Evaluation and Comparison of Anatomical Landmark Detection Methods for Cephalometric X-Ray Images: A Grand Challenge

Ching-Wei Wang*, *Member, IEEE,*
Cheng-Ta Huang, Meng-Che Hsieh, Chung-Hsing Li, Sheng-Wei Chang, Wei-Cheng Li, Rémy Vandaele, Raphaël Marée, Sébastien Jodogne, Pierre Geurts, Cheng Chen, Guoyan Zheng, *Member, IEEE,* Cheng-wen Chu, Hengameh Mirzaalian, Ghassan Hamarneh, *Senior Member, IEEE,* Tomaž Vrtovec and Bulat Ibragimov

***Abstract*—Cephalometric analysis is an essential clinical and research tool in orthodontics for the orthodontic analysis and treatment planning. This paper presents the evaluation of the methods submitted to the *Automatic Cephalometric X-Ray Landmark Detection Challenge*, held at the IEEE International Symposium on Biomedical Imaging 2014 with an on-site competition. The challenge was set to explore and compare automatic landmark detection methods in application to cephalometric X-ray images. Methods were evaluated on a common database including cephalograms of 300 patients aged six to 60 years, collected from the Dental Department, Tri-Service General Hospital, Taiwan, and manually marked anatomical landmarks as the ground truth data, generated by two experienced medical doctors. Quantitative evaluation was performed to compare the results of a representative selection of current methods submitted to the challenge. Experimental results show that three methods are able to achieve detection rates greater than 80% using the 4 mm precision range, but only one method achieves a detection rate greater than 70% using the 2 mm precision range, which is the acceptable precision range in clinical practice. The study provides insights into the performance of different landmark detection approaches under real-world conditions and highlights achievements and limitations of current image analysis techniques.**

***Index Terms*—Challenge, evaluation, cephalometric analysis, landmark detection, dental X-ray images**

*C.-W. Wang, C.-T. Huang, M.-C. Hsieh are with the Graduate Institute of Biomedical Engineering, National Taiwan University of Science and Technology, Taipei, 106 Taiwan. E-mail: {cweiwang,chengta.haung}@mail.ntust.edu.tw. *Asterisk indicates the corresponding author.*

*C.-W. Wang is also with the Department of Biomedical Engineering, National Defense Medical Center, Taiwan.

C.-H. Li, S.-W. Chang and W.-C. Li are with the Orthodontics and Pediatric Dentistry Division, Dental Department, Tri-Service General Hospital, Taiwan.

C.-H. Li is also with the School of Dentistry and Graduate Institute of Dental Science, National Defense Medical Center, Taipei, Taiwan.

R. Vandaele, R. Marée and P. Geurts are with the Systems and Modeling, Department of EE and CS & GIGA-R, University of Liège, Belgium.

R. Vandaele and S. Jodogne are with the Department of Medical Physics, C.H.U., University of Liège, Belgium.

C. Chen, G. Zheng and C. Chu are with the Institute for Surgical Technology and Biomechanics, University of Bern, Switzerland.

H. Mirzaalian and G. Hamarneh are with the Department of Computing Science, Simon Fraser University, Canada.

B. Ibragimov and T. Vrtovec are with the Faculty of Electrical Engineering, University of Ljubljana, Slovenia.

I. INTRODUCTION

Analysis of two-dimensional (2D) X-ray images of the craniofacial area, i.e. cephalograms, is the backbone of modern orthodontics, orthognathic treatment and maxillofacial surgery. Marking of anatomical landmarks in cephalometric analysis is necessary as it provides the interpretation of patients' bony structures and gives the whole picture for surgeons in orthognathic surgery (OGS). Clinical applications of cephalometric analysis in surgical planning include diagnosis and treatment of obstructive sleep apnea [1]–[3], assessment of mandible/lower jaw [4]–[6] and assessment of soft facial tissue [7], [8]. In 1982, Rakosi [12] defined 90 landmarks, which have been used by orthodontists for clinical research. Among these, 19 landmarks were commonly adopted in recent studies [13], [18]–[20], [23] and for clinical practice. In this grand challenge, these 19 landmarks are selected in the standard evaluation framework.

In clinical practice, manual marking is required, and orthodontists usually trace out craniofacial structure contours on X-ray images first, and then extract landmarks from linear and angular reference lines and other geometrical shapes. However, manual marking is time-consuming and subjective. An experienced orthodontist requires up to 20 minutes to perform one X-ray cephalometric analysis [13], and this tends to suffer from intra- and inter-observer errors [14]. In order to deal with issues of manual marking, many clinical research studies focused on landmark identification problems [9]–[11].

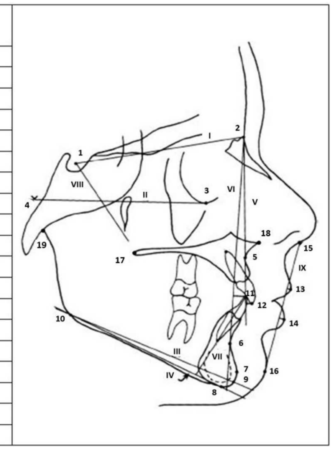
Automated landmark detection for cephalometry can be the solution to facilitate these issues. However, auto-identification of anatomical landmarks is difficult and poorly explored due to the complexity and variability of cephalometric X-ray images. Challenges of auto-identification of anatomical landmarks in cephalometric X-ray images include variations on individual skeletal structures, image blurs caused by device-specific projection magnifications, and image complexity due to the overlapping contralateral structures [15]–[17]. Some studies focused on investigating methods for automatic localization/identification of cephalometric landmarks [18]–[20]. In 2006, Yue et al. [18] built a modified active shape model to detect 12 anatomical landmarks, achieving a 71% success rate of landmark detection within 2.0 mm and 88% within 4.0

mm. In 2009, Kafieh et al. [19] combined neural networks with modified active shape models and developed a technique with 93% landmark detection accuracy for bony structures within 5.0 mm. Kaur and Singh [20] proposed an automatic cephalometric landmark detection method in 2013, which uses Zernike moment-based features for initial landmark estimation and computes small expectation window, and resulted in a mean error of 1.84 mm and standard deviation is 1.24 for 18 observed landmarks. However, variations in image data (e.g. image quality, imaging radiation time, subject age and gender) and evaluation approaches (e.g. the size of the manually obtained ground truth data from one pixel to a square region) greatly influence the evaluation outcome.

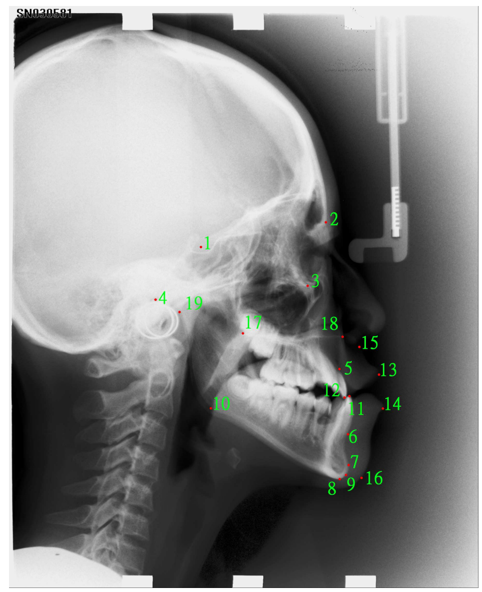
This paper presents the evaluation and comparison of a representative selection of current methods presented during the Automatic Cephalometric X-Ray Landmark Detection Challenge (ACXRLDC) [22] held in conjunction and with the support of the IEEE International Symposium on Biomedical Imaging 2014 (ISBI 2014) (Website: http://www-o.ntust.edu.tw/~cweiwang/celph/). The aim of this challenge was to investigate suitable automatic landmark detection technologies for cephalometric X-ray images and provide a standard evaluation framework with a real clinical data set, containing cephalograms of 300 patients aged six to 60 years. Nineteen commonly popularly adopted anatomical landmarks were used as targeting landmarks to detect. Extensive quantitative evaluation was performed to assess the performance of methods with respect to manual markings. This is, to our knowledge, the first image analysis challenge undertaken in the dental X-ray imaging field, providing a reference publication from which to gauge how well a representative selection of current methods works, and encouraging others to work in this area. Figure 1 presents the 19 anatomical landmarks used in this challenge.

The remainder of the paper is organized as follows. In Section II, challenge aims, image data sets used within the challenge and participants to the challenge are described. Evaluation approaches to compare the detection results are presented in Section III. Section IV summarizes the methodologies presented to the challenge. In Section V, quantitative results and analysis are presented. Finally, conclusions are given in Section VI.

