scientific reports

OPEN

Cephalometric landmark detection without X-rays combining coordinate regression and heatmap regression

Kaisei Takahashi[1]✉, Yui Shimamura[2], Chie Tachiki[2], Yasushi Nishii[2] & Masafumi Hagiwara[1]

Fully automated techniques using convolutional neural networks for cephalometric landmark detection have recently advanced. However, all existing studies have adopted X-rays. The problem of direct exposure of patients to X-ray radiation remains unsolved. We propose a model for detecting cephalometric landmarks using only facial profile images without X-rays. First, the model estimates the landmark coordinates using the features of facial profile images through high-resolution representation learning. Second, considering the spatial relationship of the landmarks, the model refines the estimated coordinates. The estimated coordinates are input into fully connected networks to improve the accuracy. During the experiment, a total of 2000 facial profile images collected from 2000 female patients were used. Experiments results suggested that the proposed method may perform at a level equal to or potentially better than existing methods using cephalograms. We obtained an MRE of 0.61 mm for the test data and a mean detection rate of 98.20% within 2 mm. Our proposed two-stage learning method enables a highly accurate estimation of the landmark positions using only facial profile images. The results indicate that X-rays may not be required when detecting cephalometric landmarks.

Quantitative maxillo-facial morphology evaluation is one of the essential steps in orthodontic treatment. In particular, a cephalometric analysis[1] is crucial to evaluate dentofacial proportions, clarify the anatomic basis for malocclusion and establish orthodontic treatment planning. Moreover, that can recognize and evaluate changes brought about by orthodontic treatment. Although dedicated software is usually employed for such an analysis, tracing the maxillo-facial structure and pointing out the landmarks on the lateral cephalograms must be conducted manually by orthodontic specialists. However, these manual procedures are time-consuming and lead to intra- and inter-person variations[2]. Furthermore, the wrong diagnosis caused by inaccurate tracing should induce serious treatment results.

Automated cephalometric analysis systems have recently been developed[2–20]. Previous studies have adopted knowledge bases[5] and pattern matching[3,4] for landmark detection. However, the detection accuracies of these studies are not clinically acceptable[6]. Recently developed algorithms can be divided into two categories: random forest and convolutional neural networks (CNNs). The IEEE International Symposium on Biomedical Imaging (ISBI) , held in 2014 and 2015[12,13] , posed the task of detecting 19 landmarks from lateral cephalograms. At ISBI, most of the studies[14,16] used classical machine learning focused on a random forest. A random forest is usually complex and vulnerable to an overfitting[21]. In the past few years, deep learning methods for landmark detection have outperformed methods using a random forest. In particular, CNN-based methods[2,6–11,17–20] have achieved remarkable results. A CNN is a deep machine learning technique inspired by visual biological recognition, and has been demonstrated to be effective in cephalometric landmark detection[17]. CNN-based methods are often implemented in several stages[2,6–9,11,18–20]. In the first stage, candidate landmarks are identified by searching for local patterns in the cephalograms. In the next stage, the landmarks are finetuned to improve the accuracy. This approach suffers from a performance gap[19] between coordinate-[6,7,10,17,18,20] and heatmap-[2,8,9,11,19], based methods. Coordinate regression methods adopt a regression model to directly predict the *x*- and *y*-coordinates of the landmarks, and it can be expected to make predictions that incorporate the structural knowledge of the landmarks; however they are not as accurate as a heatmap regression method. By contrast, the heatmap

[1]Department of Information and Computer Science, Faculty of Science and Technology, Keio University, Kanagawa 223-8522, Japan. [2]Department of Orthodontics, Tokyo Dental College, Tokyo 101-0061, Japan. ✉email: kaise64taka84@keio.jp

regression methods formulate landmark detection as a regression problem that estimates a heatmap of the landmark locations. Although they have achieved high accuracy, having the ability to exploit local features of the images, they have difficulty incorporating structural knowledge among the different landmarks. In those studies, several challenges remain, such as vulnerability to image distortions like occlusion and the difficulty of detecting certain cephalometric landmarks[7,8,13,19]. Therefore, a few studies[9,10,19] have proposed training methods that consider the local and anatomical features simultaneously. However, a significant enhancement in accuracy has yet to be achieved, and efforts to solve this problem have recently commenced. Furthermore, there have been recent attempts to investigate the potential of cephalometric landmark detection by applying originally created datasets to deep learning models[11,20].

As these previous studies were based on the assumption that cephalograms would be used, X-ray exposure is inevitable. As the exposure dose of cephalography is $2-3$ µSv, cone beam computed tomography (CBCT) is $20-850$ µSv for the maxillofacial procedures[22]. Therefore, these doses would be below the limits. Nevertheless, the risk of X-ray exposure to patients, especially children and pregnant women, in dental practice still exists. There is a demand for research on alternative techniques requiring no or low exposure to X-rays[22].

Facial landmark detection localizes predefined facial landmarks such as the eyes, nose, mouth, and chin from facial images. Previous studies have adopted the active shape model (ASM)[23] or constrained local model (CLM)[24] for detecting landmarks under certain restrictions. However, the robustness of these studies needs to be enhanced against various changes in appearance. A cascade regression approach[25] was studied to address this problem. However, a cascade regression is limited in deepening the structure for increased accuracy[26]. Deep neural networks (DNN) have recently been adopted as a powerful alternative[27]. In addition, CNN-based approaches have exhibited remarkable results. In particular, models with an hourglass structure[28] and heatmap-based regression[29–31] achieve high accuracy. Sun et al.[31] proposed HRNetV2, which uses a high-resolution network (HRNet)[32] for learning high-resolution representations, with the hourglass structure[28] being the mainstream. The structure of HRNetV2 connected high- to low-resolution convolutions in parallel, and it was possible to maintain a high-resolution throughout. Consequently, the performance of HRNetV2 was equal to or greater than that of conventional methods, while reducing the number of parameters and the computational cost. There is also study[33,34] focused on learning algorithms that do not depend on model structure. In fact, ADNet[29] is based on LAB[33] and Awing[34] and has shown high performance on many datasets.

In this paper, we present a novel cephalometric landmark detection method that incorporates a highly accurate facial landmark detection model. The proposed method is trained without cephalograms. We adopt HRNetV2[31], which achieves a high accuracy in a wide range of visual tasks by generating a high-resolution representation with accurate spatial information. In addition, we combine the heatmap regression model with a coordinate regression model. This solves the problem using conventional heatmap regression models, which incurs difficulty in learning anatomical features between landmarks. The proposed method comprises two stages: cephalometric landmark localization using HRNetV2 and refinement of landmark positions using multilayer perceptron (MLP). MLP contributes to the estimation reflecting the spatial relationship between landmarks. The inputs of the model are not cephalograms, as in the past, but facial profile images. To the best of our knowledge, this is the first study on cephalometric landmark detection without the use of cephalograms. During the experiment, we used facial profile images provided by the Tokyo Dental College. Two clinical orthodontists (with 3 and 15 years of experience, respectively) plotted a total of 23 landmarks on each of 2000 images based on the cephalograms, and one clinical orthodontist (with 36 years of experience) reviewed the results. The proposed method achieves a mean radial error (MRE) significantly below the clinically acceptable error of 2.0 mm[6]. Specifically, we obtained an MRE of 0.61 mm for the test data and a mean detection rate 98.20% for the threshold. While a direct comparison is difficult, we have suggested the possibility of achieving performance equal to or potentially better than existing methods using cephalograms. The main contributions of the proposed method are presented as follows:

- We proposed a cephalometric landmark detection without the use of cephalograms, and demonstrated performance at a level that is considered clinically acceptable.
- We achieved a significant improvement in accuracy by combining the CNN-based method with a conventional MLP.

Materials and methods

The study protocol was approved by the Institutional Review Board (IRB) of our institutes (Tokyo Dental College, No. 1091, 2021-12-17). All methods were carried out in accordance with the Helsinki Declaration principles and relevant guidelines and regulations. Informed consent was obtained from all the patients, and written informed consent was obtained from three patients, including the facial images, allowing us to use their information/images in an online open-access publication.

