LETTER TO THE EDITOR

On automatic landmarking

Dentomaxillofacial Radiology (2014) 43, 20130399. doi: 10.1259/dmfr.20130399

Cite this article as: Rakhshan V. On automatic landmarking. *Dentomaxillofac Radiol* 2014; 43: 20130399.

I read with interest the intriguing article of Shahidi et al,¹ which described in a good report the production of a rather effective computer program designed to identify lateral cephalometric landmarks automatically. The authors also tested its accuracy. The article was of good scientific merit. There were numerous positive points that have already been described by the authors in their article. However, some points remained unaddressed in their report. I would much appreciate it if the authors could possibly elaborate on the following.

Methods

Software development

- (1) How knowledge-based approaches (without any edge-detection algorithm) were used to locate the landmarks gnathion, pogonion, A point, anterior nasal spine, B point, nasion and the tip of the upper first incisor, even though these landmarks are all defined on clear contours (edges) and can benefit considerably from edge-enhancement and edge-detection algorithms?
- (2) The knowledge- or template-based approaches need some previously computed (or manually entered) reference points, to be able to work. However, the report lacks any such reference points. How could such referenceless knowledge- and template-based methods even work, let alone be accurate?
- (3) Regarding the sentence: "An inclination-finding mechanism, based on the edge enhancement approach...". Common sense indicates that landmarks are first located by easier and more accurate edge-detection algorithms. If this is not possible, other methods (that are usually more sophisticated or difficult and sometimes less accurate) are used. However, in this study, first, the other algorithms have been used, and then, for the rest of landmarks, the edge-enhancement approach

(which is a part of edge detection) has been performed. Perhaps a clarification might be of value.

- (4) Also it would be better if the authors could elaborate on their reason to completely separate the edge-enhancing and edge-detection algorithms. To my knowledge, these are related algorithms and are usually used together.
- (5) Additionally, it might help if the authors could elaborate on the "inclination finding" mechanism. A reader might be interested to know what "inclination" means here. Why should one determine it? How does it work? and Which landmarks benefited from it?

Software testing

- (6) Page 4: I wonder what exactly were the definitions of "thresholds of the points at the ends of lines" and "degree thresholds" and "millimetre thresholds"? Also, what was their role? There were some general and vague explanations on their use ("In order to make it easy for our results to be compared with other studies and also the unity of the reports"), but it was not explained how these "thresholds" could help the unity of results? Also the definitions of P1 and P2 and x are missing, both in the Figure 1 legend and in the formula. What are these?
- (7) My understanding from the second paragraph on page 4 was that the inclinations of four different lines were "also" computed *instead* of landmarks? Could the authors please elaborate on this issue? Have they not found 16 landmarks, but 12 landmarks and 4 lines? Or what else?
- (8) A critical issue not addressed in the article was that the authors have used the *same set of radiographs* for both (A) "developing/optimizing their program" and (B) "testing the same program". Testing a program using the very image that was used originally to optimize the program can falsely increase (and bias) the observed accuracy of the software and reduce the generalizability of the results. It is not known whether the software could work with any randomly given new cephalograph. Besides, image outputs of different radiography

Correspondence to: Vahid Rakhshan, E-mail: vahid.rakhshan@gmail.com

Received 1 November 2013; revised 15 December 2013; accepted 18 December 2013

Table 1. A modified table 5 of Shahidi et al¹ now including the results of one-sample *t*-test, comparing each mean error with the constant error values 1 and 2 mm (computed by this author)

Landmarks	Mean error (mm)	SD (mm)	p-value (1 mm)	p-value (2 mm)
S	1.71	2.96	0.137	0.539
N	2.34	3.49	0.020	0.541
Or	2.74	2.84	0.000	0.107
A	2.49	3.29	0.007	0.352
B	3.33	4.21	0.001	0.053
ANS	4.44	2.58	0.000	0.000
PNS	1.99	2.26	0.009	0.978
Pog	2.48	4.46	0.042	0.500
Me	1.72	4.26	0.292	0.680
Gn	3.02	4.68	0.010	0.176
UIT	2.25	2.89	0.009	0.587
Total	2.59	3.45	—	—

A, A point; ANS, anterior nasal spine; B, B point; Gn, gnathion; Me, menton; N, nasion; Or, orbitale; PNS, posterior nasal spine; Pog, pogonion; S, sella; SD, standard deviation; UIT, tip of upper first incisor.

devices can differ in many properties, such as the contrast, resolution, greyscale histogram, etc. Moreover, the shapes of cephalostat rulers and the distance of head contours from the top and the right borders of the images can differ in various devices. These factors might confound the knowledge-based algorithms. So, it might be possible that using a cephalograph taken with another device from the same patient probably leads to different results; and by testing the software on the “training image set”, the effects of these limitations could be excluded, positively biasing the results.