No.	Anatomical Landmarks
1	sella turcica
2	nasion
3	orbitale
4	porion
5	subspinale
6	supramentale
7	pogonion
8	menton
9	gnathion
10	gonion
11	lower incisal incision
12	upper incisal incision
13	upper lip
14	lower lip
15	subnasale
16	soft tissue pogonion
17	posterior nasal spine
18	anterior nasal spine
19	articulate



(a) Cephalometric tracing of anatomical structures.



(b) The 19 anatomical landmarks used in this challenge.

Fig. 1. Cephalometric tracing and anatomical landmarks.

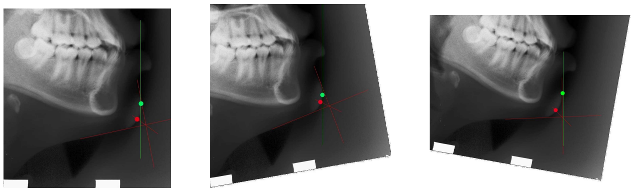
II. Automatic Cephalometric X-Ray Landmark Detection Challenge

A. Organization

The goal of the challenge was to investigate automatic methods for identifying anatomical landmarks of cephalometric X-ray images. The challenge was opened to teams from academia and industry and held in conjunction and with the support of ISBI 2014. Nineteen landmarks were chosen in this challenge, including the sella, the nasion, the orbitale, the porion, the subspinale, the supramentale, the pogonion, the menton, the gnathion, the gonion, the lower incisal incision, the upper incisal incision, the upper lip, the lower lip, the subnasal, the soft tissue pogonion, the posterior nasal spine, the anterior nasal spine, the anterior nasal spine and the articulate. The results of individual teams were automatically compared to the ground truth data. There were two stages in the challenge. In stage 1, a training data set and a first test data set were released for method development. In stage 2, an on-site competition was organized for which a second test data set was used.

B. Description of image data sets

There were 300 cephalometric X-ray images collected from 300 patients aged six to 60 years. The images were acquired in TIFF format with Soredex CRANEX® Excel Ceph machine (Tuusula, Finland) and Soredex SorCom software (3.1.5, version 2.0), and the image resolution was 1935 × 2400 pixels. Regarding the ground truth data for evaluation, 19 landmarks were manually marked in each image and reviewed by two experienced medical doctors. An ethical approval was obtained to conduct the study with IRB Number 1-102-05-017, which was approved by the research ethics committee of the Tri-Service General Hospital in Taipei, Taiwan. The whole data set was equally and randomly divided into three subsets as Training data, Test1 data and Test2 data for the challenge with



(a) 0 degree (b) 10 degree counter-clockwise rotation (c) 10 degree clockwise rotation

Fig. 2. The positions of the soft tissue pogonion by the elder marking method (highlighted with the green point) and the newer robust method (highlighted with the red point) in three situations with 0 degree, 10 degree counter-clockwise rotation and 10 degree clockwise rotation to simulate different natural head positions (NHPs). The location of the red point remains the same even for different NHPs. In comparison, the location of the green point changes in three different situations.

two stages.

In generating the ground truth data, there were two ways to mark Landmark 16 (the soft tissue pogonion), and a robust marking method was adopted. An elder definition of Landmark 16 was as the most anterior point on the contour of the chin, but the location may vary if the natural head position was not performed correctly [47]. Therefore, in clinical practice a more robust marking approach was used by drawing two lines parallel to the chin, extending the bisector of these two lines to the chin, and identifying the intersection point as Landmark 16 (Figure 2). Hence, in this work, the latter marking approach was adopted.

Two experts with different degrees of expertise participated in defining the landmark ground truth data twice. The average of all four annotations was used as the ground truth. The experts for this challenge had the following level of expertise.

- Expert 1: Clinician (orthodontist) with 15 of years experience.
- Expert 2: Clinician (orthodontist) with six years of experience.

The intra- and inter-observer variabilities were calculated independently for each expert using the metric defined in Section III and are presented in Table I.

TABLE I
INTRA- AND INTER-OBSERVER VARIABILITY OF MANUAL LANDMARKING IN TERMS OF MEAN RADIAL ERROR (MRE) $\pm$ STANDARD DEVIATION (SD)

	Intra-Observer Variability		**Inter-Observer Variability**
	Expert 1	**Expert 2**	**Expert 1 vs Expert 2**
MRE (mm)	1.73±1.35	0.90±0.89	1.38±1.55

C. Participation in the challenge

A total of 17 teams (from 14 countries) registered in this challenge, and initially six teams submitted their results on the Test1 data. However, one team obtained a detection rate of less than 30%, and therefore only five teams were accepted in stage 1 and invited to participate at the on-site competition in stage 2.

1) **Ibragimov** *et al.* [26], Automatic cephalometric X-ray landmark detection by applying game theory and random forests (Slovenia).
2) **Chu** *et al.* [27], Fully automatic cephalometric X-ray landmark detection using random forest regression and sparse shape composition (Switzerland).
3) **Chen and Zheng** [28], Fully-automatic landmark detection in cephalometric X-ray images by data-driven image displacement estimation (Switzerland).
4) **Mirzaalian and Hamarneh** [29], Automatic globally-optimal pictorial structures with random decision forest based likelihoods for cephalometric X-ray landmark detection (Canada).
5) **Vandaele** *et al.* [30], A machine learning tree-based approach (Belgium).

III. EVALUATION METHODS

Two main criteria are considered to evaluate the performance of submitted methods. The first evaluation criterion is the mean radial error with the associated standard deviation. The radial error R is formulated as follows:

$$R = \sqrt{\triangle x^2 + \triangle y^2} \tag{1}$$

where $\triangle x$ is the absolute distance in x-direction between the obtained and the reference landmark, and $\triangle y$ is the absolute distance in y-direction between the obtained and the reference landmark. The mean radial error (MRE) and the associated standard deviation (SD) are defined as follows.

$$MRE = \frac{\sum_{i=1}^{N} R_i}{N} \tag{2}$$

$$SD = \sqrt{\frac{\sum_{i=1}^{N} (R_i - MRE)^2}{N-1}} \tag{3}$$

The second evaluation criterion is the success detection rate with respect to the 2.0 mm, 2.5 mm, 3.0 mm and 4.0 mm precision range (see Figure 3). For each landmark, medical doctors marked the location of a single pixel instead of an area as a reference landmark location. If the absolute difference between the detected and the reference landmark is no greater than z, the detection of this landmark is considered as a successful detection; otherwise, it is considered as a misdetection. The success detection rate (SDR) p_z with precision less than z is formulated as follows:

$$p_z = \frac{\#\{j : \parallel L_d(j) - L_r(j) \parallel < z\}}{\#\Omega} \times 100\% \tag{4}$$

where L_d and L_r represent the locations of the detected landmark and the referenced landmark, respectively; z denotes four precision ranges used in the evaluation, including 2.0 mm, 2.5 mm, 3.0 mm and 4.0 mm; $j \in \Omega$, and $\#\Omega$ represents the number of detections made.

IV. METHODS

This section summarizes the methods that were submitted to the challenge. The main advantages and drawbacks of all methods are discussed later in Section V-D. We refer the reader to individual papers for more details.

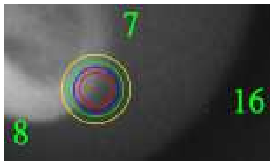


Fig. 3. Four precision ranges of a landmark: 2.0 mm (red circle), 2.5 mm (blue circle), 3.0 mm (green circle) and 4.0 mm (yellow circle).

A. Ibragimov et al. [26]

Ibragimov et al. developed a framework that describes the intensity appearance of each landmark by Haar-like features and applies the theory of random forests (RFs) to combine these features into a landmark candidate point detector. Image pixels with maximal detector response are used as landmark candidate points. To select the optimal candidate points, the framework relies on spatial relationships among pairs of landmarks, modelled by Gaussian kernel density estimation, and the most representative pairs are found by the optimal assignment-based shape representation [24]. The obtained intensity appearance and shape models are combined to optimize the position of landmark candidate points by applying concepts from game theory [25]. The framework also models spatial relationships for subsets of landmarks by RFs and refines the detected optimal candidate points [26], which represent landmarks in the target cephalometric X-ray image. In the training procedure, for each image, Ibragimov et al. manually placed 44 additional auxiliary landmarks at visually distinctive locations such as the nose tip and corners of cervical vertebrae (Fig. 4).

