a medium degree of accuracy (1.39 mm and 65.2%; 1.41 mm and 63.0%, respectively). PNS had a good mean error value (1.19 mm) and a high degree of accuracy (72.7%).

Pogonion, Menton, and Articulare showed an excellent mean error value and a very high degree of accuracy (0.79 mm and 91.3%; 0.77 mm and 93.5%; and 0.77 mm and 93.5%, respectively). B point showed a good mean error value (1.15 mm) and a high degree of accuracy (77.2 %).

Evaluation of dental landmarks (Table 3)

Mx1C showed an excellent mean error value (0.44 mm) and a very high degree of accuracy (97.8%), while Mx6D had a good mean error value (1.43 mm) and a medium degree of accuracy (64.1%). On the other hand, Mx1R and Mx6R had a fair mean error value and a medium degree of accuracy (1.55 mm and 57.6%; 1.68 mm and 51.6%, respectively).

Md1C demonstrated an excellent mean error value (0.49 mm) and a very high degree of accuracy (97.3%), while Md1R had a fair mean error value (1.57 mm) and a medium degree of accuracy (58.2%). Md6D had a fair mean error value (1.67 mm) and medium accuracy (51.6%), and Md6R exhibited an acceptable mean error value (2.03 mm) and a low degree of accuracy (41.3%).

Table 3. The absolute values of error, distribution of error, accuracy percentage, and degree of accuracy for each landmark

Compartment			Landmark	Absolute value of error (Err)	Distribution					Accuracy	
				Mean ± SD (mm)	Excellent (Err <1.0 mm)	Good (1.0 ≤ Err < 1.5 mm)	Fair (1.5 ≤ Err < 2.0 mm)	Acceptable (2.0 ≤ Err < 2.5 mm)	Unacceptable (2.5 mm ≤ Err)	Accuracy percentage	Degree of accuracy
Skeletal landmark	Cranial base		Nasion	0.59 ± 0.48	157 (85.3)	18 (9.8)	4 (2.2)	3 (1.6)	2 (1.1)	175 (95.1)	Very high
			Sella	0.46 ± 0.23	180 (97.8)	4 (2.2)	0 (0.0)	0 (0.0)	0 (0.0)	184 (100)	Very high
			Porion	1.07 ± 0.69	103 (56.0)	37 (20.1)	24 (13.0)	14 (7.6)	6 (3.3)	140 (76.1)	High
			Orbitale	1.21 ± 1.01	92 (50.0)	44 (23.9)	21 (11.4)	12 (6.5)	15 (8.2)	136 (73.9)	High
			Basion	1.64 ± 1.61	82 (44.6)	34 (18.5)	21 (11.4)	13 (7.1)	34 (18.5)	116 (63.1)	Medium
	Maxilla	Anterior	ANS	1.39 ± 1.01	72 (39.1)	48 (26.1)	23 (12.5)	14 (7.6)	27 (14.7)	120 (65.2)	Medium
			A point	1.41 ± 0.99	86 (46.7)	30 (16.3)	22 (12.0)	18 (9.8)	28 (15.2)	116 (63.0)	Medium
		Posterior	PNS	1.19 ± 0.89	97 (52.7)	37 (20.1)	24 (13.0)	13 (7.1)	13 (7.1)	134 (72.7)	High
	Mandible	Anterior	B point	1.15 ± 0.96	106 (57.6)	36 (19.6)	21 (11.4)	8 (4.3)	13 (7.1)	142 (77.2)	High
			Pogonion	0.79 ± 0.68	140 (76.1)	28 (15.2)	7(3.8)	1 (0.5)	8 (4.3)	168 (91.3)	Very high
		Bottom	Menton	0.77 ± 0.44	143 (77.7)	29 (15.8)	8 (4.3)	3 (1.6)	1 (0.5)	172 (93.5)	Very high
		Posterior	Articulare	0.77 ± 0.45	138 (75.0)	34 (18.5)	10 (5.4)	1 (0.5)	1 (0.5)	172 (93.5)	Very high
Dental landmark	Maxillary dentition	Anterior	Mx1C	0.44 ± 0.37	178 (96.7)	2 (1.1)	2 (1.1)	1 (0.5)	1 (0.5)	180 (97.8)	Very high
			Mx1R	1.55 ± 1.09	63 (34.2)	43 (23.4)	36 (19.6)	12 (6.5)	30 (16.3)	106 (57.6)	Medium
		Posterior	Mx6D	1.43 ± 1.08	76 (41.3)	42 (22.8)	23 (12.5)	17 (9.2)	26 (14.1)	118 (64.1)	Medium
			Mx6R	1.68 ± 1.08	51 (27.7)	44 (23.9)	35 (19.0)	19 (10.3)	35 (19.0)	95 (51.6)	Medium
	Mandibular dentition	Anterior	Md1C	0.49 ± 0.64	172 (93.5)	7 (3.8)	0 (0.0)	2 (1.1)	3 (1.6)	179 (97.3)	Very high
			Md1R	1.57 ± 1.04	64 (34.8)	43 (23.4)	29 (15.8)	18 (9.8)	30 (16.3)	107 (58.2)	Medium
		Posterior	Md6D	1.67 ± 1.24	54 (29.3)	41 (22.3)	33 (17.9)	30 (16.3)	26 (14.1)	95 (51.6)	Medium
			Md6R	2.03 ± 1.35	46 (25.0)	30 (16.3)	25 (13.6)	28 (15.2)	55 (29.9)	76 (41.3)	Low
Total				1.17 ± 1.04	2,100 (57.1)	631 (17.1)	368 (10.0)	227 (6.2)	354 (9.6)	2,731 (74.2)	High