Figure 1 presents an overview of the proposed method. The proposed method comprises two stages: HRNetV2[31] and MLP. Unlike with existing studies, we used the facial profile images as input instead of a cephalogram. First, all landmark locations in the input images are estimated through heatmap regression using HRNetV2. In this step, the model learns the relationship between the local features of the image and landmarks. As described later in *Ablation Study* section, the accuracy of the estimation is insufficient, however. Therefore, we introduced a coordinate regression by MLP, which significantly improved the accuracy of landmark estimation. MLP was adopted to estimate the spatial location of the landmarks, i.e., their relative location. This two-stage approach, coarse estimation using heatmap regression and a fine estimation using MLP, enables the accurate detection of all landmarks.

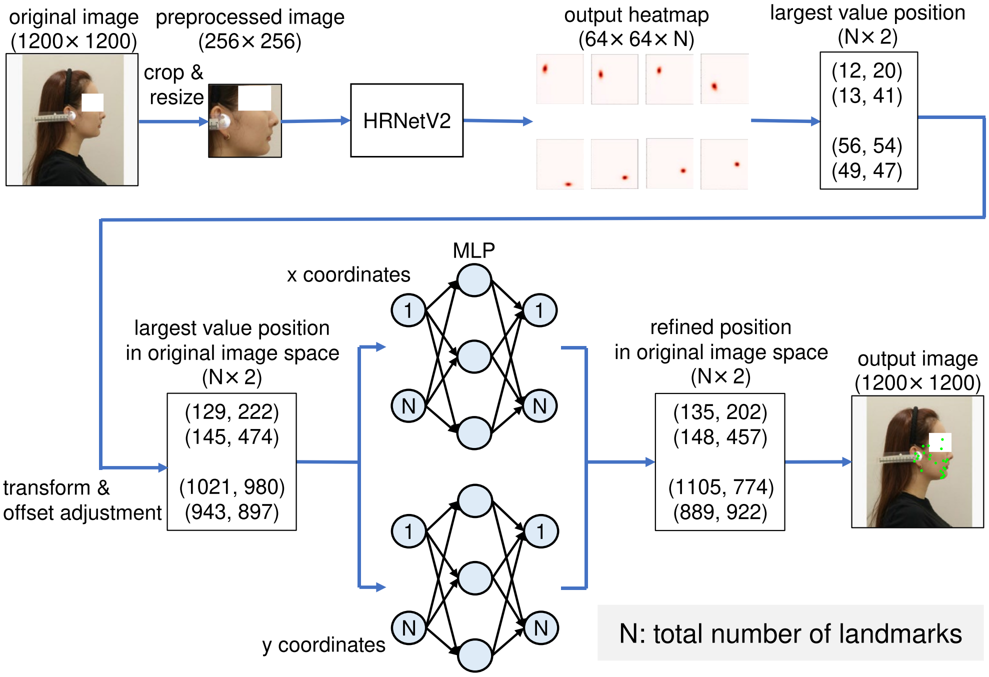

Figure 1. Overview of the proposed model. The model is divided into HRNetV2 and MLP. After HRNetV2 applies the heatmap regression based on the feature extraction, MLP estimates the coordinates based on the spatial relationship of the landmarks.

Dataset

A total of 2000 lateral cephalograms and profile photograph images collected from 2000 female patients aged 14 to 69 years (mean age of 27 years) were provided by the Tokyo Dental College. These images were acquired in jpeg format using a CX-150S (ASAHIROENTGEN, Tokyo, Japan) and an EOS Kiss X90 (Canon, Tokyo, Japan). The standard for taking lateral cephalogram is set worldwide. The film should be kept parallel to the mid-sagittal plane of the head and set the head with ear rods so that the center line of the X-ray beam passes through the axes of the left and right ear rods. The distances from the X-ray tube to the mid-sagittal plane of the head and from the mid-sagittal plane of the head to the film is assumed to be 150 cm and 15 cm, respectively. The method of taking a profile photograph image is the same as that of a lateral cephalogram. That is, the distance from the camera to the midsagittal plane of the head is 150 cm, and the head is fixed with ear rods. The skeletal information obtained from the lateral cephalograms was superimposed on the profile photograph images using Quick Ceph Studio (ver. 5.0.2, Quick Ceph Systems, San Diego, California). The 23 landmarks illustrated in Fig. 2 were manually plotted by two orthodontists and finally checked by an experienced orthodontist. The two of them are the board members of Orthodontic specialists. These 23 landmarks were selected from the measurement

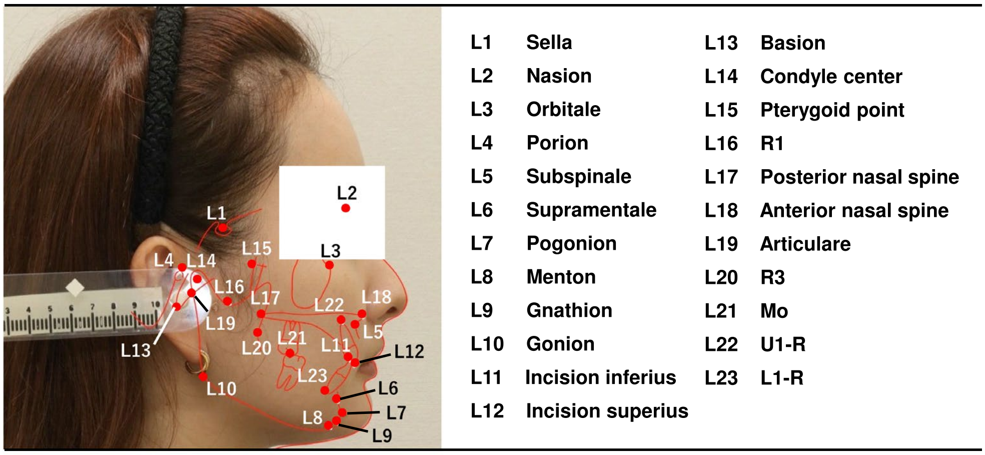

Figure 2. Illustration of 23 cephalometric landmarks.

points of the Downs method, Northwestern method, and a Ricketts analysis. Coordinates of landmarks on the lateral image were identified by imglab, an annotation tool included with Dlib[35]. Each image has 1200 × 1200 pixels and the pixel spacing for the device was 0.35. We present the averaged results of the experiment based on fivefold cross-validation, each containing 400 test and 1600 training profile photograph images, respectively.

Superimposition of the cephalometric tracing on the facial profile image

Figure 3 shows the procedure for plotting landmarks and superimposing the cephalometric tracing on the facial profile image conducted as follows:

1. Plot 23 landmarks and trace the profile (from forehead to upper lip) on the cephalogram within the computer screen of Orthodontic analysis software (Quick Ceph Studio).
2. Separate the set of the landmark plots and the cephalometric tracing on the screen.
3. Superimpose the cephalometric tracing on the facial profile image by manually matching the line from nose to upper lip of the tracing with that of the facial profile image on another screen.

Five sets of the landmark plots and the cephalometric tracing created in step 1 were randomly selected and the same set was superimposed on the facial profile image by each observer according to steps 2 and 3. We measured the intra- and inter-observer landmark distance errors for five landmarks (Sella, Porion, Menton, Gonion, Basion) to confirm the accuracy and reliability of the superimposition. The reason of selecting the five landmarks, it can be considered that the further away from the profile tracing line the larger the error because the superimposition is conducted based on the profile tracing line as shown in step 3. Therefore, the five points farthest from the profile tracing line were selected. To evaluate intra-observer variability, superimpositions were conducted by one orthodontist three times with an interval of two weeks. To evaluate inter-observer variability, superimpositions were conducted by three orthodontists. The two of them are the board member of Orthodontic specialist. The mean and standard deviation of the intra- and inter-observer landmark distance error and the intraclass correlation coefficient (ICC) for the landmark plots were calculated for each set with 95% confidence intervals. In calculating the ICC, the Shapiro-Wilk test was used to determine whether the data were normally distributed. All statistical evaluations were conducted by SPSS software (ver. 27.0, IBM, Armok, New York).