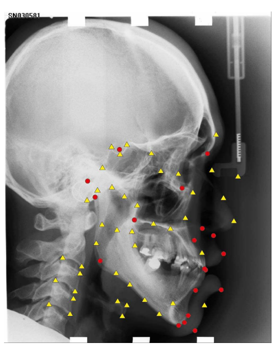


Fig. 4. A cephalometric X-ray image with anatomical landmarks of interests (red circles) and auxiliary landmarks (yellow triangles) used in Ibragimov et al.'s method [26].

B. Chu et al. [27]

Chu et al. built a fully automatic algorithm for landmark detection on 2D cephalometric X-ray images using RF regression. Chu et al.'s algorithm consists of two phases: 1) landmark detection phase and 2) landmark modification phase. To modify landmark positions, which are not correctly detected, a sparse shape composition model is built. The landmark detection phase can be further divided into two steps: a) RF-based landmark detector training step, b) landmark prediction step. RF-based landmark detector training is performed only once on 100 training images and is independent from any test image. Chu et al. train a separate RF-based landmark detector for each landmark [31]–[33], and therefore they obtain 19 detectors for all landmark positions. Given a new test image, they apply trained landmark detectors to the test image to get 19 candidates for pre-defined landmark positions. Then, in the second step for determination of optimal landmark positions, Chu et al. further add prior shape constraints to modify landmark positions. This is accomplished by using a sparse shape composition model, which is computed from a set of pre-aligned training shapes [34]. They iteratively update the shape by alternatively regularizing the shape and updating the shape pose. Finally, they find a sparse and optimized combination of the training shape to regularize the original shape obtained from landmark detection results. Figure 5 presents the flowchart of Chu et al.'s method.

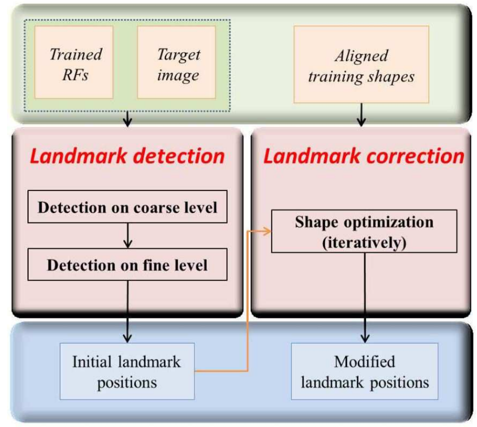


Fig. 5. The flowchart of Chu et al.'s [27] method which includes two phases: (1) landmark detection phase and (2) landmark modification phase.

C. Chen and Zheng [28]

The overview of Chen and Zheng's landmark detection framework is given in Figure 6. In the training step, Chen and Zheng randomly sample a set of square image patches around the ground-truth landmark location. The visual features and displacements to the landmark of these patches comprise the training data. Then, given a new image, they randomly sample a set of image patches and calculate their features. To estimate the displacement of each patch, they take a data-driven approach by defining an objective function on the displacements to enforce the following constraints: (1) Patches with similar features should have similar displacements. (2) The displacements of test patches should be consistent with the triangle structure (i.e. the difference between displacements from the respective centers of two patches to the landmark should be close to the coordinate difference of the two patch centers). In this way, each patch makes a vote on the landmark position. Chen and Zheng treat each vote as a small Gaussian distribution and aggregate them into a probability map, the mode of which is returned as the detected landmark location. For more details, please refer to [35].

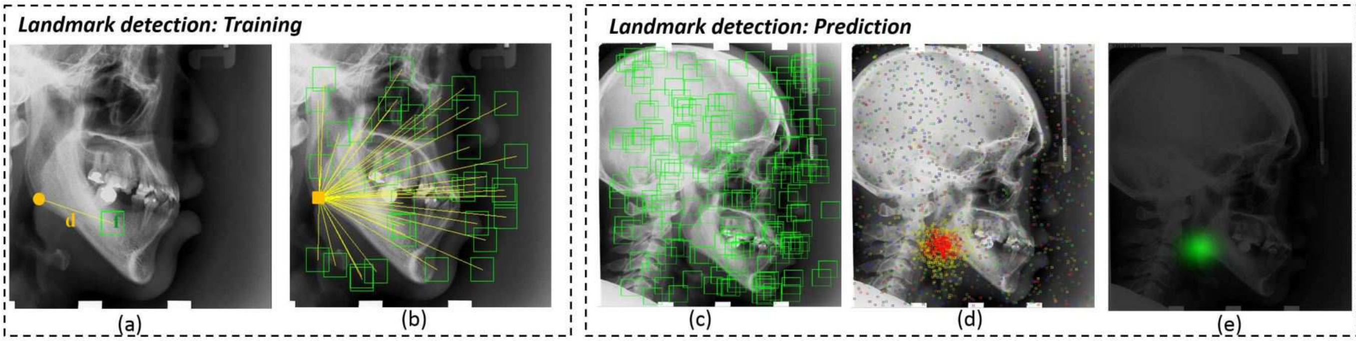


Fig. 6. Overview of Chen and Zheng's [35] landmark detection algorithm. During the training stage, a set of patches are sampled around the ground-truth landmark position (a)(b). During the testing stage, given an image, a set of patches are randomly sampled (c), each patch makes a prediction of the landmark position (d), and then predictions are aggregated to produce a response image (e).

D. Mirzaalian and Hamarneh [29]

Mirzaalian and Hamarneh formulate the problem of finding the cephalometric X-ray landmarks $\mathcal{L}_1 ... \mathcal{L}_{19} \in \mathcal{L}$ as an optimization problem:

$$\hat{\mathcal{L}}(\mathcal{I}) = \text{argmin}_{\mathcal{L}} \sum_{ij} \phi^o(\mathcal{L}_i, \mathcal{L}_j) + \sum_{i=1}^{19} \phi^i(\mathcal{L}_i, \mathcal{I}) \qquad (5)$$

where $\phi^o(\mathcal{L}_i, \mathcal{L}_j)$ is a pairwise regularization term that encourages the vector $\mathbf{L} = \mathcal{L}_i - \mathcal{L}_j$ to conform to a learnt graphical model (calculated using the Mahalanobis distance $(\mathbf{L} - \bar{\mathbf{L}})\mathcal{C}^{-1}(\mathbf{L} - \bar{\mathbf{L}})$; $\bar{\mathbf{L}}$ and $\mathcal{C}$ are the sample mean and covariance) and $\phi^i(\mathcal{L}_i, \mathcal{I})$, $i = 1...19$, are the data (unary) terms that penalize locating the i^{th} landmark at $\mathcal{L}_i$, where each is captured by using local binary pattern-based features [36] of the intensity (1 feature), x and y spatial coordinates of pixels (2 features), and Frangi et al. [37] filtered response (2 scales). Considering intensity information of 5×5 neighbouring pixels, the final feature vector to compute the data terms ϕ^i is of size 125 (=25×[1+2+2]). Examples of the measured ϕ^i, i=1...19, are shown in Figure 7. Mirzaalian and Hamarneh's method has two stages: a training and a testing stage. In the training stage, they extract the aforementioned features from a set of windows centered around the position of landmarks as *true positive* samples, and a set of randomly selected windows *not* near the position of landmarks as *true negative* samples. These positive and negative samples are used to train a RF classifier [31] with 22 trees of depth 10. For a new image, Mirzaalian and Hamarneh apply the pictorial structure algorithm [38] to globally optimize the cost function in Equation (5). Note that the pictorial structure requires computing the first term over an acyclic graph (tree) to connect landmarks. The problem can be solved by applying the minimum spanning tree over a complete graph with the 19 landmarks as vertices and with edge weights set as the covariance values of the edges [38].

E. Vandaele et al. [30]

Following [41], Vandaele et al. address the problem as 19 separate binary pixel classification problems, one for each landmark. One pixel belongs to the positive class if its distance to the landmark position is less than R, where R is a method parameter. Otherwise, it belongs to the negative class. A (x, y) pixel is described by raw pixel values in 16×16 windows centered at translated positions $(x + t_x, y + t_y)$ and extracted at six different resolutions. Parameters t_x and t_y allow to detect the landmark based on a structure which is not centered at the pixel. Vandaele et al. use extremely randomized trees [39] as the pixel classifier. Training pixels are randomly extracted in a radius of at most 4 cm to the landmark. During the prediction phase, they predict the class of N pixel positions drawn from a $\mathcal{N}(\mu_l, \Sigma_l)$ distribution, where $\mu_l \in \mathbb{R}^2$ and $\Sigma_l \in \mathbb{R}^{2\times2}$ are respectively the average and the covariance matrix of landmark positions in training images, and $N = 16\sigma_{x_l}\sigma_{y_l}$ with σ_{x_l} and σ_{y_l} are the diagonal elements of Σ_l. The median position of pixels classified as positive with the highest confidence are then returned as the final landmark position. All method parameters are tuned via a 10-fold cross-validation.