Values are presented as number (%).
Err, absolute value of error; SD, standard deviation; Accuracy percentage (AP); error range within 1.5 mm was considered accurate.
The degree of accuracy was defined as very high (90% ≤ AP), high (70% ≤ AP < 90%), medium (50% ≤ AP < 70%), and low (AP < 50%).
See Table 2 for definitions of each landmark.

Table 4. Comparison of mean errors among the four time-points (T0, T1, T2 and T3) and between two time-points ([T0, T1] vs. [T2, T3])

Compartment			Landmark	Initial stage (T0)	Pre-surgery stage (T1)	Post-surgery stage (T2)	Debonding stage (T3)	Multiple comparison			
								Among T0, T1, T2 and T3 stages†		(T0, T1) vs. (T2, T3)‡	
				Mean ± SD (mm)	Mean ± SD (mm)	Mean ± SD (mm)	Mean ± SD (mm)	*p*-value	Tukey HSD test	*p*-value	Contrast matrix with the MANOVA
Skeletal landmark	Cranial base		Nasion	0.58 ± 0.42	0.59 ± 0.48	0.55 ± 0.39	0.64 ± 0.60	0.698		0.852	
			Sella	0.48 ± 0.23	0.43 ± 0.19	0.41 ± 0.20	0.50 ± 0.27	0.155		0.986	
			Porion	1.04 ± 0.55	1.07 ± 0.76	1.17 ± 0.85	1.01 ± 0.57	0.493		0.566	
			Orbitale	1.19 ± 0.88	1.15 ± 0.88	1.39 ± 1.40	1.10 ± 0.76	0.454		0.618	
			Basion	1.41 ± 1.32	1.59 ± 1.60	1.95 ± 1.94	1.61 ± 1.52	0.148		0.092	
	Maxilla	Anterior	ANS	1.07 ± 0.78	1.22 ± 0.97	1.78 ± 1.22	1.49 ± 0.87	0.003**	T0^{a}, T1^{a}, T2^{b}, and T3^{ab}	0.003**	(T0, T1) < (T2, T3)
			A point	1.27 ± 0.89	1.28 ± 0.78	1.50 ± 1.07	1.59 ± 1.16	0.151		0.040*	(T0, T1) < (T2, T3)
		Posterior	PNS	1.16 ± 0.79	1.14 ± 0.87	1.29 ± 1.09	1.17 ± 0.82	0.823		0.587	
	Mandible	Anterior	B point	1.00 ± 0.97	1.01 ± 0.61	1.29 ± 1.24	1.31 ± 0.91	0.142		0.008**	(T0, T1) < (T2, T3)
			Pogonion	0.66 ± 0.48	0.80 ± 0.72	0.82 ± 0.69	0.86 ± 0.77	0.277		0.260	
		Bottom	Menton	0.83 ± 0.52	0.70 ± 0.39	0.74 ± 0.38	0.82 ± 0.45	0.298		0.786	
		Posterior	Articulare	0.76 ± 0.39	0.75 ± 0.42	0.73 ± 0.40	0.84 ± 0.58	0.540		0.616	
Dental landmark	Maxillary dentition	Anterior	Mx1C	0.48 ± 0.37	0.49 ± 0.55	0.41 ± 0.26	0.38 ± 0.18	0.355		0.096	
			Mx1R	1.83 ± 1.24	1.37 ± 1.14	1.56 ± 1.02	1.54 ± 1.17	0.166		0.714	
		Posterior	Mx6D	1.66 ± 1.18	1.63 ± 1.31	1.20 ± 0.80	1.23 ± 0.88	0.032*	T0^{b}, T1^{ab}, T2^{a}, and T3^{ab}	0.008**	(T2, T3) < (T0, T1)
			Mx6R	1.89 ± 1.15	1.65 ± 1.08	1.57 ± 1.03	1.60 ± 1.05	0.349		0.194	
	Mandibular dentition	Anterior	Md1C	0.48 ± 0.45	0.62 ± 1.10	0.47 ± 0.39	0.37 ± 0.31	0.215		0.096	
			Md1R	1.77 ± 1.22	1.52 ± 1.00	1.37 ± 0.96	1.63 ± 0.95	0.303		0.400	
		Posterior	Md6D	2.15 ± 1.79	1.71 ± 0.91	1.51 ± 1.14	1.33 ± 0.75	0.008**	T0^{b}, T1^{ab}, T2^{a}, and T3^{a}	0.003**	(T2, T3) < (T0, T1)
			Md6R	2.21 ± 1.64	2.08 ± 1.36	1.94 ± 1.25	1.89 ± 1.12	0.579		0.242	
Total				1.20 ± 1.10	1.14 ± 1.02	1.18 ± 1.09	1.15 ± 0.95	0.376		0.895	