Repeatability and reproducibility test of manual landmark plotting

To compare the repeatability and reproducibility of orthodontists' landmark prediction errors, a total of five cephalograms from five patients were randomly selected and 23 landmarks were plotted. To evaluate intra-observer variability, cephalometric landmark plotting was conducted by one orthodontist three times with an interval of two weeks. To evaluate inter-observer variability, cephalometric landmark plotting was conducted by

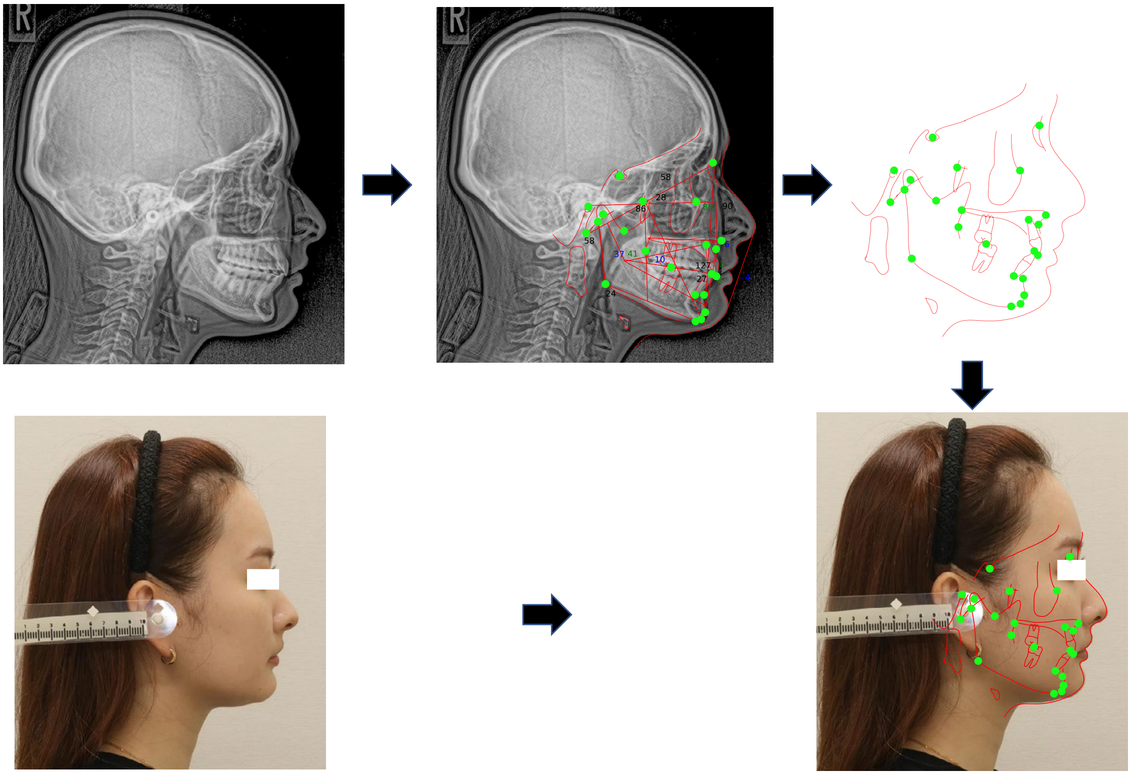

Figure 3. Procedure for plotting landmarks and superimposing the cephalometric tracing on the facial profile image.

three orthodontists. The mean and standard deviation of the intra- and inter-observer prediction error and the ICC for each landmark were calculated for each set with 95% confidence intervals. In calculating the ICC, the Shapiro-Wilk test was used to determine whether the data were normally distributed. All statistical evaluations were conducted by SPSS software.

Heatmap regression

To estimate the location of landmarks, we used HRNetV2 to generate heatmaps of the landmarks. Figure 4 shows the structure of HRNetV2[31], which starts with a high-resolution subnetwork (stem) and adds high- to low-resolution subnetworks successively to form a stage. The multi-resolution subnetworks are connected in parallel to form a total of four stages. The second, third, and fourth stages are set up by repeating the modularized multi-resolution blocks. The exchange unit, as illustrated in Fig. 5, aggregates the information of each resolution from other subnetworks[32]. Strided 3×3 convolutions with stride 2 and a simple nearest neighbor sampling are applied for downsampling and upsampling, respectively. The last exchange unit outputs a feature map where the low-resolution representation is upsampled and concatenated into a high-resolution representation. The heatmap is regressed from the high-resolution representation. A previous study[31] reported that the performance is enhanced by employing all resolution representations in comparison to solely using a high-resolution representation.

The ground truth heatmaps are defined as 2D gaussian functions with standard deviation of σ centered on the ground truth location of $\mathbf{L}_i^G$:

$$h_i^G(\mathbf{x};\sigma) = \frac{1}{2\pi\sigma^2}\exp\left(-\frac{\|\mathbf{x}-\mathbf{L}_i^G\|_2^2}{2\sigma^2}\right), \tag{1}$$

where, $\mathbf{x}$, $i = \{1, ...N\}$, and $N(= 23)$ represent the heatmap pixel, index of the landmarks, and total number of landmarks, respectively. To converge the estimated landmarks as close as possible to the ground truth landmarks, the loss function employs the L_2 loss as follows:

$$L_c = \sum_{k=1}^{N} \|h_i^G(\mathbf{x};\sigma) - h_i^P(\mathbf{x};\mathbf{w})\|_2^2, \tag{2}$$

where $h_i^P(\mathbf{x};\mathbf{w})$ represents the estimated heatmap, and $\mathbf{w}$ denotes the weights and biases of the network. The loss function allows the network to learn the relationship between the local features and landmarks. The heatmap comprises a 2D matrix with 23 channels corresponding to the number of landmarks. The estimated coordinates $\mathbf{L}_i^P$ of each landmark are predicted by transforming from the reduced space to the original image space. We adjusted the offset of the largest value position from the largest value to the second largest value[36]. Accordingly, the estimated position can be expressed as a set of 2D coordinates $\{\mathbf{x}_i^P, \mathbf{y}_i^P\}_{i=1}^N$.

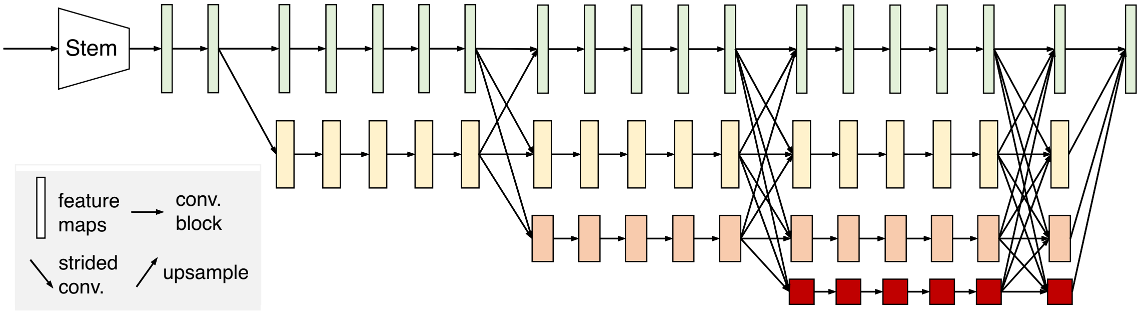

Figure 4. HRNetV2 structure[31]. HRNetV2 comprises four stages connected in parallel by high- to low-resolution subnetworks. The horizontal and vertical sizes of the feature maps correspond to the resolution and number of channels, respectively.

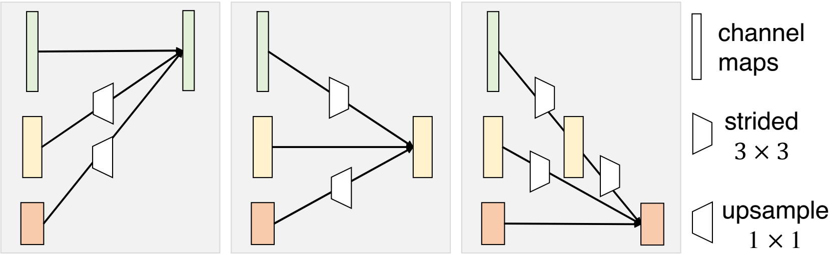

Figure 5. Illustration of how to aggregate information of multiple resolutions using an exchange unit[32]. A stride 3×3 convolution is adopted for downsampling. Nearest neighbor sampling followed by a 1×1 convolution is used for upsampling.