V. Experimental results and analysis

Quantitative evaluation of the detection accuracy and computing efficiency is presented in this section. All the proposed methods are evaluated against the ground truth on 200 testing images, including 100 Test1 images and 100 Test2 images, and the quantitative evaluation results are presented in Section V-A and Section V-B, respectively. Computation times and hardware/software specifications are presented in Section V-C, method advantages and drawbacks are given in Section V-D, and a discussion of method capabilities in provided in Section V-E.

A. Stage 1 with 100 Test1 images

Figure 8 compares the overall results of MRE, SD and SDR using four precision ranges in detection of the 19 anatomical landmarks between five teams on Test1 data. It is observed that in overall, Ibragimov et al.'s method obtains the highest SDRs (72.74% and 78.79%) using 2mm and 2.5mm precision ranges, and the lowest MRE (1.82 mm) and SD (1.61 mm). On the other hand, Vandaele et al.'s method performs best according to SDRs (83.47% and 88.53%) using 3 mm and 4 mm precision ranges. Based on the 4 mm precision range, three methods are able to achieve SDRs greater than 80%, but only two methods (by Ibragimov et al. and Vandaele et al.) obtain SDR more than 70% using the 2 mm precision range,

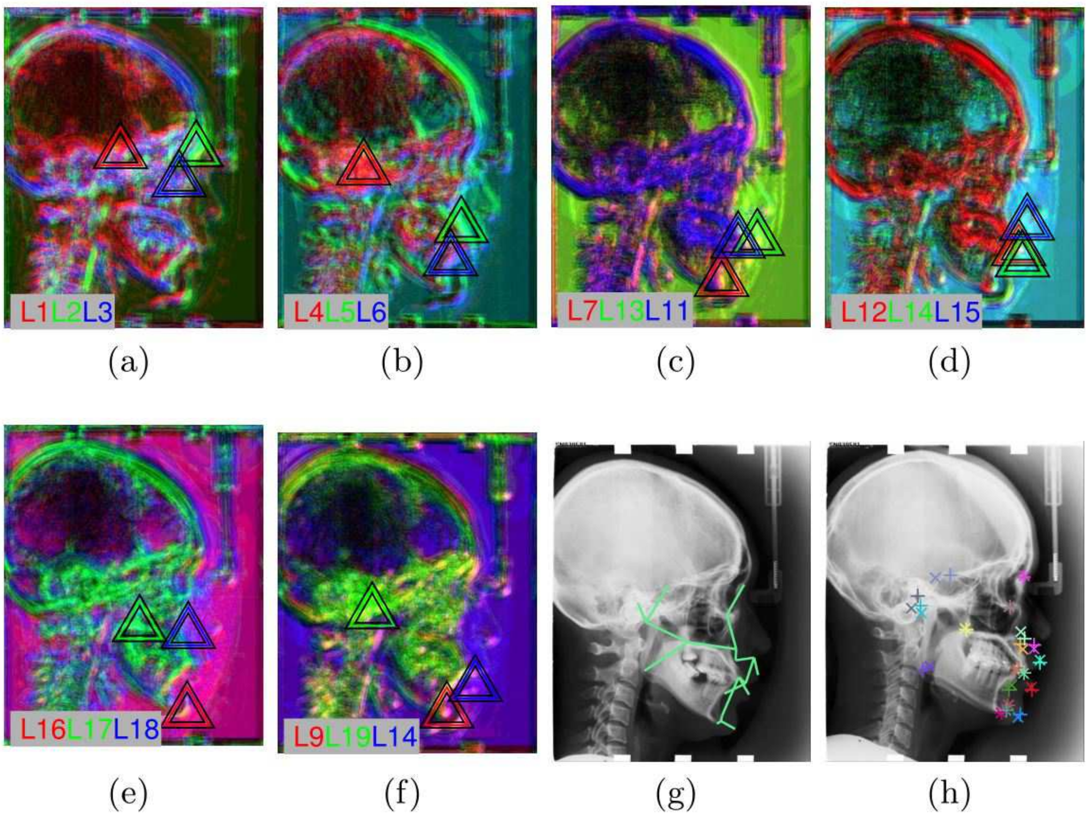


Fig. 7. Details of Mirzaalian and Hamarneh's [29] landmark detection algorithm. Computed likelihoods of three different landmarks for each sub-figure are encoded into the RGB channels (a-f). The centers of triangles represent the ground truth positions of landmarks. (g) The tree resulting from applying the minimum spanning tree algorithm computed in the training phase, which is used in the pictorial algorithm. (h) Detected landmarks (x) and their ground truth (+).

which is the acceptable precision range in clinical practice. The best result on Test1 data is achieved by Ibragimov et al.'s method with 72.74% SDR in 2 mm. The experimental results show that this is still an unsolved and challenging task and needs further investigation.

B. Stage 2 on-site competition with 100 Test2 images

In stage 2, an on-site competition was held with only five accepted teams participating and a separate 100 Test2 images. Figure 9 compares the overall results of MRE, SD and success detection rates using four precision ranges in detection of the 19 anatomical landmarks between five teams on Test2 data. It is observed that in overall, Ibragimov et al.'s method obtains the highest SDRs for all four precision ranges (70.21%, 76.37%, 87.58% and 88.16%), and the lowest MRE (1.92 mm) and SD (1.88 mm). Based on the 4 mm precision range, four methods are able to achieve SDRs higher than 80%, but only one method (Ibragimov et al.) obtains a SDR greater than 70% using the 2 mm precision range.

C. Computer specification and efficiency

The computer specification and efficiency of five accepted teams are summarized below.

1) Ibragimov et al.'s method [26]
   The landmark detection framework was implemented in C#, and executed on a personal computer with Intel Core i7 processor at 2.8 GHz and 8 GB RAM without graphics processing unit-assisted acceleration. Annotation of one cephalogram of size 1935×2400 pixels takes on average 18 seconds.
2) Chu et al.'s method [27]
   The computation time for detecting all 19 landmarks takes on average $60 \sim 120$ seconds per image. All tests were executed on a computer with 3.0 GHz CPU and 12 GB RAM under the OS of Windows 7. The algorithm was implemented using the C++ language.
3) Chen and Zheng's method [28]
   Chen and Zheng implemented their method on a PC with 3.0 GHz CPU and 16 GB RAM running Windows 7. Chen and Zheng's impementation is in Matlab. Under this configuration, their method takes on average 90 seconds to detect all 19 landmarks on one image.
4) Mirzaalian and Hamarneh's method [29]
   The running time of the classifier to compute 19 appearance models per image is $T_1 = 140 \pm 3$ seconds (using Matlab on intel Xeon 8 cores with 2.3 GHz CPU), and the running time to find the best positions is $T_2 = 139 \pm 2$ seconds. Note that the first step, applying the classifier, is parallelizable, e.g. it can be reduced down to 7 ($\approx$ 140/19) seconds, which means that the computation time for detecting all 19 landmarks takes

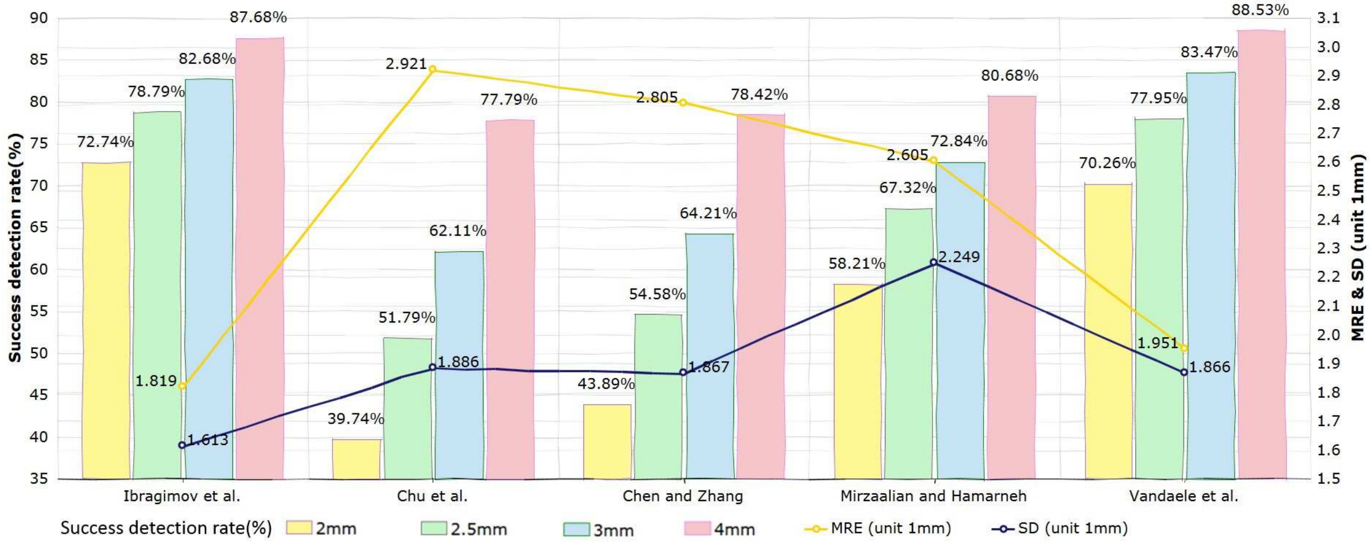


Fig. 8. Mean radial errors (MREs), standard deviations (SDs) and success detection rates (SDRs) using four precision ranges, including 2.0 mm (yellow), 2.5 mm (green), 3.0 mm (blue) and 4.0 mm (red), of five submitted methods on Test1 data.