$^{a, b, ab}$: 'a' and 'b' indicate statically significant difference between a and b; while, 'ab' indicates that there was no significant difference between a and ab and between ab and b.
T0, initial stage; T1, pre-surgery stage (presence of orthodontic brackets [OBs]); T2, post-surgery stage (presence of OBs and surgical plates and screws [S-PS]); T3, debonding stage (presence of S-PS and fixed retainers); SD, standard deviation; HSD, honestly significant difference.
**p* < 0.05; ***p* < 0.01.
†Repeated measures analysis of variance (ANOVA) test and post-hoc test for 'within-subject' by Tukey's adjustment for multiple comparisons were performed.
‡Repeated measures multivariate analysis of variance (MANOVA) test was performed.
See Table 2 for definitions of each landmark.

Comparison of the mean errors among the four timepoints (T0, T1, T2, and T3) (Table 4)

No significant difference was found in the overall mean errors ($p > 0.05$). Only three landmarks, namely ANS, Mx6D, and Md6D showed a significant difference in the mean errors among the four timepoints (ANS, increase in the mean error from T0 and T1 to T2, $p < 0.01$; Mx6D, decrease in the mean error from T0 to T2, $p < 0.05$; Md6D, decrease in the mean error from T0 to T2 and T3, $p < 0.01$).

Comparison of the mean errors between the two timepoints ([T0, T1] vs. [T2, T3]) (Table 4)

ANS, A point, and B point showed an increase in the mean error after TJ-OGS (ANS, $p < 0.01$; A point, $p < 0.05$; B point, $p < 0.01$), while Mx6D and Md6D showed a decrease in the mean error after TJ-OGS (all $p < 0.01$).

Comparison of the mean errors between the genioplasty and non-genioplasty groups (Table 5)

No significant difference in the mean errors in the landmarks located adjacent to the genioplasty area (B point, Pogonion, Menton, Md1C, and Md1R) existed in each timepoint between the two groups, except Md1R at T1 ($p < 0.05$).

Table 5. Comparison of mean errors in each time-point (from T0 to T3) between the genioplasty and non-genioplasty groups