Coordinate regression

MLP is a refinement of coordinates estimated using spatial relationships between landmarks. It has a simple three-layer structure with an input layer, a hidden layer, and an output layer. Because this is a simple regression task, we adopted a simple MLP to reduce the possibility of an overfitting. In[2], a linear filter was also employed to refine the landmark positions; however, only some landmarks were adjusted, and the inputs were the combination of outputs from the two models when considering spatial information. The estimated coordinates $\{\mathbf{x}_i^P, \mathbf{y}_i^P\}_{i=1}^N$ are divided into $\{\mathbf{x}_1^P, ..., \mathbf{x}_N^P\}$ and $\{\mathbf{y}_1^P, ..., \mathbf{y}_N^P\}$. The *x*- and *y*-coordinates are input into separate models. We adopted the L_2 loss as the loss function to get closer to the ground truth landmarks by using the positional relationship of the landmarks:

$$L_n^x = \sum_{k=1}^{N} \|\mathbf{x}_i^G - \mathbf{x}_i^R\|_2^2, \tag{3}$$

$$L_n^x = \sum_{k=1}^{N} \|\mathbf{y}_i^G - \mathbf{y}_i^R\|_2^2, \tag{4}$$

where $\mathbf{x}_i^G$ and $\mathbf{y}_i^G$ represent the ground truth coordinates, and $\mathbf{x}_i^R$ and $\mathbf{y}_i^R$ denote the refined coordinates. Losses in the MLP were not propagated to HRNetV2, and were learned independently. The number of training epochs for HRNetV2 and MLP differed. For each epoch used for training HRNetV2, we trained the MLP with multiple epochs, provided that we initialized the parameters only at the beginning of training. This procedure allowed us to train the entire model efficiently, while fine-tuning the MLP to match the training phase of HRNetV2.

Implementation details

We trained and carried out testing using a GeForce GTX 1080, 3.70-GHz Intel(R) Core(TM) i7-8700K CPU, and 16GB of memory. The training and testing were conducted in Pytorch. The input images were cropped and resized to 256 × 256, according to the center positions of the boxes.

The HRNetV2 network starts with a stem comprising two strided 3 × 3 convolutions, which reduces the resolution to 1/4. As the inputs pass through the four subsequent stages, the resolution is gradually reduced by half, and the number of channels is accordingly doubled. The first stage contains four residual units, each of which was formed by a bottleneck with a width of 64. This stage is followed by a 3 × 3 convolution, which reduces the width of the feature map to 18. Thus, the number of channels for the four resolutions is 18, 36, 72, and 144, respectively. The second, third, and fourth stages contain one, four, and three multi-resolution blocks, respectively. One multi-resolution block contains four residual units. Each unit contains two 3 × 3 convolutions for each resolution and an exchange unit across resolutions. The four resolution representations from the fourth stage are concatenated and used to predict heatmaps with a width of 64, following two 1 × 1 convolutions. We trained HRNetV2 with 60 epochs and a batch size of 16. The model was pre-trained using WFLW[37]. The base learning rate was set as 0.0001, and then was reduced to 0.00001 and 0.000001 at 30 and 50 epochs, respectively. The loss function was minimized using the Adam optimizer.

The MLP network comprises three layers: input, hidden, and output layers. The number of neurons in the middle layer was set to 500. The number of inputs and outputs was set to 23 for splitting the coordinates transformed from the heatmaps and predicted by HRNetV2 into *x*- and *y*-coordinates. We trained the MLP using 100 epochs and a batch size of 16 every time HRNetV2 was trained for 1 epoch. The loss function was minimized using the Adam optimizer with a learning rate of 0.00001 and a weight decay (L2 regularization) factor of 0.0001.

Evaluation metrics

We evaluated the proposed method in terms of the mean radial error (MRE), successful detection rate (SDR), and successful classification rate (SCR), according to a previous benchmark study[13]. The radius error was defined by $R = \sqrt{\Delta x^2 + \Delta y^2}$, where Δx and Δy represent the Euclidean distances between the estimated landmarks and the ground truth landmarks of the *x*- and *y*-axes, respectively. The MRE and the standard deviation (SD) are defined as follows:

$$\text{MRE} = \frac{\sum_{i=1}^{N} R_i}{n}, \tag{5}$$

$$\text{SD} = \sqrt{\frac{\sum_{i=1}^{N} (R_i - MRE)^2}{n-1}}, \tag{6}$$

where R_i represents the radial error of the *i*th landmark, and *n* denotes the total number of landmarks to be detected. The SDR is the ratio of estimated landmarks that are within a reference threshold, and is defined as follows:

$$SDR = \frac{n_d}{n} \times 100\%, \tag{7}$$

where n_d represents the number of successfully detected landmarks, and the threshold values are 2.0, 2.5, 3.0, and 4.0 mm, as typically used. The SCR is the classification accuracy of anatomical face types based on eight clinical measures (ANB, SNB, SNA, overbite depth indicator (ODI), anteroposterior dysplasia indicator (APDI), facial height index (FHI), frankfurt mandibular angle (FMA), modified wits (MW)). Facial images are classified into three anatomical types under clinical measures. Note that geometric criteria such as the angles and distances between landmarks listed in Table 1 are considered.

Results

Accuracy and reliability of the superimposition of cephalometric tracing on the facial profile image

All distance errors were considered to follow a normal distribution by the Shapiro-Wilk test (p>0.05; not shown for details). The mean landmark distance error, standard deviation and the ICC values for each set are shown in Table 2. The mean intra- and inter-observer landmark distance error and standard deviations were 0.32 mm ± 0.07 mm and 0.41 mm ± 0.11 mm, respectively, indicating superiority over previous studies[18]. The ICC of the mean landmark distance error for each patient ranged from 0.993 to 0.998 (< 0.00, poor; 0.00–0.20, slight; 0.21–0.40, fair; 0.41–0.60, moderate; 0.61–0.80, substantial; 0.81–1.00, almost perfect)[38], indicating a high degree of intra- and inter-observer agreement. The 95% confidence interval also supports intra- and inter-observer agreement. In this way, the accuracy and reliability of the superposition was confirmed.

Result of repeatability and reproducibility test of manual landmark plotting

All distance errors were considered to follow a normal distribution by the Shapiro-Wilk test (p>0.05; not shown for details). The mean prediction error, standard deviation and the ICC values for each landmark are shown in Table 3. The mean intra- and inter-observer prediction error and standard deviations were 0.38 mm ± 0.11 mm and 0.37 mm ± 0.11 mm, respectively, indicating superiority over previous studies[18]. The ICC of the prediction error for each landmark ranged from 0.992 to 0.999 (< 0.00, poor; 0.00–0.20, slight; 0.21–0.40, fair; 0.41–0.60, moderate; 0.61–0.80, substantial; 0.81–1.00, almost perfect)[38], indicating a high degree of intra- and inter-observer agreement. The 95% confidence interval also supports intra- and inter-observer agreement. In this way, intra- and inter-observer agreement was confirmed.

Performance of cephalometric landmark detection for the proposed model and comparison with existing methods

Table 4 presents a comparison between the results of the proposed and existing methods. The proposed method achieves the best performance under all metrics. Because no existing studies have solely adopted facial profile images for training, we provide comparisons with a study adopting X-rays. The comparison methods were selected based on their recency and proximity to the proposed method. The performance of the proposed

Method	type1	type2	type3
ANB	Class 1 (normal): 3.2° – 5.7°	Class 2: > 5.7°	Class 3: < 3.2°
SNB	Normal mandible: 74.6° – 78.7°	Retrognathic mandible: > 74.6°	Prognathic mandible: < 78.7°
SNA	Normal mandible: 79.4° – 83.2°	Prognathic maxilla: > 83.2°	Retrognathic maxilla: < 79.4°
ODI	Normal: 78.4° – 80.5°	Deep bite tendency: > 80.5°	Open tendency: < 68.4°
APDI	Normal: 77.6° – 85.2°	Class 2 tendency: < 77.6°	Class 3 tendency: > 85.2°
FHI	Normal: 0.65 – 0.75	Short face tendency: > 0.75	Long face tendency: < 0.65
FMA	Normal: 26.8° – 31.4°	Mandible high angle tendency: > 31.4°	Mandible lower angle tendency: < 26.8°
MW	Normal: 2 – 4.5mm	Edge to edge: 0 mm	Large over jet: > 4.5 mm

Table 1. Criteria for eight clinical measures of anatomical face-type classification in SCR. ANB is the angle between L5, L2, and L6. SNB is the angle between L1, L2, and L6. SNA is the angle between L1, L2, and L5. ODI is the arithmetic sum of the angle between the lines from L5 to L6 and from L8 to L10, and the angle between the lines from L3 to L4 and from L17 to L18. APDI is the arithmetic sum of the angle between the lines from L3 to L4 and from L2 to L7, the angle between the lines from L2 to L7 and from L5 to L6, and the angle between the lines from L3 to L4 and from L17 to L18. FHI is the ratio of the distance from L1 to L10 (PFH) to the distance from L2 to L8 (AFH). FMA is the angle between the line from L1 to L2 and the line from L10 to L9. MW is the distance between L12 and L11.