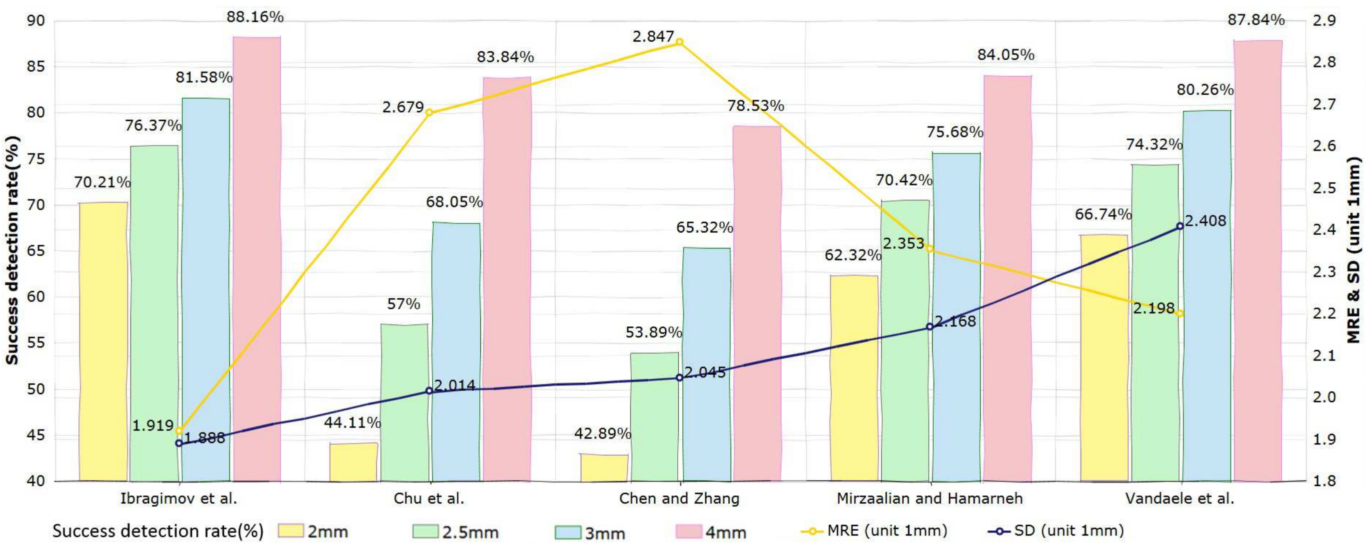


Fig. 9. Mean radial errors (MREs), standard deviations (SDs) and success detection rates (SDRs) using four precision ranges, including 2.0 mm (yellow), 2.5 mm (green), 3.0 mm (blue) and 4.0 mm (red), of five submitted methods on Test2 data.

on average 7+139 = 146 seconds.

5) Vandaele et al.'s method [30]
Vandaele et al. use the implementation of the extremely randomized trees method in scikit-learn [40] and their own Python code for pixel and feature computation. As an illustration, on Intel Core i7-3630QM 2.4 GHz $\times$ 8 and 12 GB RAM with OS Fedora 19, the detection of all 19 landmarks in a single image takes on average 42 seconds (with unoptimized code).

D. Analysis of advantages and drawbacks of all methods

The advantages and drawbacks of all methods are summarized in Table II. From the results of Section V-A and Section V-B, Ibragimov et al.'s method shows the best performance in this challenge. The crucial difference that makes Ibragimov et al.'s framework superior to other competitors is the shape model. To confirm this assumption, Ibragimov et al. compared the framework performance with and without the shape model (see Table III). As shown in Table III, the results indicate that the usage of the proposed shape model reduces the detection error from 3.65 mm to 1.82 mm, i.e. more than 50% of improvement, measured in terms of MRE. Similarly, considerable improvement in landmark detection accuracy is also observed for SDR with respect to all detection precision ranges.

E. Discussion

To analyze the capabilities of methods in detection of individual landmarks, Figure 10 and Figure 11 compare the MRE values between five teams on individual landmarks using Test1

TABLE II

SUMMARY OF THE ADVANTAGES AND DRAWBACKS OF ALL METHODS

Method	Advantages	Drawbacks
Ibragimov et al. [26]	• Computationally efficient and accurate intensity model, strong shape model.	• Intensity model is based on local landmark appearance.
Chu et al. [27]	• A fully automatic method for landmark detection without any user intervention. • Combining local likelihood calculated by random forest regressors and global shape information encoded by the sparse shape composition model to handle outliers.	• Computational complexity increases with the number of patches. • Only considers global shape information without the pair-wised landmark dependence, leading to less accurate results.
Chen and Zheng [28]	• Considers the geometric relation between different patches. • Convex optimization formulation which tends to find the global optimum.	• Computational complexity increases with the number of patches.
Mirzaalian and Hamarneh [29]	• Fast: the running time is a linear function of possible locations of landmarks. • Regularized: the pairwise regularization term encourages detecting landmarks which conform to a learnt graphical model. • Globally optimized: the objective function is globally optimized using the pictorial structure algorithm.	• Needs to be trained: it requires training data to set the hyperparameters. • Not invariant to scale and rotation: the landmark detection objective function is not sensitive to small variations in scale and orientation.
Vandaele et al. [30]	• Extracted features (i.e. raw pixels) are very simple, making both the extraction and the detection fast. • Because of the introduction of translation parameters, the detection at a given pixel can be done on the basis of nearby structures. This is helpful when the landmark is positioned in a low variance region. • The extraction of features at multiple resolutions allows to better grasp the context of the landmark position, and thus avoid errors due to pixel neighborhood similarities at the same resolution. • Landmarks are treated independently of each other, without relying on the whole landmark variability or requiring additional landmarks. It is therefore easy to add a new landmark to the system.	• The method depends on several parameters that need to be tuned by cross-validation. • Each pixel is described by a large number of features ($6\times16\times16$) and each landmark requires to train a separate tree ensemble from a large number of training examples (50000). • Performing the detection of landmarks one by one does not allow to take full advantage of the relations that might exist between them.

TABLE III

COMPARISON OF IBRAGIMOV ET AL.'S FRAMEWORK PERFORMANCE WITH AND WITHOUT THE SHAPE MODEL MEASURED IN TERMS OF THE MEAN RADIAL ERROR (MRE) AND THE SUCCESS DETECTION RATE (SDR). THE RESULTS WERE OBTAINED ON TEST1 DATABASE.

	MRE (mm)	SDR (%)			
		2.0 mm	2.5 mm	3.0 mm	4.0 mm
No shape model	3.65±5.71	66.7	71.6	75.4	79.6
Ibragimov et al. proposed shape model	1.82±1.61	72.7	78.8	82.7	87.7

TABLE IV

INTRA-OBSERVER VARIABILITY OF MANUAL MARKING OF THE LANDMARK 4, 10 AND 19 ON TEST1 DATA AND TEST2 DATA, REPORTED IN TERMS OF MEAN RADIAL ERROR (MRE) ± STANDARD DEVIATION (SD)

MRE (mm)	**Landmark 4**		**Landmark 10**		**Landmark 19**	
	Expert 1	Expert 2	Expert 1	Expert 2	Expert 1	Expert 2
Test1 Data	4.50±2.37	1.83±2.05	2.28±1.82	0.9±0.85	3.01±2.98	0.78±1.13
Test2 Data	2.47±1.43	1.61±1.79	2.60±1.61	0.11±0.89	2.24±2.39	0.85±1.58

data and Test2 data, respectively. It is observed that the method of Ibragimov et al. generally performs best, but among the 19 landmarks, MREs for few landmarks, including the landmark 4 (porion), 10 (gonion) and 19 (articulate), are especially high even for the best performing method (Figure 12). The reason why these landmarks are difficult to be identified precisely may be due to image complexity caused by the difference of X-ray projection between the left and right side of the head structure. As discussed by Kwon et al. [45], there exists a face asymmetry problem because it is common that the left side of the head structure is not identical to the right side. Furthermore, as the data was randomly extracted from the clinical database at the Tri-Service General Hospital in Taipei, the variations on the data with respect to patients' age, gender and image quality are large. All collected data was then randomly divided into three image sets, including training, Test1 and Test2. Using the ground truth data generated by experienced medical experts (see Table IV), it was found that the intra-observer variability of the landmark 4 is larger than the landmark 10 and 19, and the intra-observer variability of the landmark 4 on Test1 data is larger than on Test2 data. These results suggest that even for medical experts marking the landmark 4 is more challenging than marking the landmark 10 and 19, and marking the landmark 4 on Test1 data is more difficult than on Test2 data.