		Genioplasty group	Non-genioplasty group	*p*-value†
B point	Initial stage (T0)	0.87 ± 0.46	1.13 ± 1.30	0.386
	Pre-surgery stage (T1)	0.99 ± 0.60	1.03 ± 0.63	0.855
	Post-surgery stage (T2)	1.21 ± 0.70	1.37 ± 1.63	0.670
	Debonding stage (T3)	1.25 ± 0.82	1.37 ± 1.01	0.682
	p-value‡	0.184	0.543	
Pog	Initial stage (T0)	0.61 ± 0.38	0.71 ± 0.57	0.468
	Pre-surgery stage (T1)	0.66 ± 0.39	0.95 ± 0.93	0.171
	Post-surgery stage (T2)	0.81 ± 0.82	0.84 ± 0.56	0.898
	Debonding stage (T3)	0.95 ± 0.87	0.77 ± 0.66	0.436
	p-value‡	0.109	0.463	
Menton	Initial stage (T0)	0.71 ± 0.35	0.95 ± 0.63	0.117
	Pre-surgery stage (T1)	0.71 ± 0.47	0.68 ± 0.28	0.813
	Post-surgery stage (T2)	0.68 ± 0.38	0.79 ± 0.38	0.334
	Debonding stage (T3)	0.82 ± 0.48	0.81 ± 0.43	0.926
	p-value‡	0.578	0.186	
Md1C	Initial stage (T0)	0.58 ± 0.53	0.39 ± 0.33	0.149
	Pre-surgery stage (T1)	0.38 ± 0.18	0.87 ± 1.52	0.138
	Post-surgery stage (T2)	0.39 ± 0.25	0.56 ± 0.48	0.155
	Debonding stage (T3)	0.30 ± 0.14	0.44 ± 0.41	0.137
	p-value‡	0.062	0.156	
Md1R	Initial stage (T0)	1.64 ± 1.13	1.90 ± 1.32	0.484
	Pre-surgery stage (T1)	1.22 ± 0.68	1.83 ± 1.18	0.040*
	Post-surgery stage (T2)	1.61 ± 1.03	1.12 ± 0.84	0.082
	Debonding stage (T3)	1.53 ± 0.88	1.73 ± 1.03	0.484
	p-value‡	0.380	0.091	

Values are presented as mean ± standard deviation.
*$p < 0.05$.
†Comparison between genioplasty and nongenioplasty groups by independent *t*-test.
‡Comparison mean error among T0, T1, T2, and T3 stages by repeated measures analysis of variance (ANOVA).
See Table 2 for definitions of each landmark.

https://doi.org/10.4041/kjod21.248

DISCUSSION

Since TJ-OGS induces the position change and bone remodeling in the skeletal structures and produces the metallic images of the OB, SP-S, and FR, the accuracy and reliability of cephalometric landmark identification in serial lateral cephalograms are important for assessment of treatment outcomes.[18]

As total landmarks exhibited a good mean error value and a high degree of accuracy (1.17 mm and 74.2%, respectively, Table 3) without significant difference among the four time-points (T0, 1.20 mm; T1, 1.14 mm; T2, 1.18 mm; T3, 1.15 mm; $p > 0.05$, Table 4), accuracy of the AI-assisted digitization was not significantly affected by the presence of OB, SP-S, FR, and bone remodeling change during orthodontic treatment and TJ-OGS.

Regardless of the degree of accuracy of each landmark (Table 3), none of the five cranial base landmarks exhibited a significant difference in the mean errors among the four time-points (T0, T1, T2, and T3) and between the two time-points ([T0, T1] vs. [T2, T3]) (Table 4). Accuracy of the cranial base landmarks can be regarded as baseline for comparison of serial lateral cephalograms because the positions of these cranial base landmarks are not affected by TJ-OGS.

Three error patterns were found in the maxillary skeletal landmarks. First, the mean errors of ANS were different among the four time-points (T0, 1.07 mm; T1, 1.22 mm; T2, 1.78 mm; T3, 1.49 mm; $p < 0.01$, Table 4) and presented an increased error value after TJ-OGS than before it ([T0, T1] vs. [T2, T3]; $p < 0.01$, Table 4). This suggested that the metal image of the SP-S adjacent to ANS as well as surgical shape modification of ANS[19,20] (Figure 1) could affect the accuracy of AI-assisted landmark detection. Second, although the error of A point was not significantly different among the four time-points (T0, 1.27 mm; T1, 1.28 mm; T2, 1.50 mm; T3, 1.59 mm; Table 4), it presented an increase in the mean error value after TJ-OGS than before it ([T0, T1] vs. [T2, T3]; $p < 0.05$, Table 4). This occurred because A point might be less affected by the metal image of the SP-S installed at the maxilla and have a lower chance for surgical shape modification, compared to ANS (Figure 1). Furthermore, A point might be less affected by the metal image of SP-S installed lateral to the pyriform aperture in the maxilla and have a lower chance for surgical shape modification relative to ANS. Third, in case of posterior impaction and/or anteroposterior movement of the maxilla, the position of PNS had to be changed. However, for PNS, no significant difference was found either among the four time-points (T0, 1.16 mm; T1, 1.14 mm; T2, 1.29 mm; T3, 1.17 mm; $p > 0.05$, Table 4) or between the two time-points ([T0, T1] vs. [T2, T3]; $p > 0.05$, Table 4). No significant difference in accuracy between time points means that the amount error of landmark at four or two timepoints was neither significantly increased nor decreased. This might be due to (1) an absence of the metal image of the SP-S within the region of interest of PNS and (2) an easily defined the end point of the hard palate.