Set	The intra-observer landmark distance error (mm)	The inter-observer landmark distance error (mm)	ICC(1, 3) with 95% confidence intervals	ICC(2, 3) with 95% confidence intervals
1	0.37 ± 0.06	0.27 ± 0.04	0.996 (0.984–1.000)	0.993 (0.968–0.999)
2	0.32 ± 0.06	0.41 ± 0.10	0.996 (0.983–1.000)	0.998 (0.992–1.000)
3	0.32 ± 0.07	0.41 ± 0.09	0.996 (0.982–1.000)	0.998 (0.987–1.000)
4	0.32 ± 0.07	0.46 ± 0.10	0.996 (0.983–1.000)	0.998 (0.991–1.000)
5	0.29 ± 0.07	0.49 ± 0.07	0.997 (0.973–0.999)	0.996 (0.980–0.999)

Table 2. Intra- and inter-observer mean landmark distance error, standard deviation and the ICC values with 95% confidence intervals for five sets.

Landmark	The intra-observer prediction error (mm)	The inter-observer prediction error (mm)	ICC(1, 3) with 95% confidence intervals	ICC(2, 3) with 95% confidence intervals
L1	0.32 ± 0.06	0.39 ± 0.08	0.994 (0.971–0.999)	0.993 (0.970–0.999)
L2	0.34 ± 0.06	0.40 ± 0.07	0.996 (0.981–1.000)	0.995 (0.979–0.999)
L3	0.36 ± 0.08	0.40 ± 0.11	0.997 (0.985–1.000)	0.993 (0.968–0.999)
L4	0.38 ± 0.06	0.39 ± 0.09	0.993 (0.968–0.999)	0.994 (0.975–0.999)
L5	0.36 ± 0.10	0.43 ± 0.10	0.996 (0.982–1.000)	0.992 (0.964–0.999)
L6	0.35 ± 0.10	0.46 ± 0.09	0.998 (0.991–1.000)	0.993 (0.966–0.999)
L7	0.41 ± 0.09	0.48 ± 0.05	0.997 (0.987–1.000)	0.994 (0.975–0.999)
L8	0.44 ± 0.07	0.44 ± 0.10	0.994 (0.972–0.999)	0.993 (0.965–0.999)
L9	0.42 ± 0.08	0.45 ± 0.06	0.995 (0.979–0.999)	0.996 (0.979–1.000)
L10	0.39 ± 0.10	0.33 ± 0.07	0.993 (0.971–0.999)	0.993 (0.969–0.999)
L11	0.25 ± 0.05	0.22 ± 0.05	0.992 (0.965–0.999)	0.994 (0.967–0.999)
L12	0.25 ± 0.06	0.24 ± 0.04	0.994 (0.971–0.999)	0.993 (0.969–0.999)
L13	0.46 ± 0.08	0.30 ± 0.04	0.993 (0.967–0.999)	0.993 (0.967–0.999)
L14	0.52 ± 0.06	0.35 ± 0.06	0.993 (0.967–0.999)	0.998 (0.992–1.000)
L15	0.51 ± 0.06	0.33 ± 0.08	0.995 (0.975–0.999)	0.998 (0.981–1.000)
L16	0.55 ± 0.06	0.39 ± 0.10	0.994 (0.974–0.999)	0.997 (0.984–1.000)
L17	0.34 ± 0.05	0.29 ± 0.07	0.994 (0.972–0.999)	0.996 (0.981–1.000)
L18	0.37 ± 0.07	0.29 ± 0.07	0.993 (0.968–0.999)	0.997 (0.986–1.000)
L19	0.28 ± 0.06	0.33 ± 0.09	0.993 (0.968–0.999)	0.998 (0.991–1.000)
L20	0.46 ± 0.06	0.43 ± 0.11	0.994 (0.973–0.999)	0.999 (0.994–1.000)
L21	0.47 ± 0.06	0.46 ± 0.08	0.995 (0.977–0.999)	0.998 (0.989–1.000)
L22	0.23 ± 0.05	0.46 ± 0.07	0.992 (0.966–0.999)	0.997 (0.985–1.000)
L23	0.22 ± 0.06	0.35 ± 0.08	0.996 (0.981–1.000)	0.996 (0.990–1.000)

Table 3. Intra- and inter-observer mean prediction error, standard deviation and the ICC values with 95% confidence intervals for 23 landmarks.

Method	MRE (mm)	SDR (%)			
		2.0 mm	2.5 mm	3.0 mm	4.0 mm
Ibragimov et al.[14]	1.96	68.13	74.63	79.77	86.87
Lindner et al.[15]	1.77	71.65	76.93	82.17	89.85
Arik et al.[18]	-	72.30	78.21	82.24	86.81
Gilmour et al.[9]	1.14	83.81	89.14	93.22	97.13
Li et al.[10]	1.20	83.72	89.34	92.72	96.78
Kwon et al.[2]	1.24	83.01	88.78	92.21	96.59
Oh et al.[19]	1.29	82.08	88.06	92.34	96.92
Proposed	**0.85**	**96.35**	**98.80**	**99.64**	**99.99**

Table 4. Comparison of MRE and SDR for the automated cephalometric analysis systems. All methods were trained by the ISBI2015 dataset[13]. The averages using test1 and test2 are reported. Significant values are in bold.

model trained on facial profile images is presented in Table 7. Table 5 presents the MRE, SD, and SDR for each of the 23 landmarks. The MRE for all landmarks is less than 2 mm, which is clinically acceptable[6]. Among the landmarks, Basion exhibited the best MRE. However, the MRE of Sella and Mo are large, and the SDR at 2 mm is low. The results of the proposed method indicates that Orbitale, Subspinale, Pogonion, Gonion, and Articulare, which have been known to handle landmarks that are difficult to accurately estimate in previous studies[7,8,13,19], can be estimated within 1 mm. In addition, it can be seen that the error in the y-coordinate is larger than that in the x-coordinate. Table 6 presents the SCR of the proposed method and the existing approachs. In terms of classification, the proposed method outperforms the existing methods under all metrics. The classification accuracy exceeds 90% in six out of eight metrics. Fig. 6 shows the MRE and loss of training and test according to the number of cycles. We defined one cycle as the process of training one epoch of HRNetV2 followed by 100 epochs of MLP.

Ablation study

We evaluated the contribution of MLP to the proposed method. We compared the performance of HRNetV2[31] with the performance of HRNetV2 combined with MLP. Table 7 presents the MRE, SD, and SDR for each model. Accordingly, it was deduced that the application of MLP significantly improves the performance. This

Landmark	MRE (mm)	SD (mm)	*x*-direction		*y*-direction		SDR (%)			
			Δx (mm)	SD (mm)	Δy (mm)	SD (mm)	2.0 mm	2.5 mm	3.0 mm	4.0 mm
L1	1.13	1.39	0.44	0.31	0.91	1.37	89.95	95.35	97.50	99.25
L2	0.67	0.33	0.28	0.19	0.56	0.35	99.70	99.90	99.95	100
L3	0.53	0.30	0.26	0.19	0.40	0.30	99.80	99.90	100	100
L4	0.39	0.16	0.30	0.16	0.20	0.13	100	100	100	100
L5	0.63	0.36	0.43	0.35	0.37	0.29	99.40	99.80	99.95	99.95
L6	0.64	0.35	0.32	0.25	0.48	0.34	99.50	99.95	99.95	99.95
L7	0.72	0.41	0.38	0.29	0.53	0.42	99.00	99.70	99.95	100
L8	0.61	0.37	0.36	0.27	0.41	0.35	99.45	99.85	99.90	99.95
L9	0.70	0.40	0.41	0.31	0.48	0.31	98.90	99.80	99.90	100
L10	0.63	0.69	0.19	0.14	0.56	0.71	98.85	99.65	99.80	99.90
L11	0.66	0.37	0.43	0.34	0.41	0.31	99.45	99.70	99.90	99.95
L12	0.59	0.36	0.37	0.31	0.38	0.31	99.80	99.95	99.95	99.95
L13	0.30	0.15	0.17	0.13	0.21	0.14	100	100	100	100
L14	0.68	0.54	0.45	0.35	0.41	0.51	99.50	99.85	99.90	99.90
L15	0.42	0.14	0.23	0.14	0.31	0.14	100	100	100	100
L16	0.46	0.28	0.36	0.28	0.22	0.18	99.90	99.90	99.90	99.95
L17	0.47	0.20	0.33	0.18	0.27	0.18	100	100	100	100
L18	0.71	0.39	0.52	0.39	0.38	0.30	98.60	99.75	99.90	99.95
L19	0.51	0.29	0.37	0.29	0.28	0.20	99.90	99.95	99.95	100
L20	0.60	0.29	0.39	0.26	0.37	0.26	99.85	100	100	100
L21	0.95	0.62	0.45	0.36	0.75	0.63	92.45	96.95	99.00	99.85
L22	0.45	0.25	0.28	0.22	0.30	0.23	99.80	99.95	100	100
L23	0.50	0.29	0.37	0.25	0.27	0.24	99.35	99.75	99.80	99.95
Average	0.61	0.53	0.35	0.30	0.41	0.51	98.83	99.55	99.79	99.93

Table 5. MRE, SD, and SDR for each landmark.

