PAPER

An Efficient Deep Learning Based Coarse-to-Fine Cephalometric Landmark Detection Method

Yu SONG[†*], **Xu QIAO**[††*], ***Nonmembers***, **Yutaro IWAMOTO**[†], **Yen-Wei CHEN**[†a)], ***Members***, ***and*** **Yili CHEN**[†††b)], ***Nonmember***

SUMMARY Accurate and automatic quantitative cephalometry analysis is of great importance in orthodontics. The fundamental step for cephalometry analysis is to annotate anatomic-interested landmarks on X-ray images. Computer-aided automatic method remains to be an open topic nowadays. In this paper, we propose an efficient deep learning-based coarse-to-fine approach to realize accurate landmark detection. In the coarse detection step, we train a deep learning-based deformable transformation model by using training samples. We register test images to the reference image (one training image) using the trained model to predict coarse landmarks' locations on test images. Thus, regions of interest (ROIs) which include landmarks can be located. In the fine detection step, we utilize trained deep convolutional neural networks (CNNs), to detect landmarks in ROI patches. For each landmark, there is one corresponding neural network, which directly does regression to the landmark's coordinates. The fine step can be considered as a refinement or fine-tuning step based on the coarse detection step. We validated the proposed method on public dataset from 2015 International Symposium on Biomedical Imaging (ISBI) grand challenge. Compared with the state-of-the-art method, we not only achieved the comparable detection accuracy (the mean radial error is about 1.0–1.6mm), but also largely shortened the computation time (4 seconds per image).

key words: *cephalometric landmark, x-ray, deep learning, registration, deformable transformation*

1. Introduction

Cephalometry analysis is of great importance for doctors to make diagnosis and treatment plans [1]–[3]. It has a long history, which can date back to 1931. Usually, skeletal X-ray images are widely used for this analysis due to its high resolution. In order to do cephalometry analysis, anatomical cephalometric landmarks need to be annotated first. One typical example of 19 cephalometric-interested landmarks is shown as Fig. 1.

With the development of machine learning and deep learning techniques, research on automatic cephalometric landmark detection have been increased sharply. Especially in 2014 and 2015, International Symposium on Biomedical Imaging (ISBI) launched two grand challenges on cephalometry, aiming to recruit computer-aided methods to automatic detect cephalometric landmarks in high accuracy [4], [5]. Several classic methods have been proposed. In 2015's ISBI grand challenge, the best result was achieved by Lindner et.al [6]. By using a random-forest based method, they achieved a 74.84% SDR (Successful Detection Rate) for a 2mm precision range [6]. Ibragimov et.al. achieved the second-best results by using harr-like feature extraction with random-forest regression [7].

Deep learning has presented unprecedented performance in computer vision problems since the success of AlexNet in 2012 ImageNet Challenge [8]. Compared with conventional image processing methods, as well as other machine learning methods, they have achieved great improvements in problems like image classification [9], image segmentation [10], [11] and so on. Many state-of-the-art deep learning-based methods have also been proposed on the cephalometric landmark detection problems. In 2017, Arik et al. improved their previous work by replacing random-forest regression with convolutional neural network (CNN) to do binary classification, then refined it with shape model [12]. In 2017, Hansang Lee et al. proposed a deep learning method to directly output landmarks'

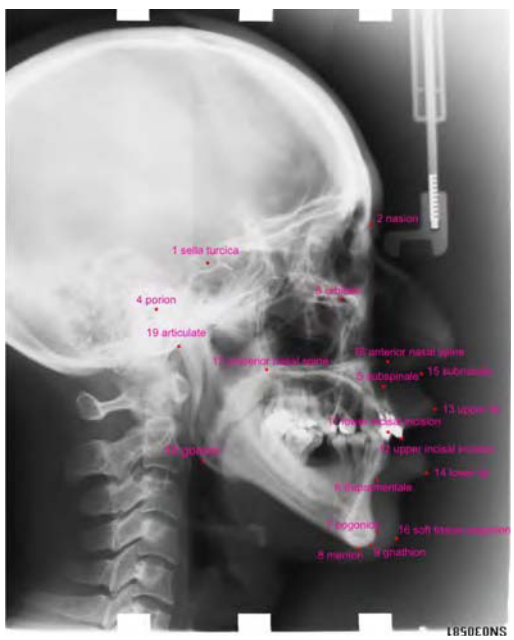


Fig. 1 Example of 19 cephalometric-interested landmarks.

Manuscript received January 4, 2021.
Manuscript revised March 20, 2021.
Manuscript publicized May 14, 2021.