VI. CONCLUSIONS

This paper presents a representative selection of current methods which were submitted to the Automatic Cephalometric X-Ray Landmark Detection Challenge held at ISBI 2014. The 300 cephalometric X-ray images with resolution 1935 by 2400 pixels were provided by the Tri-Service General Hospital in Taipei, Taiwan (IRB Number 1-102-05-017). Regarding the ground truth data for evaluation, 19 landmarks were manually marked in each image and reviewed by two experienced medical doctors. A total of six teams submitted their results in Stage 1, and five teams were accepted to attend Stage 2 (on-site competition).

The quantitative results of the automatic cephalometric X-ray landmark detection challenge show that overall (combining the results of Test1 and Test2 data) SDR for 2mm, 2.5mm, 3mm and 4mm precision ranges for the best performing teams are 71.48% (by Ibragimov et al.), 77.58% (by Ibragimov et al.), 82.13% (by Ibragimov et al.) and 88.19% (by Vandaele et al.), respectively. In clinical diagnosis, 2mm is the acceptable precision range [18], and the results show that this is still a challenging problem as only 71.48% a SDR in the 2mm

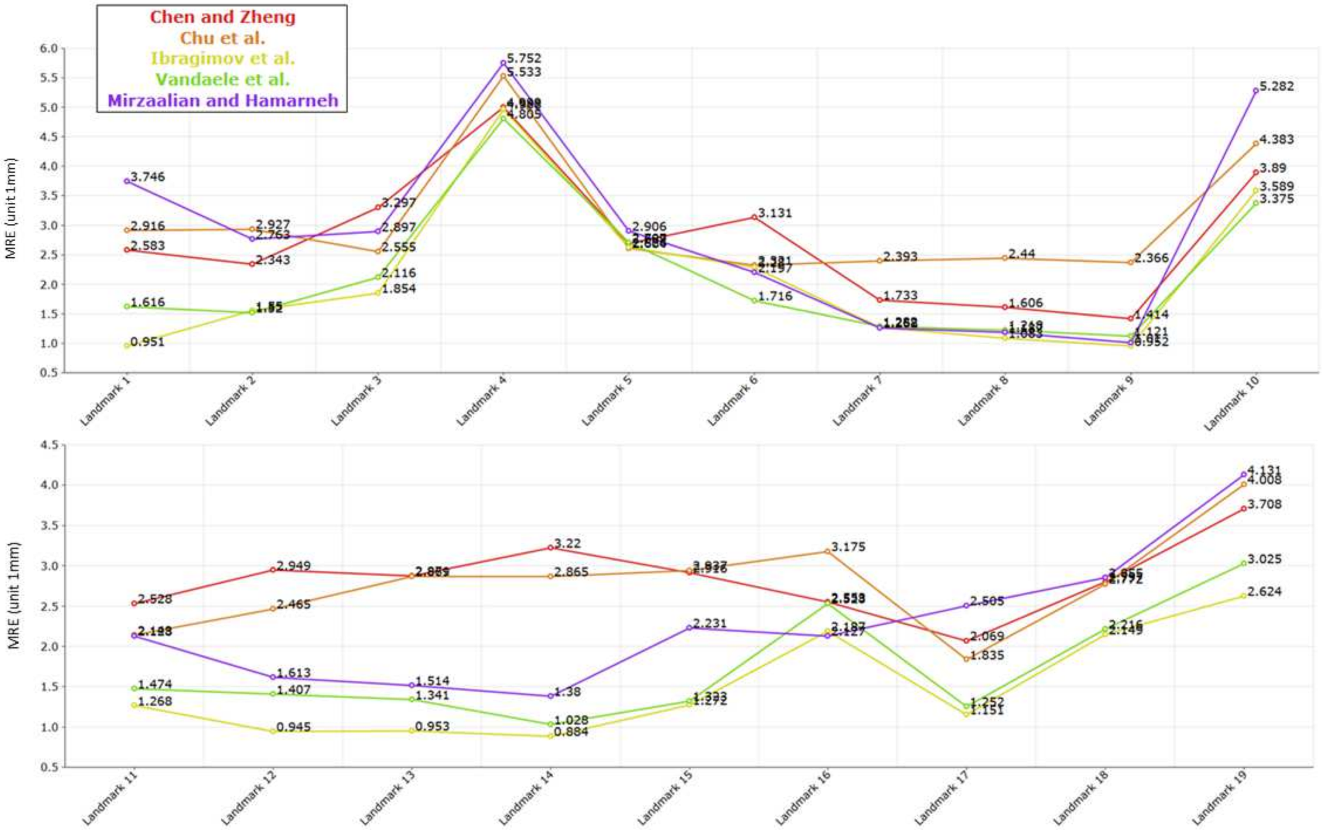


Fig. 10. Mean radial errors (MREs) with 1 mm as the measurement unit in detection of individual landmarks on Test1 data. MREs for Landmark 4, Landmark 10 and Landmark 19 are particularly high, showing that these landmarks are especially difficult for detection.

precision range is achieved by the best performing team. In addition, we found that detection of landmarks 4 (porion), 10 (gonion) and 19 (articulate) is especially difficult. In conclusion, this is still an unsolved and challenging task and needs further investigation.

Automated methods save time and manual costs, and avoid problems caused by intra- and inter-observer variations or errors due to fatigue. Regarding the processing time, detecting 19 landmarks by the best method only takes 18 seconds. In comparison, it generally takes up to 20 minutes to perform one X-ray cephalometric analysis by an experienced orthodontist. In the future, we expect that more sophisticated methods will be developed and higher success detection rates will be obtained.

ACKNOWLEDGMENT

This work was jointly supported by Tri-Service General Hospital-National Taiwan University of Science and Technology Joint Research Program (TSGH-NTUST-103-02 and TSGH-NTUST-102-02) and Ministry of Science and Technology under a grant (MOST-101-2628-E-011-006-MY3). The work of Ibargimov and Vrtovec was supported by the Slovenian Research Agency, under grants (P2-0232 and L2-4072). Rémy Vandaele is funded by a FNRS-Télévie grant in Belgium. Raphaël Marée is supported by Cytomine research grant n°1017072 of the Wallonia (Belgium).

REFERENCES

[1] Bernard Deberry-Borowiecki, Andrzej Kukwa, and Robert H. I. Blanks, "Cephalometric analysis for diagnosis and treatment of obstructive sleep apnea," The Laryngoscope, vol. 998, no. 2, pp. 226-234, 1988.

[2] C. Guilleminault, M.A. Quera-Salva, M. Partinen, and A. Jamieson, "Women and the obstructive sleep apnea syndrome," CHEST, vol. 93, no. 1, pp. 104-109, 1988.

[3] W. Hochban and U. Brandenburg, "Morphology of the viscerocranium in obstructive sleep apnoea syndrome-cephalometric evaluation of 400 patients," Journal of Cranio-Maxillo-Facial Surgery, vol. 22, no. 4, pp. 205-213, 1994.

[4] R.S. Nanda and R.M. Merrill, "Cephalometric assessment of sagittal relationship between maxilla and mandible," American Journal of Orthodontics and Dentofacial Orthopedics, vol. 105, no. 4, pp. 328-344, 1994.

[5] S.L. Lake, R.W. McNeill, R.M. Little, and R.A. West, "Surgical mandibular advancement: a cephalometric analysis of treatment response,"American Journal of Orthodontics, vol. 80, no. 4, pp. 376-394, 1981.

[6] A.K. Melnik, "A cephalometric study of mandibular asymmetry in a longitudinally followed sample of growing children,"American Journal of Orthodontics and Dentofacial Orthopedics, vol. 101, no. 4, pp. 355-366, 1992.