- (9) The authors stated that their sample was random. However, I believe, owing to the above-mentioned limitations, this sample might not be quite random.

Results

Statistical analyses

- (10) Table 3: I could not comprehend the definition of “equivalent mean error (mm)” and why the authors have compared “mean error (degree)” with “equivalent mean error (mm)”? What did this statistical comparison imply?
- (11) Table 4: It is not stated that the *p*-value is calculated by comparing the error value with which value? What was the purpose of the comparison? Which statistical test was used?
- (12) Table 5: A major scientific concern might be that there are no statistical analyses for the comparisons of landmarks with the errors 1, 2 and 4 mm, whereas the Results and Discussion sections heavily rely on those *unsubstantiated* comparisons. This can undermine their values. Without *statistical substantiation*, any raw comparison lacks internal validity (*i.e.* is not reliable). The comparisons could be of value if validated, as landmarking errors <1 mm can be considered ideal, and those <2 mm can be regarded as clinically acceptable.^{1–4} However, it is not known (using statistical validation) whether any of those results reported by Shahidi et al¹ as <2 mm error were only <2 mm because of chance.

Moreover, extensively discussing these non-validated results (without addressing this crucial limitation) might be misleading, and such results should have been approached with more caution. I suggest that the authors statistically compare each mean error with the constant values 1 and 2 mm, using a one-sample *t*-test,² to verify their results statistically. They could also use other approaches such as equivalence testing.⁵ I computed the one-sample *t*-test’s *p*-values for the mean errors reported in table 5 (Table 1 of this letter). The analyses indicated that all their results <2 mm were actually non-significant (and thus not necessarily verified). I think a revision might favour the text.

Potential inconsistencies

- (13) In the Results section (page 5), it is stated that seven landmarks had errors <2 mm (sella, gonion, tip of lower first incisor, apex of lower first incisor, menton, porion and tip of upper first incisor). Also, two landmarks (apex of lower first incisor and tip of lower first incisor) had errors <1 mm. However, in table 5, only three landmarks had errors <2 mm and none were <1 mm. I would appreciate it if the authors could clarify the reason for this difference between the text and table 5.
- (14) Table 5: Landmarks apex of upper first incisor, apex of lower first incisor and porion are mentioned in the table legend but not in the table itself. Also the authors had identified 16 landmarks, not 11 landmarks as stated in table 5, and not even 14 landmarks as mentioned in its legend. So, what happened to the remaining five landmarks?

V Rakhshan
Department of Dental Anatomy and Morphology
Dental Branch
Islamic Azad University
Tehran
Islamic Republic of Iran

References

1. Shahidi S, Oshagh M, Gozin F, Salehi P, Danaei SM. Accuracy of computerized automatic identification of cephalometric landmarks by a designed software. *Dentomaxillofac Radiol* 2013; **42**: 20110187. doi: [10.1259/dmfr.20110187](https://doi.org/10.1259/dmfr.20110187)
2. Rakhshan V, Rakhshan H, Sheibaninia A. Developing an automatic lateral cephalometric landmark identification program and evaluating its performance. *Int J Comput Dent* 2009; **12**: 327–43.
3. Rakosi T. The introduction of cephalometry to orthodontics. In: Rakosi T, ed. *An atlas of manual cephalometric radiography*. London, UK: Wolfe Medical Publications; 1982. p. 7–19.
4. Rudolph DJ, Sinclair PM, Coggins JM. Automatic computerized radiographic identification of cephalometric landmarks. *Am J Orthod Dentofacial Orthop* 1998; **113**: 173–9.
5. Kazemi M, Geramipناه F, Negahdari R, Rakhshan V. Active tactile sensibility of single-tooth implants versus natural dentition: a split-mouth double-blind randomized clinical trial. *Clin Implant Dent Relat Res March* 2013. Epub ahead of print. doi: [10.1111/cid.12053](https://doi.org/10.1111/cid.12053)