[†]The authors are with the Ritsumeikan University, Kusatsu-shi, 525–0058 Japan.
[††]The author is with the Shandong University, China.
[†††]The author is with the Zhejiang University, China.
[*]Yu SONG and Xu QIAO contributed equally to this paper.
a) E-mail: chen@is.ritsumei.ac.jp (Corresponding author)
b) E-mail: 3194086@zju.edu.cn (Corresponding author)

DOI: 10.1587/transinf.2021EDP7001

coordinates [13]. They achieved comparable results on resized small images. In 2019, Jianhong Qian et al. proposed a network structure named Cephanet and achieved relatively high detection accuracy compared with other state-of-the art methods [14]. In our previous work, we proposed a two-step method to detect cephalometric landmarks with high detection accuracy [15]. In the coarse detection step, we used a rigid registration method to register the test image to the training image to detect the landmark roughly. Then we used deep learning models to detect landmarks precisely based on the extracted regions of interest (ROIs). Since the rigid registration used in the coarse detection step is not possible to achieve a good match if two images are quite different, we need to register the test image to all training images and find the best matched image. It should also be noted that the transform parameters for each test image registration are obtained based on an optimization algorithm such as gradient decent, which is an iterative method. So the coarse detection step in the previous method [15] takes very large computation time.

In this paper, we propose a coarse-to-fine method to detect cephalometric landmarks, therefore, reducing the large computation cost in the coarse detection step. We first train a deep learning-based deformable transformation model by using training samples for the coarse detection step. We choose a training image as the reference image and use other training images as moving images. In the test phase, we just need to input the test image (as a moving image) and the reference image to the deformable transformation model and we can obtain a displacement field as an output of the model to transform the test image to the reference image. We then inversely transform the reference image's landmarks to the test image, which can be considered as coarse estimation or coarse detection of the test image. Since we use a trained model to estimate the displacement field (parameters) for each test image, the coarse detection is very fast and efficient. In addition, the deformable transformation model is trained for non-rigid registration, we do not need to find the best matched training image for coarse detection. The fine step is the same as our previous method [15]. For each landmark, we train one model to detect the landmark in the ROI. In all, we have 19 models for 19 landmarks, all the models share the same architecture but with different weights. Based on the coarse landmark locations in the coarse detection step, we crop small patches (ROIs) and use the trained deep neural network models to detect landmarks' locations in those ROIs precisely (fine detection).

The following of this paper will be arranged as follows: In Sect. 2, we will introduce our proposed method in detail. In Sect. 3, we are going to present our experiments and comparisons. Finally, we will make a conclusion and discussion in Sect. 4.

2. Materials and Methods

2.1 Overview

Since it is difficult to accurately detect all cephalometric landmarks at once [15], we propose an efficient deep learning-based coarse-to-fine approach to realize accurate landmark detection in this paper. The overview of the proposed method is shown in Fig. 2.

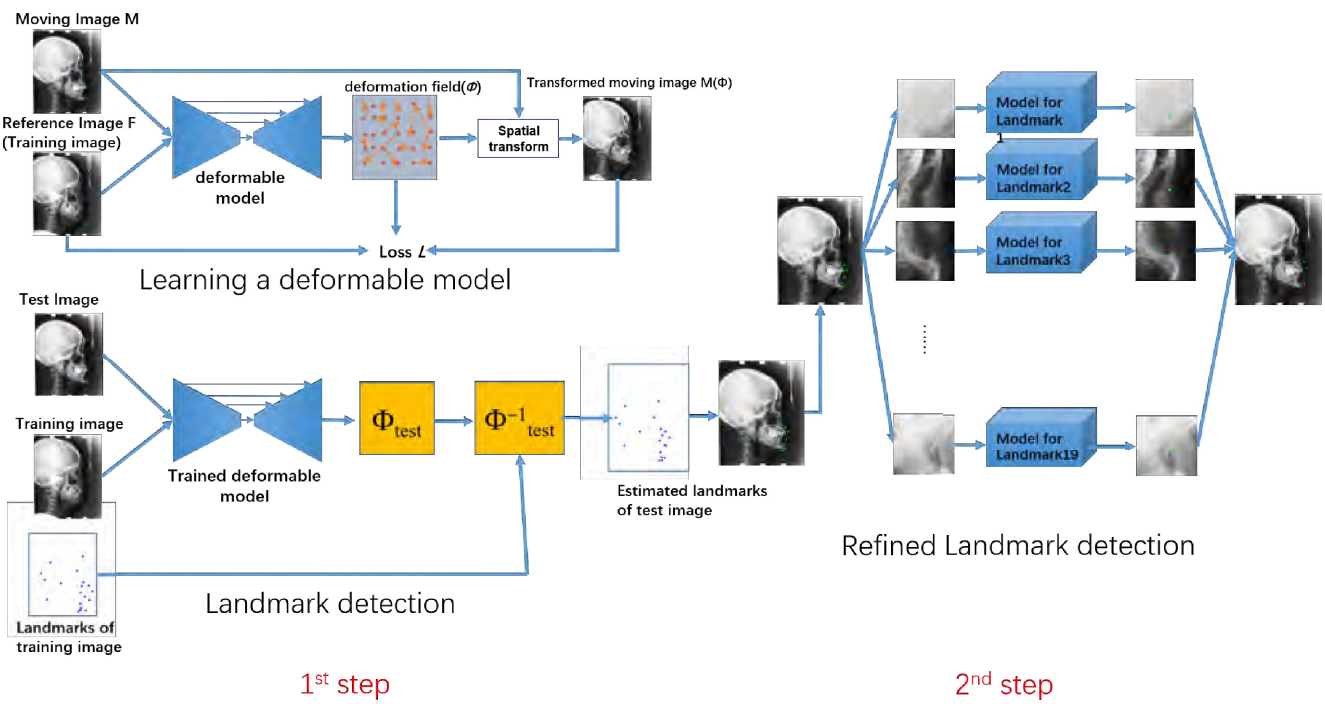


Fig. 2 Overview of the proposed coarse-to-fine method.