Method	ANB	SNB	SNA	ODI	APDI	FHI	FMA	MW
Ibragimov et al.[14]	66.31	72.75	63.50	72.31	80.07	70.08	77.78	81.38
Lindner et al.[15]	69.33	83.48	72.26	79.29	84.18	77.11	77.59	82.16
Arik et al.[18]	67.81	69.99	64.83	73.94	84.30	67.22	75.54	79.77
Kwon et al.[2]	81.78	84.14	72.91	85.26	87.47	86.40	85.75	87.50
Oh et al.[19]	80.90	85.23	68.98	79.01	86.15	82.48	80.98	87.73
Proposed	**90.45**	**87.30**	**83.50**	**93.75**	**91.50**	**93.55**	**93.55**	**98.65**

Table 6. Comparison of SCR for the automated cephalometric analysis systems. Numbers are given in percentages (%). Significant values are in bold.

indicates that heatmap regression, followed by coordinate regression using MLP, works effectively. The MRE for each landmark is presented in Fig. 7. The accuracies for Sella, Porion, Gonion, Basion, and Articulare, where HRNetV2 exhibits a poor accuracy, were also significantly improved. Table 8 presents the SCR for each. From this Table, it is evident that the accuracy is significantly improved, except for MW, which is consistently high. Figure 8 presents a visualization of the estimations by HRNetV2 and HRNetV2 + MLP, including the ground truth locations of the landmarks.

Discussion

The proposed method is novel because it estimates the location of cephalometric landmarks without lateral cephalograms. Table 7 shows that the MRE of the proposed method is 0.61 mm. While a direct comparison is difficult due to differences in datasets, the proposed method demonstrates higher performance compared to the previously reported error[18] between expert clinicians and existing methods. The experimental results suggest that the proposed method may perform at a level equal to or potentially better than existing methods using cephalograms. Although a manual analysis is limited by inter- and intra-person errors, the proposed model achieves an SD of 1 mm for all landmarks except for Sella. This implies that the proposed model can provide a stable estimation with small deviations. Furthermore, the proposed method achieves a significant SDR and is free

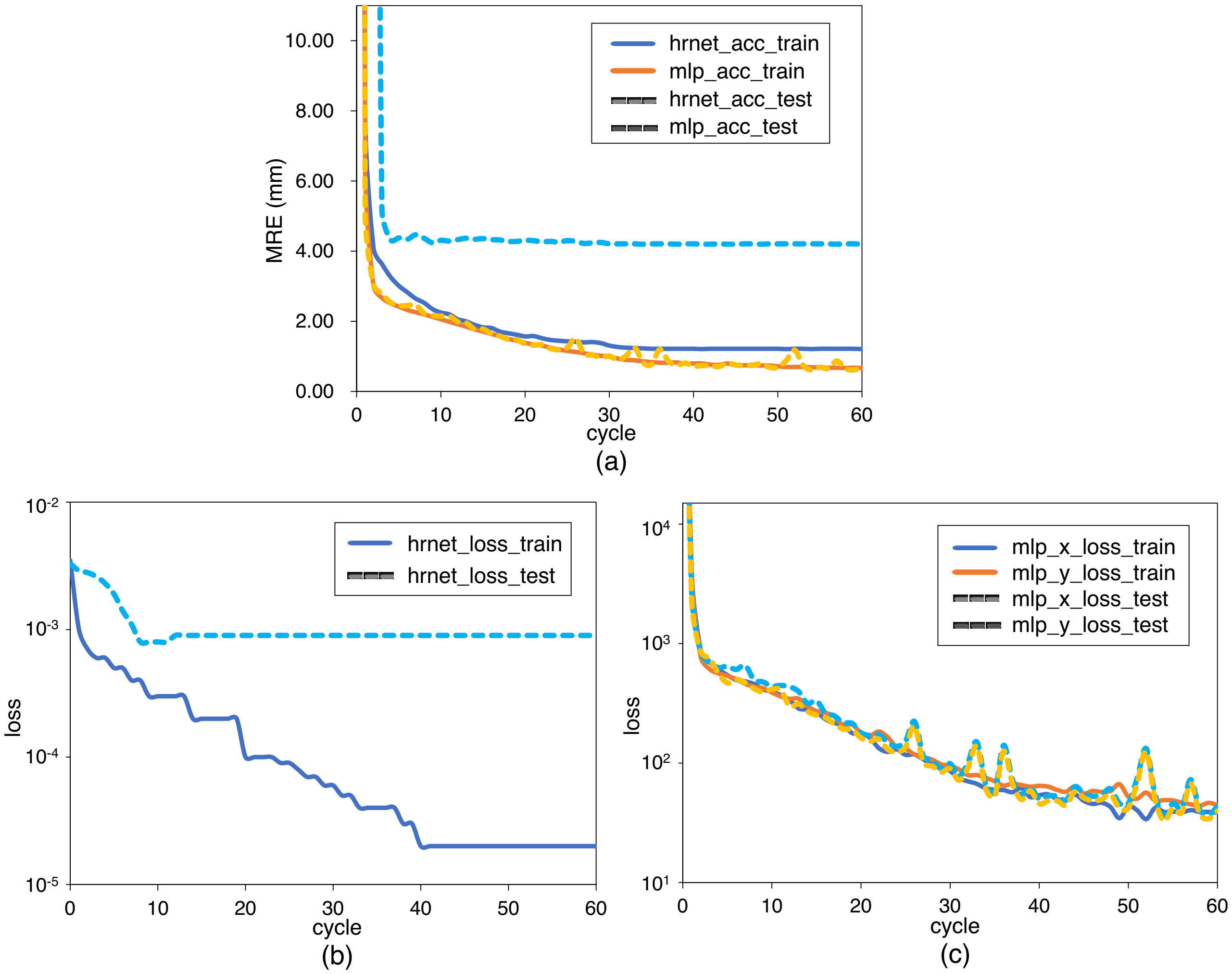

Figure 6. Performances of deep convolutional neural network-based AI model. (**a**) MRE of training and test according to the number of cycles. (**b**) HRNetV2 Loss of training and test according to the number of cycles. (**c**) MLP Loss of training and test according to the number of cycles. Cycle means the process of training one epoch of HRNetV2 followed by 100 epochs of MLP.

Method	MRE (mm)	SD (mm)	SDR (%)			
			2.0 mm	2.5 mm	3.0 mm	4.0 mm
HRNetV2[31]	4.11	3.21	22.42	32.23	40.74	58.20
HRNetV2[31]+MLP	0.61	0.53	98.20	99.55	99.79	99.93

Table 7. Performance of the proposed method and comparison of MRE, SD, and SDR, with and without MLP.

from the effects of extreme outliers[18,19] which have limited several previous studies. Therefore, a cephalometric analysis using AI may replace a human analysis in the future.