[7] M.Y. Hajeer, A.F. Ayoub, and D.T. Millett, "Three-dimensional assessment of facial soft-tissue asymmetry before and after orthognathic surgery," British Journal of Oral and Maxillofacial Surgery, vol. 42, no. 5, pp. 396-404, 2004.

[8] J.P. McDonnell, R.W. McNeill, and R.A. West, "Advancement genioplasty: a retrospective cephalometric analysis of osseous and soft tissue changes," Journal of oral surgery, vol. 35, no. 8, pp. 640-647, 1977.

[9] A.E. de Oliveira, L.H. Cevidanes , C. Phillips, A. Motta, B. Burke, and D. Tyndall, "Observer reliability of three-dimensional cephalometric landmark identification on cone-beam computerized tomography," Oral Surgery, Oral Medicine, Oral Pathology, Oral Radiology, vol. 107, no. 2, pp. 256-265, 2009.

[10] P.W. Major, D.E. Johnson, K.L. Hesse, and K.E. Glover, "Effect of head

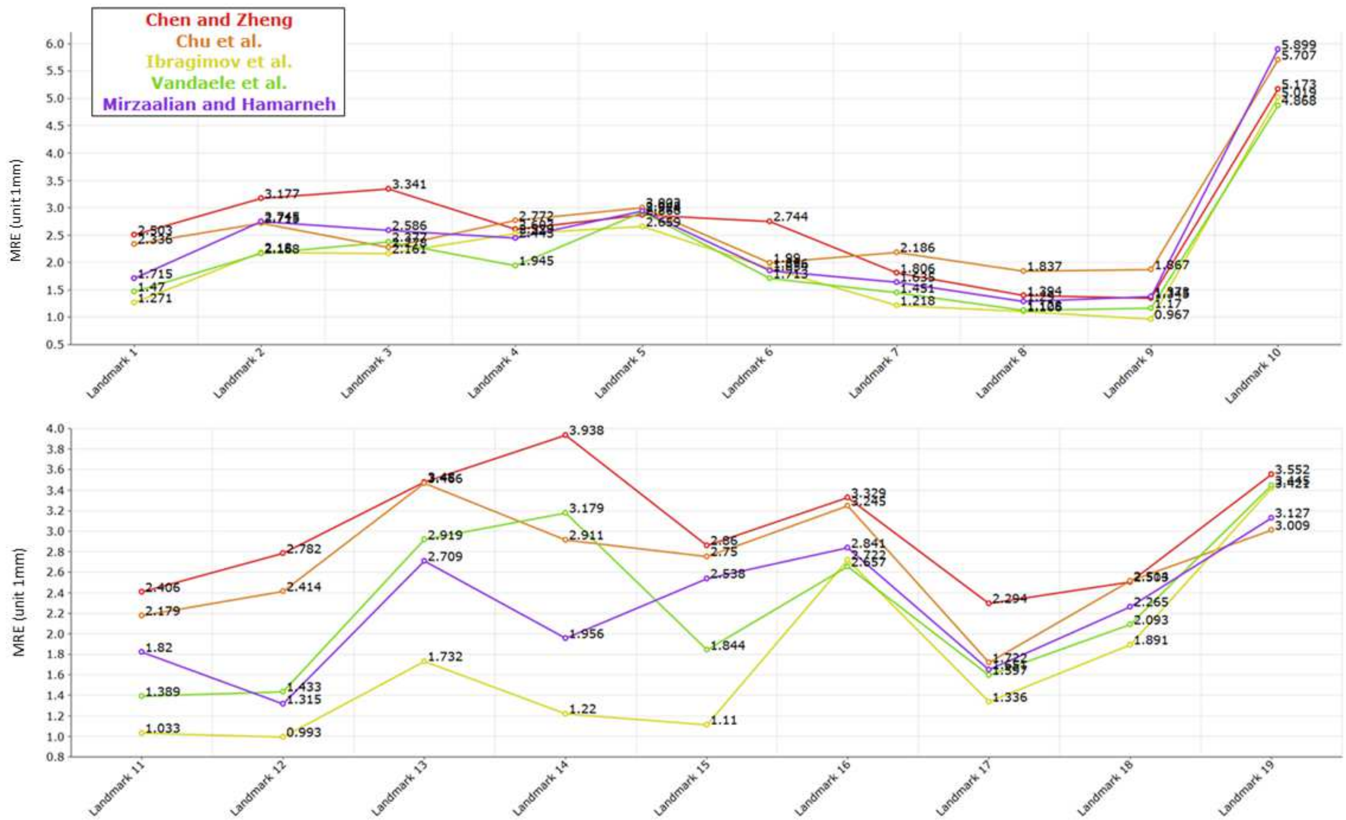


Fig. 11. Mean radial errors (MREs) with 1 mm as the measurement unit in detection of individual landmarks on Test2 data. MREs for Landmark 10 and Landmark 19 are particularly high, showing that these landmarks are especially difficult for detection.

orientation on posterior anterior cephalometric landmark identification," The angle orthodontist, vol. 66, no. 1, pp. 51-60, 1996.

[11] Y.J. Chen, S.K. Chen, H.F. Chang, and K.C. Chen, "Comparison of landmark identification in traditional versus computer-aided digital cephalometry," The angle orthodontist, vol. 70, no. 5, pp. 387-392, 2000.

[12] T. Rakosi, "An Atlas and Manual of Cephalometric Radiology," Wolfe Medical Publications, London, 1982.

[13] I. El-Feghi, M.A. Sid-Ahmed, and M. Ahmadi, "Automatic localization of craniofacial landmarks for assisted cephalometry," Pattern Recognition, vol. 37, no. 3, pp. 609-621, 2004.

[14] A. Kamoen, L. Dermaut, and R. Verbeeck, "The clinical significance of error measurement in the interpretation of treatment results," European Journal of Orthodontics, vo. 23, no. 5, pp. 569-578, 2001.

[15] A. Kaur and C. Singh, "Cephalometric X-Ray Registration using Angular Radial Transform," IJCA Proceedings on International Conference on Recent Advances and Future Trends in Information Technology (iRAFIT 2012), iRAFIT(9), pp. 18-22, 2012.

[16] S.R. McClure, P.L. Sadowsky, A. Ferreira and A. Jacobson, "Reliability of Digital Versus Conventional Cephalometric Radiology: A Comparative Evaluation of Landmark Identification Error," Seminars in Orthodontics, vol. 11, no. 2, pp. 98-110, 2005.

[17] E.L. Delamare, G.S. Liedke, M.B. Vizzotto, H.L.D. da Silveira, J.L.D. Ribeiro and H.E.D. Silveira, "Influence of a programme of professional calibration in the variability of landmark identification using cone beam computed tomography-synthesized and conventional radiographic cephalograms," Dentomaxillofac Radiol., vol. 39, no. 7, pp. 414-423, 2010.

[18] W. Yue, D. Yin , C. Li , G. Wang , and T. Xu, "Automated 2-D Cephalometric Analysis on X-ray Images by a Model-based Approach," IEEE Transactions on Biomedical Engineering, vol. 53, no. 8, pp. 1615V1623, 2006.

[19] R. Kafieh, S. Sadri, A. Mehri, and H. Raji, "Discrimination of Bony Structures in Cephalograms for Automatic Landmark Detection," Communications in Computer and Information Science, vol. 6, no. 1, pp. 609-620, 2009.

[20] A. Kaur and C. Singh, "Automatic Cephalometric Landmark Detection Using Zernike Moments and Template Matching," Signal, Image and Video Processing, vol.9, no. 1, pp.117-132, 2015.

[21] J.T.L. Ferreira and C.D.S. Telles, "Evaluation of the Reliability of Computerized Profile Cephalometric Analysis," Brazilian Dental Journal, vol. 13, no. 3, pp. 201-204, 2002.

[22] C.W. Wang, C.T. Huang, C.H. Li, and S.W. Chang, "A Grand Challenge for Automated Detection of Critical Landmarks for Cephalometric X-Ray Image Analysis," IEEE International Symposium on Biomedical Imaging, 2014.

[23] P. Vucinic, Z. Trpovski, and I. Scepan, "Automatic landmarking of cephalograms using active appearance models," European Journal of Orthodontics, vol. 32, no. 1, pp. 233-241, 2010.

[24] B. Ibragimov, B. Likar, F. Pernuš, and T. Vrtovec, "Shape representation for efficient landmark-based segmentation in 3D," *IEEE Transactions on Medical Imaging*, vol. 33, no. 4, pp. 861-874, 2014.