In the coarse detection step, we first train a deep learning-based deformable transformation model by using training samples. Then we register test image to reference image to predict the coarse landmarks' locations of the test image and extract a region of interest (ROI) patch for each landmark centered on its predicted position. In the fine detection step, we utilize trained CNNs with ResNet backbone, to detect every landmark in its corresponding patch. In other words, the fine step can be considered as making refinements based on the coarse detection step.

For the fine step, we use the same strategy with our previous method [15], aim to do refinements by locating the landmark in small patch images. In the training phase, we cut out small patches from training images, doing data augmentation, and training deep CNNs to detect landmarks in the small patches. We train one model for each landmark, which means that we have 19 models to detect 19 landmarks, where every model shares the same architecture but with different weights. In the test phase, since we already get the coarse landmark locations in the coarse detection step, we cut out a patch centered at that coarse location from test image, input into our corresponding trained models to detect landmarks. The result is our final prediction results, which can be considered as the refined results.

2.2 Coarse Landmark Detection

In the coarse detection step, we propose to use a deep-learning-based deformable transformation model, to register the test image to the reference image [16], [17].

The backbone of our architecture is 2D U-Net [11]with encoders and decoders, as shown in Fig. 3. We concatenate reference image and moving image into a two-channel image as input. After the encoder layers, the image's size is reduced to 1/16 of its original size. Then, the decoder layers upsample the small feature maps to the original size. The output of the decoders is the displacement field u between the reference image and the moving image. The displacement field has the size of w x h x 2, where w and h represent the input image's width and height respectively. For each pixel p, $u(p)$ is a displacement field to make $f(p)$ and $M(\Phi)(p)$ correspond to similar anatomical locations, which means a shift is added to every pixel p. After transforming the moving image M using the displacement field Φ, we obtained the transformed moving image $M(\Phi)$. Since we don't have any ground-truth displacement field, our aim is to make this transformed moving image be as similar as possible to the reference image F, so that we can consider the reference image's landmarks as transformed moving image's landmarks. In order to calculate the similarity between transformed moving image and reference image, we choose to calculate the intensity difference between them. We calculate the pixel-wise difference between these two images, back propagating through the network to make the difference become smaller during training process, until achieving convergence. One example of moving image, fixed image, their corrsponding displacement field and transformed moving image is shown in Fig. 4.

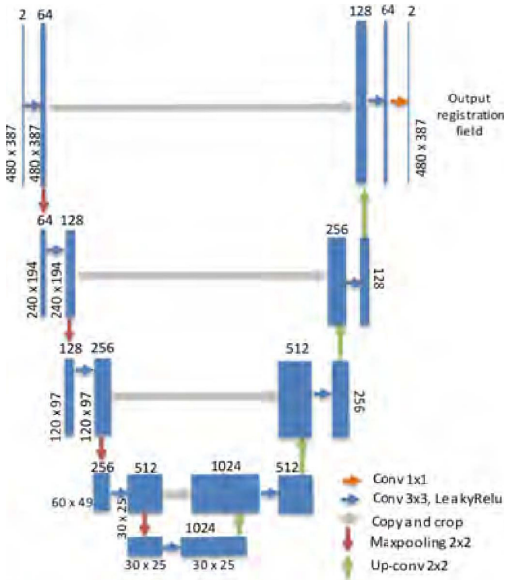
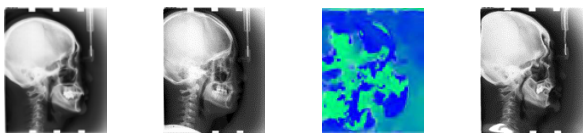


Fig. 3 Unet architecture used in the proposed method.



(a) Moving image M (b) Reference image F (c) Displacement Field Φ (d) Transformed Moving image $M(\Phi)$

Fig. 4 Example of Moving image M, Reference Image F, their displacement field Φ and Transformed moving image $M(\phi)$.

In the training phase, we choose a training image as the reference image and use other training images as moving images. In order to calculate the similarity between transformed moving image $M(\Phi)$ and reference image F, we choose mean squared error (MSE) between $M(\Phi)$ and F as the loss function. In addition, we also use a Laplacian of the displacement field Φ as a regularization term to penalize local spatial variations in Φ. The loss function L can be written as Eq. (1):

$$L = \frac{1}{n * m} \sum_{i=1}^{n} \sum_{j=1}^{m} (F_{i,j} - M(\Phi)_{i,j})^2 + \lambda \sum_{i=1}^{n} \sum_{j=1}^{m} \left\| \Delta\Phi(i,j) \right\| \tag{1}$$

where $M(\Phi)$ represents the transformed moving image, F represents the reference image, i and j are pixel coordinates, n and m represent width and height. In the test phase, we just need to input the test image (as a moving image) and the reference image to the trained deformable transformation model and we can obtain a displacement field Φ_{test} between the test image and the reference image. We then inversely transform the reference image's landmarks to the test image using (Φ_{test}^{-1}), which can be considered as coarse landmark