The accuracy is also high for landmarks that have been considered difficult to estimate accurately in existing studies. Figure 7 indicates that the estimations of Gonion and Articulare by HRNetV2 are still inaccurate. This may be due to the influence of the ear rod used to fix the head position, which is commonly included in both lateral cephalograms and facial profile images. This supports the idea that CNN-based heatmap regression methods learn each landmark independently. In particular, HRNetV2 strictly learns the positional relationship of each landmark with the surrounding features owing to the parallel distributed processing with many convolutional layers applied. Following the estimation with HRNetV2, the refinement of the landmark locations using MLP contributes to an improvement of the estimation accuracy of all landmarks. It is difficult to learn intricate positional relationships using only MLP. Note that the estimation in HRnetV2 described in the previous section may include the potential structural relationships among landmarks. Because MLP is fully connected from the input layer to the output layer and the input location information is processed comprehensively, it is possible to explicitly learn this potential positional relationship. This should allow MLP to incorporate the structural

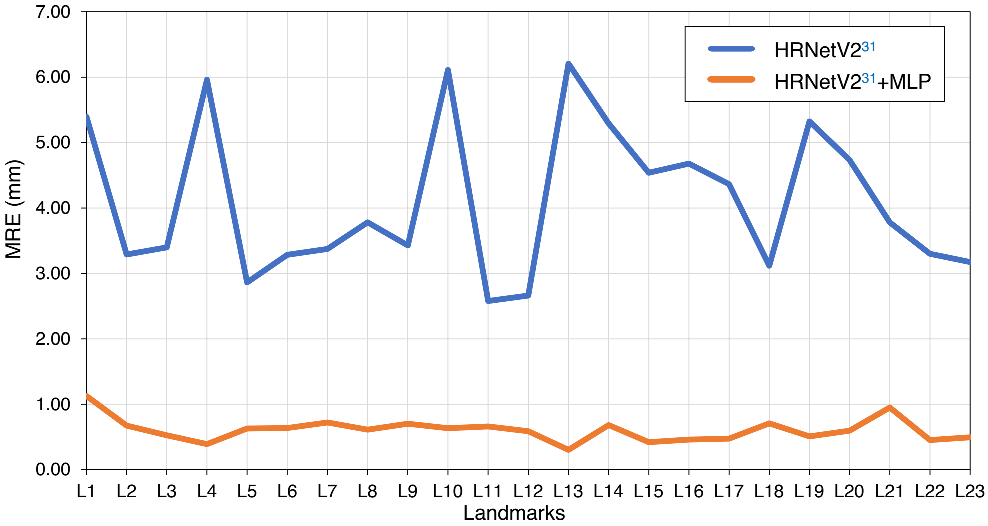

Figure 7. MRE for each landmark, with and without MLP.

Method	ANB	SNB	SNA	ODI	APDI	FHI	FMA	MW
HRNetV2[31]	66.30	56.75	44.60	65.70	64.00	76.70	79.35	97.65
HRNetV2[31]+MLP	90.45	87.30	83.50	93.75	91.50	93.55	93.55	98.65

Table 8. SCR comparison, with and without MLP.

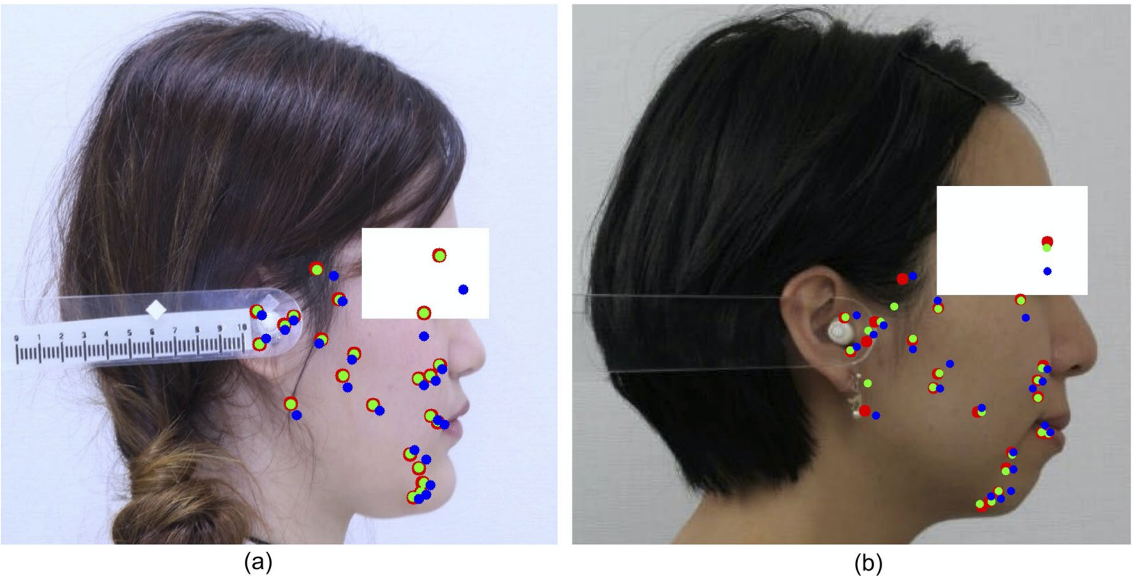

Figure 8. Visualization of landmarks, with and without MLP. (**a,b**) The best and worst result of the model with MLP, respectively. Blue and green dots show the estimations by HRNetV2 and HRNetV2+MLP, respectively, and the red dots indicate the ground truth.

features among landmarks into the estimation, which should lead to a significant improvement in accuracy. The effectiveness of incorporating structural relationships among landmarks in an estimation was also shown by Oh et al.[19]. Although Fig. 6a, b show that HRNetV2, which provides the intermediate output, tends to overfit, Fig. 6a, c show that the loss of training and test data is very similar in the MLP that provides the final output, indicating almost no over-fitting issue. This suggests that MLP is effective in reducing the effects of over-fitting by HRNetV2. Increasing the number of data may reduce over-fitting in HRNetV2 and further improve overall accuracy. In addition, enhancing the heatmap regression procedure can further enhance the overall performance. We can introduce adaptive wing loss[34] or face boundary prediction[33], which achieves a high accuracy in facial landmark detection.

As demonstrated in Table 5, Sella, and Mo have a larger MRE and lower SDR at 2 mm than the other landmarks. In particular, Sella has a larger error in the y-direction. Two reasons can be given for the estimation difficulty: First, Sella is the highest position among the landmarks to be estimated in most cases, and its location is far from the jaw area where the landmarks are densely located. Second, Mo has large errors in both the x- and y-directions. It appears that the variation of the position among patients biased the model. The error tends to be larger in the y-direction than in the x-direction. The shape of the face may influence the estimation. Most of the patients in this study have long faces, and landmarks are scattered in the y-direction. This may explain why the estimation error in the y-direction is larger than that in the x-direction.

Table 6 implies that the proposed method achieves a significant SCR. Because the SCR is calculated using linear and angular measurements, it is not influenced by the pixel spacing. Therefore, we can make reliable comparisons, even with different datasets. Note that the dataset used in this study did not include patients with a malocclusion. However, a clinical evaluation requires the inclusion of such data.

The performance of the proposed method tends to depend on the size and diversity of the training data. Previous studies[39] have reported that the accuracy of the proposed method increases with an increase in the number of data. Hence, adding more data is the easiest way to improve the system. However, this study also includes the risk of bias regarding the validation discussed in Schwendicke et al.[40]. Because the MLP approach is model-agnostic, there is no restriction on the heatmap regression model used for the initial estimation. The proposed method can reconcile the local features of images with the spatial relationship between landmarks. Hence, it may be widely used for landmark detection tasks such as facial landmark detection and pose estimation. In the future, it will be necessary to address 3D landmark detection.

Conclusion

In this paper, we proposed a novel cephalometric landmark detection method without the use of X-rays. The proposed framework combines the CNN-based heatmap regression model (HRNetV2) with a coordinate regression model (MLP). HRNetV2 estimates the location of landmarks by learning the relationship between local features and landmarks. However, it is limited by the same problem as conventional heatmap regression methods, and its accuracy is insufficient. The MLP following HRNetV2 can learn the spatial positional relationship between landmarks, which significantly improves the accuracy of the estimation. In experiments conducted with the created dataset, the proposed method performed remarkably with an MRE of 0.61 mm and a detection rate within 2.0 mm of 98.20%. The proposed method also achieved a high accuracy in terms of anatomical face-type classification, where a reliable comparison between different datasets is possible. In the future, in addition to replacing existing methods using X-rays, the proposed method may also replace measurements by humans. Because the proposed landmark refinement using MLP is model-agnostic, it can be combined with conventional methods. Furthermore, it has the potential to be a prominent approach applicable to various landmark detection tasks and provide significant improvements in performance.