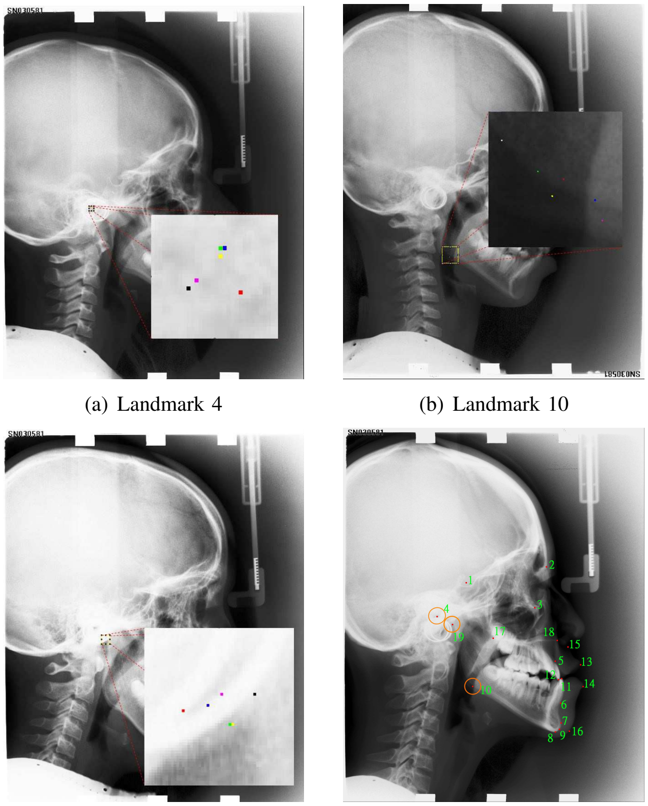
[25] B. Ibragimov, B. Likar, F. Pernuš, and T. Vrtovec, "A game-theoretic framework for landmark-based image segmentation," *IEEE Transactions on Medical Imaging*, vol. 31, no. 9, pp. 1761-1776, 2012.

[26] B. Ibragimov, B. Likar, F. Pernuš and T. Vrtovec, "Automatic cephalometric X-ray landmark detection by applying game theory and random forests," in *Proc. ISBI International Symposium on Biomedical Imaging 2014: Automatic Cephalometric X-Ray Landmark Detection Challenge*, 2014.

[27] C. Chu, C. Chen, L.-P. Nolte and G. Zheng, "Fully Automatic Cephalometric X-Ray Landmark Detection Using Random Forest Regression and Sparse shape composition," in *Proc. ISBI International Symposium on Biomedical Imaging 2014: Automatic Cephalometric X-Ray Landmark Detection Challenge*, 2014.

[28] C. Chen and G. Zheng, "Fully-Automatic Landmark Detection in Cephalometric X-Ray Images by Data-Driven Image Displacement Estimation," in *Proc. ISBI International Symposium on Biomedical Imaging 2014: Automatic Cephalometric X-Ray Landmark Detection Challenge*, 2014.

[29] H. Mirzaalian and G. Hamarneh, "Automatic Globally-Optimal Pictorial Structures with Random Decision Forest Based Likelihoods For Cephalometric X-Ray Landmark Detection," in *Proc. ISBI International*



(a) Landmark 4 (b) Landmark 10

(c) Landmark 19 (d) Overview

Fig. 12. The detection outputs by the submitted methods and the reference landmark for three challenging landmarks, including (a) Landmark 4, (b) Landmark 10 and (c) Landmark 19. (d) A cephalogram highlighted with the three challenging landmarks using orange circles. Reference landmarks are highlighted with black color in (a,c) and white color in (b); outputs of methods of Ibragimov et al., Chu et al., Chen and Zheng, Mirzaalian and Hamarneh and Vandaele et al. are marked in yellow, blue, red, magenta and green, respectively.

Symposium on Biomedical Imaging 2014: Automatic Cephalometric X-Ray Landmark Detection Challenge, 2014.

[30] R. Vandaele, R. Marée, S. Jodogne, and P. Geurts, "Automatic Cephalometric X-Ray Landmark Detection Challenge 2014: A tree-based approach," in *Proc. ISBI International Symposium on Biomedical Imaging 2014: Automatic Cephalometric X-Ray Landmark Detection Challenge*, 2014.

[31] L. Breiman, Random forests, Machine Learning, vol. 45, no. 1, pp. 5-32, 2001

[32] A. Criminisi, J. Shotton, D. Robertson, E. Konukoglu, "Regression forests for efficient anatomy detection and localization in CT studies," in Proc. MICCAI 2010 International Conference on Medical Image Computing and Computer Assisted Intervention, Lecture Notes in Computer Science, vol. 6533, no.1, pp. 106-117, 2011.

[33] C. Lindner, S. Thiagarajah, J. M. Wilkinson, arcOGEN Consortium, G. Wallis, T. F. Cootes, "Fully automatic segmentation of the proximal femur using random forest regression voting," IEEE Trans Med Imaging, vol. 32, no. 8, pp. 1462-1472, 2013.

[34] S. Zhang, Y. Zhan, M. Dewan, J. Huang, D. N. Metaxas, X. S. Zhou, "Sparse shape composition: a new framework for shape prior modeling," in *Proc. CVPR IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1025-1032, 2011.

[35] C. Chen, W. Xie, J. Franke, P.A. Grutzner, L.-P. Nolte, G. Zheng, "Automatic X-ray landmark detection and shape segmentation via data-driven joint estimation of image displacements," Medical Image Analysis, vol. 18, no. 3, pp. 487-499, 2014.

[36] T. Ojala, M. Pietikinen, and D. Harwood, "Performance evaluation of texture measures with classification based on Kullback discrimination of distributions," in Proc. of ICPR International Conference on Pattern Recognition (ICPR), vol. 1, pp. 582-585, 1994.

[37] A. Frangi, W. J. Niessen, L. Vincken, and A. Viergever, "Multiscale vessel enhancement filtering," in Proc. MICCAI 1998 International Conference on Medical Image Computing and Computer Assisted Intervention, vol. 1, pp. 130-137, 1998.

[38] P. Felzenszwalb and D. Huttenlocher, "Pictorial structures for object recognition," International Journal of Computer Vision, vol. 61, no. 1, pp. 55-79, 2005.

[39] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," Machine learning, vol. 63, no. 1, pp.3-42, 2006.

[40] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825-2830, 2011.

[41] O. Stern, R. Marée, J. Aceto, N. Jeanray, M. Muller, L. Wehenkel, and P. Geurts, "Automatic localization of interest points in zebrafish images with tree-based methods," Pattern Recognition in Bioinformatics, vol. 7036, no. 1, pp. 179-190, 2011.

[42] L. Seres, E. Jr. Varga, A. Kocsis, Z. Rasko, B. Bago, E. Varga, and J. Piffko, "Correction of a severe facial asymmetry with computerized planning and with the use of a rapid prototyped surgical template: a case report/technique article," Head Face Med., vol. 10, no. 27, doi: 10.1186/1746-160X-10-27, 2014.

[43] J. Kragskov, C. Bosch, C. Gyldensted, and S. Sindet-Pedersen, "Comparison of the reliability of craniofacial anatomic landmarks based on cephalometric radiographs and three-dimensional CT scans," Cleft Palate-Craniofacial Journal, vol. 34, no. 2, pp. 111-116, 1997.

[44] M.J. Gliddon, J.J. Xia, J. Gateno, H.T. Wong, R.E. Lasky, J.F. Teichgraeber, X. Jia, M.A. Liebschner, and J.J. Lemoine, "The accuracy of cephalometric tracing superimposition," Journal of Oral and Maxillofacial Surgery, vol. 64, no. 2, pp. 194-202, 2006.

[45] T.G. Kwon, H.S. Park, H.M. Ryoo, and S.H. Lee, "A comparison of craniofacial morphology in patients with and without facial asymmetry-a three-dimensional analysis with computed tomography," Journal of Oral and Maxillofacial Surgery, vol. 35, no. 1, pp. 43-48, 2006.

[46] W. Bholsithi, W. Tharanon, K. Chintakanon, R. Komolpis, and C. Sinthanayothin, "3D vs. 2D cephalometric analysis comparisons with repeated measurements from 20 Thai males and 20 Thai females," Biomedical Imaging and Intervention Journal, vol. 5, no. 4, doi. 10.2349/biij.5.4.e21., 2009.

[47] A.L. Pereira, L.M. De-Marchi, P.C. Scheibel, and A.L. Ramos, "Reproducibility of natural head position in profile photographs of children aged 8 to 12 years with and without the aid of a cephalostat," Dental Press Journal of Orthodontics, vol. 15, no. 1, pp. 65-73, 2010.