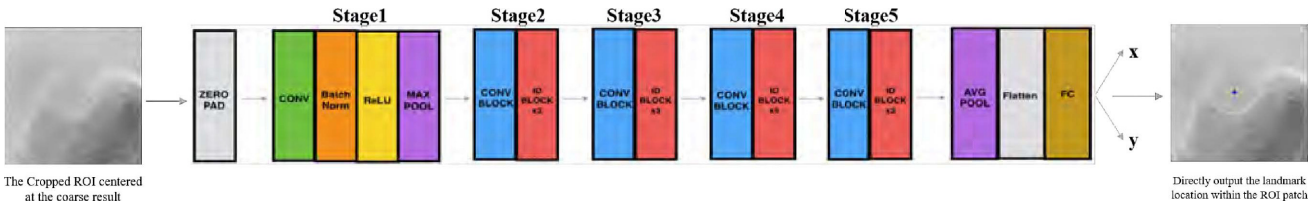


Fig. 5 The overall architecture of ResNet50.

estimation or coarse landmark detection of the test image. The inverse transformation (Φ_{test}^{-1}) is represented as Eq. (2):

$$\Phi_{test}^{-1} = \Phi_{test}(-i, -j) \tag{2}$$

Compared with our previous method [15], which used a rigid registration method to register the test image to all training images so that a best match training image for each test image can be found, the proposed deep learning-based registration method is very fast and efficient since we use a trained model to estimate the displacement field (parameters). In addition, the deformable transformation model is trained for non-rigid registration, we do not need to find the best match training image for the coarse landmark detection (extraction of ROIs).

2.3 Fine Landmark Detection

Though we can estimate or detect the landmarks roughly in the coarse detection step, it is not accurate enough. Therefore, we cut off ROIs for each landmark centered on their predicted coarse positions and perform a fine detection in each ROI using CNN models as shown in Fig. 2(the fine step). In the corarse detection step, we resize images to 1/5 of original size to reduce computational time, however, the ROIs we cropped in the fine step come from the original resolution images. We use a ResNet50 [18] to detect the exact landmark as shown in Fig. 5, adding fully connected layer for regression after feature extraction. The input is the ROI image and the output is landmark's coordinate. The reason we choose the ResNet50 is that it is one of the state-of-the-art CNNs and it is efficient when facing gradient vanishing problems. Since every different landmark is in a different anatomic structure, we train the network independently for each landmark. Thus, we have 19 models for each landmark. Note that direct regression on all landmarks is a highly non-linear mapping which is difficult to learn [19]–[21]. But in the proposed method, each landmark has its specific non-linear mapping function (model).

The loss function we use for training is Mean Squared Error. It can be written as Eq. (3).

$$MSE = \frac{1}{n}\sum_{i=1}^{n}((x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2) \tag{3}$$

Where x_i and y_i represent the ground-truth coordinate of landmark i, $\hat{x}_i$ and $\hat{y}_i$ represent the estimated coordinate of landmark i. Since we only have limited number of training data (150 in ISBI dataset), we make data augmentation to all annotated data in the training step. We randomly crop a region around the ground-truth landmark positions, every region includes the landmark and the landmark could be everywhere in the cropped region. For each landmark, we crop 200 images in one X-ray image, which means that we increase training data 200 times than before for each landmark.

The detection procedure is quite easy. We first get the coarse landmarks' locations through trained displacement field. Then we cropped ROIs centered at the coarse locations. After that, we input each ROI into corresponding trained ResNet50 model, making predictions directly.

3. Experiments and Results

3.1 Datasets

We evaluate our method using International Symposium on Biomedical Imaging (ISBI) 2015 Cephalometry X-ray image analysis Challenge dataset [5]. It includes 150 x-ray images for training, 150 images in testset 1 and 100 images in testset 2. Each image is 1935 x 2400 pixels in Tiff format, where each pixel is 0.1 x 0.1 mm. Each image has 19 landmarks to be detected, the annotations are performed by two experienced doctors. In our experiment, we calculate the average of two annotations from two doctors as our ground-truth.

3.2 Implementation Details

We use Titan-X GPU to help us accelerating training procedure. We use Python programming language, tensorflow and keras deep learning tools, to implement our experiment. For coarse landmark detection model, we choose one training image (the closest one to the average image of taining images) as reference image and all other training images as moving images to train the CNN model. For refined landmark detection models, we use all 150 annotated training images to train the CNN models, after doing data augmentation by randomly cropping 200 patches for each landmark in each image, we have 30000 training images (200*150) for each landmark.

3.3 Evaluation Measurements

According to ISBI grand challenge [5], we use the mean radial error (MRE) and successful detection rate (SDR) to evaluate the performance. Radial error is defined as follows:

$$R = \sqrt{\Delta x^2 + \Delta y^2}$$

And the MRE is defined as follows:

$$MRE = \frac{\sum_{i=1}^{n} R_i}{n}$$

where Δx and Δy are the differences of x-axis and y-axis between predicted landmark location and ground-truth, n is the total number of test images. The definition of Successful Detection is as follows: If the radial error between the predicted landmark and the ground-truth value is no greater than z mm (where z = 2, 2.5, 3, 4), the detection is considered as a successful one (Usually, 2mm range is acceptable in medical analysis). The definition of SDR is shown as follows:

$$SDR = \frac{N_a}{N} * 100\%$$

where Na indicates the number of successful detections and N indicates the number of total detections.