There are three explanations for the errors in the mandibular skeletal landmarks. First, since there were no metal images within the region of interest of Articulare and Menton, their mean errors were not significantly different among the four time-points and between the two time-points (all $p > 0.05$, Table 4). Second, the mean error of Pogonion was not significantly different among the four time-points and between the two time-points ($p > 0.05$, Table 4), which suggests that the metal image of the SP-S adjacent to Pognion (Figure 1) might not affect the accuracy of AI-assisted landmark detection. Third, although the mean errors of B point did not significantly differ among the four time-points (T0, 1.00 mm; T1, 1.01 mm; T2, 1.29 mm; T3, 1.31 mm; $p > 0.05$, Table 4), comparison of the two time-points revealed an increase in error after TJ-OGS ([T0, T1] vs. [T2, T3]; $p < 0.01$, Table 4). These findings suggest that the metal image of the SP-S adjacent to the B point (Figure 1) might affect the accuracy of AI-assisted landmark detection.

There are two sources of errors in the dental landmarks. First, regardless of the degree of accuracy in the dental landmarks (Table 3), Mx1C, Md1C, Mx1R, Md1R, Mx6R, and Md6R did not exhibit significant difference in the mean errors among the four time-points and between the two time-points (all $p > 0.05$, Table 4). Second, the mean errors of Mx6D and Md6D were significantly different among the four time-points (Mx6D: T0, 1.66 mm; T1, 1.63 mm; T2, 1.20 mm; T3, 1.23 mm; Md6D, T0, 2.15 mm; T1, 1.71 mm; T2, 1.51 mm; T3, 1.33 mm; all $p < 0.01$, Table 4) and presented decreased mean error values after TJ-OGS ([T0, T1] vs. [T2, T3]; all $p < 0.01$, Table 4). The possible reasons for these might be the following: (1) Horizontal and vertical overlapping of the right and left maxillary and mandibular first molars made it difficult to accurately locate the Mx6D and Mn6D at T0 lateral cephalogram; and (2) Orthodontic treatment and TJ-OGS improved the alignment of the maxillary and mandibular dentition and corrected the cant, shift and yaw of the maxilla and mandible, reducing the double images of the maxillary and mandibular first molars.

No significant difference was found in the mean errors in the landmarks adjacent to the genioplasty area including B point, Pogonion, Menton, Md1C, and Md1R (all $p > 0.05$, Table 5). The possible reasons for this are as follows: (1) Menton and Md1C were located relatively far from the SP-S installed at the symphysis and their shapes were not affected by orthognathic surgery; (2)

 https://doi.org/10.4041/kjod21.248

Since Pogonion and B point are the most forward and deepest points on the anterior surface of the symphysis, respectively, they can be easily identified despite the presence of the metal image of the SP-S; and (3) Although Md1R had a fair mean error value and a medium degree of accuracy (1.57 mm and 58.2%, respectively), these patterns were not aggravated at T2 and T3 despite the presence of the metal image of the SP-S.

CONCLUSIONS

• The cascade CNN algorithm proposed in this study showed a possibility of landmark identification from bony anatomies in serial lateral cephalograms despite the presence of OB, S-PS, FR, genioplasty, and bone remodeling.

• However, since Mx1R, Mx6R, Md1R, Md6D, and Md 6R showed more than 1.5 mm of error and less than 60% of AP, it is necessary to increase the accuracy and reliability of landmark identification of the dental landmarks, especially the distal root apex of the mandibular first molar.

• When the AI-assisted landmark identification is used, clinicians should consider these characteristics.

CONFLICTS OF INTEREST

No potential conflict of interest relevant to this article was reported.

ACKNOWLEDGEMENTS

This research was supported by grants from the Korea Health Technology R&D Project through the Korea Health Industry Development Institute and funded by the Ministry of Health &Welfare (HI18C1638) and the Technology Innovation Program (20006105) funded by the Ministry of Trade, Industry & Energy, Republic of Korea.

SUPPLEMENTAL VIDEO

A video presentation of this article is available at https://youtu.be/gGGYjWS7_KQ or www.e-kjo.org.