Data availability

The datasets generated and analyzed during the study are not publicly available because sensitive information in them may violate patient privacy and our institution's ethics policy. However, the datasets are available from the corresponding author on reasonable request. Data usage agreements may be required.

Received: 29 September 2022; Accepted: 7 November 2023
Published online: 16 November 2023

References

1. Broadbent, B. H. A new X-ray technique and its application to orthodontia. *Angle Orthod.* **1**, 45–66 (1931).
2. Kwon, H. J., Koo, H. I., Park, J. & Cho, N. I. Multistage probabilistic approach for the localization of cephalometric landmarks. *IEEE Access* **9**, 21306–21314 (2021).
3. Grau, V., Alcaniz, M., Juan, M., Monserrat, C. & Knoll, C. Automatic localization of cephalometric landmarks. *J. Biomed. Inf.* **34**, 146–156 (2001).
4. Yue, W., Yin, D., Li, C., Wang, G. & Xu, T. Automated 2-D cephalometric analysis on X-ray images by a model-based approach. *IEEE Trans. Biomed. Eng.* **53**, 1615–1623 (2006).
5. Levy-Mandel, A., Venetsanopoulos, A. & Tsotsos, J. Knowledge-based landmarking of cephalograms. *Comput. Biomed. Res.* **19**, 282–309 (1986).
6. Song, Y., Qiao, X., Iwamoto, Y., Chen, Y.-W. & Chen, Y. An efficient deep learning based coarse-to-fine cephalometric landmark detection method. *IEICE Trans. Inf. Syst.* **104**, 1359–1366 (2021).
7. Lee, J.-H., Yu, H.-J., Kim, M.-J., Kim, J.-W. & Choi, J. Automated cephalometric landmark detection with confidence regions using Bayesian convolutional neural networks. *BMC Oral Health* **20**, 1–10 (2020).
8. Zhong, Z., Li, J., Zhang, Z., Jiao, Z. & Gao, X. An attention-guided deep regression model for landmark detection in cephalograms. In *Proceedings of the International Conference Medicine Image Computing and Computer Assisted Intervention (MICCAI)*. 540–548 (Springer, 2019).
9. Gilmour, L. & Ray, N. Locating cephalometric x-ray landmarks with foveated pyramid attention. In *Proceeding of the International Conference on Medicine Imaging Deep Learning (MIDL)*. 262–276 (PMLR, 2020).
10. Li, W. *et al.* Structured landmark detection via topology-adapting deep graph learning. In *Proceedings of the European Conference on Computer Vision*. 266–283 (Springer, 2020).
11. Kim, H. *et al.* Web-based fully automated cephalometric analysis by deep learning. *Comput. Methods Programs Biomed.* **194**, 105513 (2020).
12. Wang, C.-W. *et al.* Evaluation and comparison of anatomical landmark detection methods for cephalometric X-ray images: A grand challenge. *IEEE Trans. Med. Imag.* **34**, 1890–1900 (2015).

13. Wang, C.-W. *et al.* A benchmark for comparison of dental radiography analysis algorithms. *Med. Image Anal.* **31**, 63–76 (2016).
14. Ibragimov, B., Likar, B., Pernus, F. & Vrtovec, T. Computerized cephalometry by game theory with shape-and appearance-based landmark refinement. In *Proceedings of the IEEE International Symposium on Biomedicine Imaging (ISBI)* (2015).
15. Lindner, C. & Cootes, T. F. Fully automatic cephalometric evaluation using random forest regression-voting. In *Proceeding of the IEEE International Symposium Biomedicine Imaging*. 1–5 (2015).
16. Lindner, C. *et al.* Fully automatic system for accurate localisation and analysis of cephalometric landmarks in lateral cephalograms. *Sci. Rep.* **6**, 1–10 (2016).
17. Lee, H., Park, M. & Kim, J. Cephalometric landmark detection in dental X-ray images using convolutional neural networks. In *Medical Imaging 2017: Computer-Aided Diagnosis*. Vol. 10134. 101341W (International Society for Optics and Photonics, 2017).
18. Arik, S. Ö., Ibragimov, B. & Xing, L. Fully automated quantitative cephalometry using convolutional neural networks. *J. Med. Imag.* **4**, 014501 (2017).
19. Oh, K. *et al.* Deep anatomical context feature learning for cephalometric landmark detection. *IEEE J. Biomed. Health. Inf.* **25**, 806–817 (2020).
20. Kim, M.-J. *et al.* Automatic cephalometric landmark identification system based on the multi-stage convolutional neural networks with cbct combination images. *Sensors* **21**, 505 (2021).
21. Domingos, P. A few useful things to know about machine learning. *Commun. ACM* **55**, 78–87 (2012).
22. Valentin, J. The 2007 recommendations of the international commission on radiological protection. *ICRP Publ.* **103**(37), 2–4 (2007).
23. Milborrow, S. & Nicolls, F. Locating facial features with an extended active shape model. In *Proceeding of the European Conference on Computer Vision*. 504–513 (Springer, 2008).
24. Cristinacce, D., Cootes, T. F. *et al.* Feature detection and tracking with constrained local models. In *Proceeding of the 17th British Machine Vision Conference*. Vol. 1. 3 (Citeseer, 2006).
25. Dollár, P., Welinder, P. & Perona, P. Cascaded pose regression. In *Proceedings of the IEEE Conference on Computer Vision Pattern Recognition*. 1078–1085 (IEEE, 2010).
26. Sun, X., Wei, Y., Liang, S., Tang, X. & Sun, J. Cascaded hand pose regression. In *Proceedings of the IEEE Conference on Computer Vision Pattern Recognition*. 824–832 (2015).
27. Khabarlak, L. & Koriashkina, K. Fast facial landmark detection and applications: A survey. *J. Comput. Sci. Technol.* **22**, e02 (2022).
28. Newell, A., Yang, K. & Deng, J. Stacked hourglass networks for human pose estimation. In *Proceedings of the European Conference on Computer Vision*. 483–499 (Springer, 2016).
29. Huang, Y., Yang, H., Li, C., Kim, J. & Wei, F. Adnet: Leveraging error-bias towards normal direction in face alignment. arXiv:2109.05721 (2021).
30. Bulat, A., Sanchez, E. & Tzimiropoulos, G. Subpixel heatmap regression for facial landmark localization. arXiv:2111.02360 (2021).
31. Sun, K. *et al.* High-resolution representations for labeling pixels and regions. arXiv:1904.04514 (2019).
32. Sun, K., Xiao, B., Liu, D. & Wang, J. Deep high-resolution representation learning for human pose estimation. In *Proceeding of the IEEE Conference on Computer Vision Pattern Recognition*. 5693–5703 (2019).
33. Wu, W. *et al.* Look at boundary: A boundary-aware face alignment algorithm. In *Proceedings of the IEEE Conference on Computer Vision Pattern Recognition*. 2129–2138 (2018).
34. Wang, X., Bo, L. & Fuxin, L. Adaptive wing loss for robust face alignment via heatmap regression. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 6971–6981 (2019).
35. King, D. E. Dlib-ml: A machine learning toolkit. *J. Mach. Learn. Res.* **10**, 1755–1758 (2009).
36. Chen, Y., Shen, C., Wei, X.-S., Liu, L. & Yang, J. Adversarial posenet: A structure-aware convolutional network for human pose estimation. In *Proceedings of the IEEE International Conference Computer Vision*. 1212–1221 (2017).
37. Yang, S., Luo, P., Loy, C.-C. & Tang, X. Wider face: A face detection benchmark. In *Proceedings of the IEEE Conference on Computer Vision Pattern Recognition*. 5525–5533 (2016).
38. Landis, J. R. & Koch, G. G. The measurement of observer agreement for categorical data. *Biometrics* **33**, 159–174 (1977).
39. Moon, J.-H. *et al.* How much deep learning is enough for automatic identification to be reliable? *Angle Orthod.* **90** (2020).
40. Schwendicke, F. *et al.* Deep learning for cephalometric landmark detection: Systematic review and meta-analysis. *Clin. Oral Investig.* **25**, 4299–4309 (2021).

Author contributions

K.T. contributed to study design, building and learning the deep learning models, experiments, and writing of the manuscript. Y.S. and C.T. contributed to study concept, data collection, and data labeling. Y.N. contributed to study concept, data collection, data labeling, and critical revision of the manuscript. M.H. contributed to study design and critical revision of the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to K.T.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2023