3.4 Performance of the Proposed Method

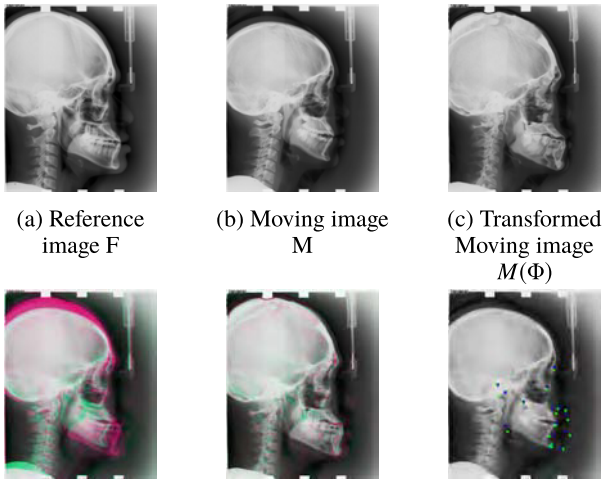
3.4.1 Coarse Landmark Detection Results

In the training phase, we choose image that is closest to the average image of training images (No.126) as our reference image. For the moving image, we use all other 149 training images. We train a U-Net based CNN to generate displacement fields. In the test phase, we input the test image and the reference image (No.126) into our trained network, the output will be the predicted displacement field. We get the coarse landmarks' locations by tracing back the displacement field. The input images' sizes are downsampled to 1/5 of original size in this step. One of the typical detection results is shown in Fig. 6. The reference image, moving image, transformed moving image, composed reference image and moving image (before transformation), composed reference image and transformed moving image (after transformation), comparison of detected landmarks (green) and ground truth (blue) are shown in Figs. 6(a)-(f), respectively. As we can see, the transformed moving image becomes more similar with reference image. The MRE of this coarse step is shown as Table 1, the MRE is calculated using the original resolution. Note that the coarse step aims to locate the landmarks' ROIs, as long as landmarks are within the ROIs, their locations can be refined in the fine detection step. Also, in Fig. 6, the predicted landmark seems to be really close to ground-truth, but this is the resized image (1/5 of original size), which means the actual distance should be 5 times larger.

3.4.2 Refined Landmark Detection

To continue refining the coarse location, we train 19 ResNet models with same architecture. In training phase, since there are only 150 training images, the number is insufficient for training a deep convolutional neural network. We do data augmentation as described in Sect. 2.3. We randomly crop 200 patch images which includes the landmark, the landmark could be everywhere in the patch. The patches are cropped from original size (1935 x 2400) x-ray images. Each cropped patch image is 512 x 512 pixels, we resize them to 256 x 256 pixels, treating them as our training dataset. The training images are 30000 (200 x 150) for each landmark. In the test phase, we first get the coarse landmark location through displacement field generated from trained U-Net weights, then cutting patches centered at the coarse landmark location.

We input the cropped patches respectively to the trained ResNet models, outputting the locations in the patch images directly. The SDR and MRE results on test dataset 1 and test dataset 2 are shown in Table 2. The results are calculated based on original resolution. One of the typical



(a) Reference image F (b) Moving image M (c) Transformed Moving image $M(\Phi)$

(d) Composed image of F and M (e) Composed image of F and $M(\Phi)$ (f) Result example G:Predicted B:GT

Fig. 6 Example of coarse landmark detection results.

Table 1 Coarse step's results of testset1 and testset2 on MRE(mean radial error)

Anatomical Landmarks	MRE on Testset1(mm)	MRE on Testset2(mm)
1. sella turcica	6.610	7.565
2. nasion	9.204	9.827
3. orbitale	8.096	10.198
4. porion	5.301	6.608
5. subspinale	7.195	6.683
6. supramentale	8.563	7.910
7. pogonion	10.499	8.707
8. menton	11.137	8.365
9. gnathion	10.951	8.557
10. gonion	10.293	9.890
11. lower incisal incision	8.029	7.687
12. upper incisal incision	7.793	7.112
13. upper lip	9.458	9.922
14. lower lip	10.336	9.325
15. subnasale	9.022	8.729
16. soft tissue pogonion	11.611	10.905
17. posterior nasal spine	6.295	7.210
18. anterior nasal spine	7.865	7.241
19. articulate	5.121	6.303
Average:	8.598	8.355

Table 2 SDR and MRE Results on test dataset1 and test dataset2 of 2mm, 2.5mm, 3mm, 4mm range