REFERENCES

1. Im DH, Kim TW, Nahm DS, Chang YI. Current trends in orthodontic patients in Seoul National University Dental Hospital. Korean J Orthod 2003;33:63-72.
2. Piao Y, Kim SJ, Yu HS, Cha JY, Baik HS. Five-year investigation of a large orthodontic patient population at a dental hospital in South Korea. Korean J Orthod 2016;46:137-45.
3. Nielsen IL. Maxillary superimposition: a comparison of three methods for cephalometric evaluation of growth and treatment change. Am J Orthod Dentofacial Orthop 1989;95:422-31.
4. Johnston C, Burden D, Kennedy D, Harradine N, Stevenson M. Class III surgical-orthodontic treatment: a cephalometric study. Am J Orthod Dentofacial Orthop 2006;130:300-9.
5. Hutton TJ, Cunningham S, Hammond P. An evaluation of active shape models for the automatic identification of cephalometric landmarks. Eur J Orthod 2000;22:499-508.
6. Leonardi R, Giordano D, Maiorana F, Spampinato C. Automatic cephalometric analysis. Angle Orthod 2008;78:145-51.
7. Leonardi R, Giordano D, Maiorana F. An evaluation of cellular neural networks for the automatic identification of cephalometric landmarks on digital images. J Biomed Biotechnol 2009;2009:717102.
8. Arık SÖ, Ibragimov B, Xing L. Fully automated quantitative cephalometry using convolutional neural networks. J Med Imaging (Bellingham) 2017; 4:014501.
9. Lee JH, Yu HJ, Kim MJ, Kim JW, Choi J. Automated cephalometric landmark detection with confidence regions using Bayesian convolutional neural networks. BMC Oral Health 2020;20:270.
10. Vandaele R, Aceto J, Muller M, Péronnet F, Debat V, Wang CW, et al. Landmark detection in 2D bioimages for geometric morphometrics: a multi-resolution tree-based approach. Sci Rep 2018;8:538.
11. Khanagar SB, Al-Ehaideb A, Vishwanathaiah S, Maganur PC, Patil S, Naik S, et al. Scope and performance of artificial intelligence technology in orthodontic diagnosis, treatment planning, and clinical decision-making - a systematic review. J Dent Sci 2021;16:482-92.
12. Kim J, Kim I, Kim YJ, Kim M, Cho JH, Hong M, et al. Accuracy of automated identification of lateral cephalometric landmarks using cascade convolutional neural networks on lateral cephalograms from nationwide multi-centres. Orthod Craniofac Res 2021;24 Suppl 2:59-67.
13. Wang CW, Huang CT, Hsieh MC, Li CH, Chang SW, Li WC, et al. Evaluation and comparison of anatomical landmark detection methods for cephalometric X-ray images: a grand challenge. IEEE Trans Med Imaging 2015;34:1890-900.
14. Wang CW, Huang CT, Lee JH, Li CH, Chang SW, Siao MJ, et al. A benchmark for comparison of dental radiography analysis algorithms. Med Image Anal 2016;31:63-76.
15. Hwang HW, Park JH, Moon JH, Yu Y, Kim H, Her SB, et al. Automated identification of cephalometric

landmarks: part 2-Might it be better than human? Angle Orthod 2020;90:69-76.
16. Lin TY, Goyal P, Girshick R, He K, Dollar P. Focal loss for dense object detection. IEEE Trans Pattern Anal Mach Intell 2020;42:318-27.
17. Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation. arXiv. 04597 [Preprint]. 2015 [cited 2020 Dec 15]. Available from: https://doi.org/10.48550/arXiv.1505.04597.
18. Roden-Johnson D, English J, Gallerano R. Comparison of hand-traced and computerized cephalograms: landmark identification, measurement, and superimposition accuracy. Am J Orthod Dentofacial Orthop 2008;133:556-64.
19. Ohba S, Nakao N, Nakatani Y, Yoshimura H, Minamizato T, Kawasaki T, et al. Effects of vertical movement of the anterior nasal spine on the maxillary stability after Le Fort I osteotomy for pitch correction. J Craniofac Surg 2015;26:e481-5.
20. Venkategowda PR, Prakash AT, Roy ET, Shetty KS, Thakkar S, Maurya R. Stability of vertical, horizontal and angular parameters following superior repositioning of maxilla by Le Fort I osteotomy: a cephalometric study. J Clin Diagn Res 2017;11:ZC10-4.