Test dataset 1 and Test dataset2										
Anatomical Landmarks	2mm(%)		2.5mm(%)		3mm(%)		4mm(%)		MRE(mm)	
1. sella turcica	97.3	94.0	98.0	94.0	98.0	94.0	98.0	94.0	0.759	1.802
2. nasion	86.0	85.0	91.3	90.0	93.3	92.0	96.0	96.0	1.212	1.096
3. orbitale	84.0	33.0	94.0	51.0	95.3	72.0	98.0	93.0	1.302	2.808
4. porion	69.3	70.0	78.0	78.0	82.0	82.0	92.7	92.0	1.849	1.973
5. subspinale	69.3	77.0	80.0	93.0	89.3	97.0	100.0	96.7	1.629	1.300
6. supramentale	85.3	34.0	93.3	48.0	97.3	63.0	99.3	87.0	1.186	2.640
7. pogonion	94.0	98.0	97.3	99.0	98.7	99.0	99.3	99.0	0.866	0.748
8. menton	88.0	95.0	94.0	97.0	95.3	98.0	95.3	99.0	1.258	0.799
9. gnathion	94.0	99.0	97.3	99.0	98.0	99.0	98.7	99.0	0.895	0.676
10.gonion	60.0	67.0	72.0	81.0	82.7	86.0	90.7	97.0	1.966	1.999
11.lower incisal incision	96.0	94.0	96.7	96.0	98.0	96.0	98.7	99.0	0.719	0.823
12.upper incisal incision	96.0	97.0	97.3	97.0	98.0	98.0	99.3	99.0	0.554	0.482
13.upper lip	80.0	7.0	93.3	29.0	97.3	59.0	99.3	96.0	1.555	2.857
14.lower lip	98.0	62.0	100.0	83.0	100.0	92.0	100.0	100.0	0.891	1.875
15.subnasale	92.0	96.0	95.3	97.0	98.0	98.0	99.3	100.0	0.990	0.939
16.soft tissue pogonion	88.7	4.0	94.0	7.0	96.0	10.0	98.0	37.0	1.127	4.397
17.posterior nasal spine	92.7	88.0	95.3	93.0	98.0	93.0	99.3	96.0	0.880	1.240
18.anterior nasal spine	87.3	94.0	92.7	96.0	96.7	98.0	97.3	100.0	1.167	0.934
19.articulate	61.3	78.0	72.0	83.0	81.3	86.0	90.7	94.0	1.871	1.821
Average:	85.2	72.2	91.2	79.5	94.4	85.0	97.2	93.5	1.194	1.643

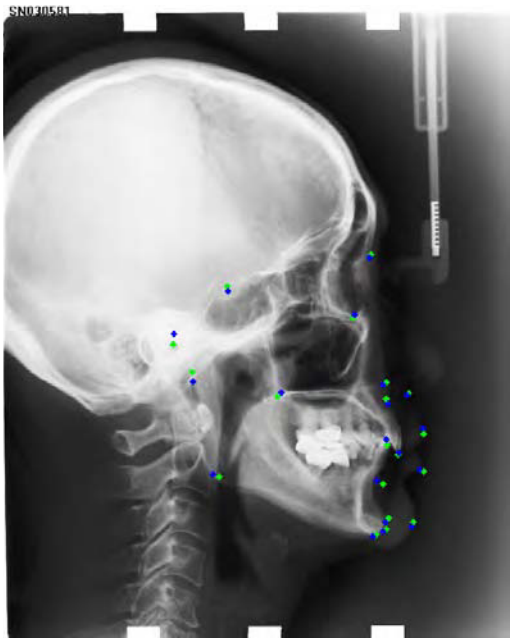


Fig. 7 Example of one result. Green:predicted result Blue:ground-truth

detection results is shown as Fig. 7.

3.4.3 Comparison

The comparison with other state-of-the-art methods is shown in Table 3. The comparison with our previous method is shown in Table 4. Notice that the reason we multiply computational time by 150 is because we need to find the best matching image as our reference image in our previous method, in other words, we register each test image to every training image to find the best reference image.

4. Conclusion and Discussion

The proposed improved coarse-to-fine method achieves satisfying performance in automatic cephalometric landmark detection. Especially, for the coarse landmark detection, we locate the ROIs in very short time. After the refined detection, the result surpasses other state-of-the-results. What's more, compared with our previous method, the computational time is largely reduced, only about 1/3000 time spent per test image, while maintaining the detection accuracy.

For the coarse location in the coarse detection step, it can be seen as a positional normalization to find the ROI of the landmark. Since coarse locations are used to locate regions of interests (ROIs) that include landmarks, it would be meaningless if landmarks are not included in the ROIs. We found that three images have the situation that landmarks are not included ROIs, which is about 1.2% (3/250) of test images. We think this is acceptable. As long as the ROI includes landmarks, our trained ResNet CNNs can detect them correctly. The accuracy of coarse detection is shown in Table 1. The MRE for test dataset1 and test dataset2 are 8.598mm and 8.355mm, respectively, while as shown in Table 3, the MRE and be significantly improved to 1.194mm and 1.613mm, respectively, by using fine detection step.

We also performed traditional method using both rigid

Table 3 SDR proposed in this paper compared with other benchmarks for ISBI 2015 grand challenge Testset1 and Testset2.

Comparisons of SDR								
Method	2mm(%)		2.5mm(%)		3mm(%)		4mm(%)	
	Test 1	Test 2	Test 1	Test 2	Test 1	Test 2	Test 1	Test 2
Ibrgimov [7]	71.9	62.7	77.4	70.5	81.9	76.5	88.0	85.1
Lindner [6]	73.7	66.1	80.2	72.0	85.2	77.6	91.5	87.4
Arik [12]	75.4	67.7	80.9	74.2	84.3	79.1	88.3	84.6
Qian [14]	82.5	**72.4**	86.2	76.2	89.3	79.7	90.6	85.9
Proposed Method	**85.2**	72.2	**91.2**	**79.5**	**94.4**	**85.0**	**97.2**	**93.5**

Table 4 MRE and Computational time per image compared with previous method

Comparisons of MRE and Computation Time			
Method	MRE of Test1(mm)	MRE of Test2(mm)	Computation Time(s)
Previous Method [15]	1.077	1.542	85.3 x 150
Proposed Method	1.194	1.643	4.0

and non-rigid registration for landmark detection to validate the effectiveness of our U-Net based coarse registration. We used affine transform for alignment first. After that, we use a displacement field transform to warp the moving image [22]. The MRE for testset1 and testset2 is 10.7mm and 11.2mm respectively. For computational time, one registration takes 280 seconds in average. Compared with our U-Net-based method, which is shown in Table 1, the conventional registration method not only takes large computational time, but also result in poor performance.

Neither our proposed U-Net-based method nor the traditional rigid and non-rigid registration method achieved satisfying result. So the second fine detection step is needed to achieve accurate landmark detection. We think the large image resolution (1980 x 2400), as well as the strict medical acceptable landmark error (within 2mm), limit the effectiveness of registration methods, thus, making registration methods only appropriate for locating coarse regions of each landmark.

Some landmark have relatively low SDR in test dataset2 compared with those in test dataset1. As we explained in the previous paper [15], due to the extreme differnet anatomical structure of test dataset2 from training dataset and test dataset1, some landmarks cannot be accurately located using the trained CNN models.

In conclusion, our proposed method is fast and accurate, which is applicable for practical use.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under the Grant No. 61603218, in part by the Grant-in Aid for Scientific Research from the Japanese Ministry for Education, Science, Culture and Sports (MEXT) under the Grant No. 20KK0234 and No. 20K21821, and also in part by the Zhejiang University Global Partnership Fund under the Grant No. 100000-11320.

References

[1] S. Albarakati, K. Kula, and A. Ghoneima, "The reliability and reproducibility of cephalometric measurements: a comparison of conventional and digital methods," Dentomaxillofacial Radiology, vol.41, no.1, pp.11–17, 2012.

[2] L. Devereux, D. Moles, S.J. Cunningham, and M. McKnight, "How important are lateral cephalometric radiographs in orthodontic treatment planning?," American J. Orthodontics and Dentofacial Orthopedics, vol.139, no.2, pp.e175–e181, Feb. 2011.

[3] P.G. Nijkamp, L.L. Habets, I.H. Aartman, and A. Zentner, "The influence of cephalometrics on orthodontic treatment planning," The European J. Orthodontics, vol.30, no.6, pp.630–635, Dec. 2008.

[4] C.W. Wang, C.T. Huang, M.C. Hsieh, C.H. Li, S.W. Chang, W.C. Li, R. Vandaele, R. Marée, S. Jodogne, P. Geurts, C. Chen, G. Zheng, C. Chu, H. Mirzaalian, G. Hamarneh, T. Vrtovec, and B. Ibragimov, "Evaluation and comparison of anatomical landmark detection methods for cephalometric x-ray images: a grand challenge," IEEE Trans. Medical Imaging, vol.34, no.9, pp.1890–1900, Sept. 2015.

[5] C.W. Wang, C.T. Huang, J.H. Lee, C.H. Li, S.W. Chang, M.J. Siao, T.M. Lai, B. Ibragimov, T. Vrtovec, O. Ronneberger, P. Fischer, T.F. Cootes, and C. Lindner, "A benchmark for comparison of dental radiography analysis algorithms," Medical image analysis, vol.31, pp.63–76, July 2016.

[6] C. Lindner and T.F. Cootes, "Fully automatic cephalometric evaluation using random forest regression-voting," IEEE Int. Symp. Biomedical Imaging, Citeseer, 2015.

[7] B. Ibragimov, B. Likar, F. Pernus, and T. Vrtovec, "Computerized cephalometry by game theory with shape-and appearance-based landmark refinement," Proc. Int. Symp. Biomedical imaging (ISBI), 2015.

[8] A. Krizhevsky, I. Sutskever, and G.E. Hinton, "Imagenet classification with deep convolutional neural networks," Advances in neural information processing systems, vol.25, pp.1097–1105, Dec. 2012.

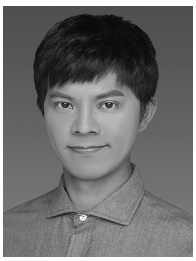
[9] D.C. Ciresan, U. Meier, J. Masci, L.M. Gambardella, and J. Schmidhuber, "Flexible, high performance convolutional neural networks for image classification," Proc. Twenty-Second International Joint Conference on Artificial Intelligence, pp.1237–1242, July 2011.

[10] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp.3431–3440, 2015.

[11] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," Int. Conf. Medi-

cal Image Computing and Computer-Assisted Intervention, LNCS, vol.9351, pp.234–241, Springer, 2015.
[12] S.Ö. Arik, B. Ibragimov, and L. Xing, "Fully automated quantitative cephalometry using convolutional neural networks," J. Medical Imaging, vol.4, no.1, p.014501, 2017.
[13] H. Lee, M. Park, and J. Kim, "Cephalometric landmark detection in dental x-ray images using convolutional neural networks," Medical Imaging 2017: Computer-Aided Diagnosis, p.101341W, International Society for Optics and Photonics, 2017.
[14] J. Qian, M. Cheng, Y. Tao, J. Lin, and H. Lin, "Cephanet: An improved faster r-cnn for cephalometric landmark detection," 2019 IEEE 16th Int. Symp. Biomedical Imaging (ISBI 2019), pp.868–871, 2019.
[15] Y. Song, X. Qiao, Y. Iwamoto, and Y.w. Chen, "Automatic cephalometric landmark detection on x-ray images using a deep-learning method," Applied Sciences, vol.10, no.7, p.2547, 2020.
[16] G. Balakrishnan, A. Zhao, M.R. Sabuncu, J. Guttag, and A.V. Dalca, "An unsupervised learning model for deformable medical image registration," Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp.9252–9260, 2018.
[17] A.V. Dalca, G. Balakrishnan, J. Guttag, and M.R. Sabuncu, "Unsupervised learning for fast probabilistic diffeomorphic registration," Int. Conf. Medical Image Computing and Computer-Assisted Intervention, LNCS, vol.11070, pp.729–738, Springer, 2018.
[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp.770–778, 2016.
[19] T. Pfister, J. Charles, and A. Zisserman, "Flowing convnets for human pose estimation in videos," Proc. IEEE Int. Conf. Comput. Vis., pp.1913–1921, 2015.
[20] C. Payer, D. Štern, H. Bischof, and M. Urschler, "Regressing heatmaps for multiple landmark localization using cnns," Int. Conf. Medical Image Computing and Computer-Assisted Intervention, pp.230–238, LNCS, vol.9901, Springer, 2016.
[21] J.J. Tompson, A. Jain, Y. LeCun, and C. Bregler, "Joint training of a convolutional network and a graphical model for human pose estimation," Advances in neural information processing systems, vol.27, pp.1799–1807, Dec. 2014.
[22] D. Rueckert, L.I. Sonoda, C. Hayes, D.L. Hill, M.O. Leach, and D.J. Hawkes, "Nonrigid registration using free-form deformations: application to breast mr images," IEEE Trans. Medical Imaging, vol.18, no.8, pp.712–721, Aug. 1999.

Yu Song received a B.S. degree in 2014 from Huaqiao University, Xiamen, China and a M.S. degree in 2020 from Ritsumeikan University, Kusatsu, Japan. He is currently a visiting researcher at Ritsumeikan University. His research interests include image processing and machine learning.



Xu Qiao received the B.S. degree in mathematical statistics and the M.S. degree from Shandong University, China, in 2004 and 2007, respectively, and the Ph.D. degree from Ritsumeikan University, Japan, in 2010. From 2010 to 2012, he was a Research Fellow with the Japan Society for the Promotion of Science (JSPS). From 2012 to 2018, he was a Lecturer with the School of Control Science and Engineering, Shandong University. Since 2018, he has been an Associate Professor of biomedical engineering with the School of Control Science and Engineering, Shandong University. His research interests include imaging diagnosis and medical image analysis.

Yutaro Iwamoto received the B.E. and M.E., and D.E. degree from Ritsumeikan University, Kusatsu, Japan in 2011 and 2013, and 2017, respectively. He is currently an Assistant Professor at Ritsumeikan University, Kusatsu, Japan. His current research interests include medical image processing and computer vision, and deep learning.

Yen-Wei Chen received a B.E. degree in 1985 from Kobe Univ., Kobe, Japan, a M.E. degree in 1987, and a D.E. degree in 1990, both from Osaka University, Osaka, Japan. From 1991 to 1994, he was a research fellow with the Institute for Laser Technology, Osaka. From October 1994 to March 2004, he was an associate Professor and a professor with the Department of Electrical and Electronic Engineering, University of the Ryukyus, Okinawa, Japan. He is currently a professor with the college of Information Science and Engineering, Ritsumeikan University, Kyoto, Japan. He is also a visiting professor with the Zhejiang Lab, China and the College of Computer Science and Technology, Zhejiang University, China. He is an associate Editor of International Journal of Image and Graphics(IJIG) and an associate Editor of the International Journal of Knowledge based and Intelligent Engineering Systems. His research interests include pattern recognition, image processing and machine learning. He has published more than 200 research papers in these fields.

Yili Chen received a B.E. degree in 1994 from China Medical University, Liaoning, China and a D.E. degree in 2010 from Tokyo University, tokyo, Japan. He is currently a director at the Department of Neurosurgery, Fourth Affiliated Hospital, School of Medicine, Zhejiang University, China.