Cephalometric Landmark Detection in Lateral X-ray Images: Structured description of the challenge design

CHALLENGE ORGANIZATION

Title

Use the title to convey the essential information on the challenge mission.

Cephalometric Landmark Detection in Lateral X-ray Images

Challenge acronym

Preferable, provide a short acronym of the challenge (if any).

CL-Detection2024

Challenge abstract

Provide a summary of the challenge purpose. This should include a general introduction in the topic from both a biomedical as well as from a technical point of view and clearly state the envisioned technical and/or biomedical impact of the challenge.

Cephalometric analysis is a fundamental examination which is routinely used in fields of orthodontics and orthognathics. The key operation in the analysis is marking craniofacial landmarks from lateral cephalograms. These landmarks serve as the datum of the succeeding qualitative assessment of angles and distances, which provide diagnosis information of the craniofacial condition of a patient and affect treatment planning decision. However, reliable landmark annotations often require experienced doctors, and even for seasoned orthodontists, manually identifying these landmarks can be a time-consuming and labor-intensive process. Hence, fully automatic and accurate landmark localization has been a long-standing area with a great deal of need.

In this challenge, we aim to promote the development of universal cephalometric landmark detection in lateral X-ray images. This is an extension of CL-Detection2023 challenge. In CL-Detection2023, the challenge task is to detect 38 landmarks. In CL-Detection2024, we will add the cervical vertebrae and ruler landmark detection task, which can use to examine the skeletal growth of a patient and assess the magnification errors during the scanning process, respectively. Specifically, the detection algorithm is expected to accurately locate 53 landmarks (13 soft tissue-related landmarks, 6 tooth-related landmarks, 19 skull-related landmarks, 13 cervical spine-related landmarks, 2 calibration ruler landmarks) in lateral X-ray images. Our challenge aims to provide a comprehensive benchmark for cephalometric landmark detection methods.

Compared to the dataset in CL-Detection2023 (600 lateral X-ray images), in CL-Detection2024, we further increase the dataset with an inclusion of data from a new medical center, resulting in a total of to 700 lateral X-ray images. The all data from this new center will serve as a test set to explore the algorithm domain generalization. To the best knowledge, this will be the most diverse and most landmark annotated public dataset for cephalometric landmark detection.

In addition, based on the results in CL-Detection2023, we found that detection models cannot achieve a good tradeoff between segmentation accuracy and efficiency. Thus, motivated by the FLARE 2022&2023 challenge, in CL-Detection2024, the challenge evaluation criteria are not limited to detection accuracy, but also include two new evaluation metrics: runtime and GPU memory consumption, which provide a comprehensive evaluation of detection accuracy and efficiency.

In summary, the CL-Detection2024 challenge has three main features:
(1) Task: this is the first challenge for cervical vertebrae landmark detection in cephalometric X-ray images.
(2) Dataset: we provide the most diverse and most meticulous lateral X-ray dataset, including 700 2D X-ray images from 4 medical centers.
(3) Evaluation: we not only focus on detection accuracy but also detection efficiency, which are in concordance with real clinical practice and requirements.

Challenge keywords

List the primary keywords that characterize the challenge.challenge_

Cephalometric, Landmark Detection, Lateral X-ray Images, Efficiency

Year

The challenge will take place in 2024

FURTHER INFORMATION FOR CONFERENCE ORGANIZERS

Workshop

If the challenge is part of a workshop, please indicate the workshop.

Advances in Cephalometric Landmark Detection.

Duration

How long does the challenge take?

Half day.

Expected number of participants

Please explain the basis of your estimate (e.g. numbers from previous challenges) and/or provide a list of potential participants and indicate if they have already confirmed their willingness to contribute.

In MICCAI CL-Detection2023, we had more than 300 participants in the challenge. Finally, 37 teams made successful docker submissions during the testing phase. Thus, we expect the 2024 participation to be similar and estimate there will be at least 40 teams in CL-Detection2024.

Publication and future plans

Please indicate if you plan to coordinate a publication of the challenge results.

An overview paper will be written with the organizing team's members. Each participating team who presented their method at the challenge session is allowed two co-authorships.

Space and hardware requirements

Organizers of on-site challenges must provide a fair computing environment for all participants. For instance, algorithms should run on the same computing platform provided to all.

Assuming that MICCAI will take place in person, we need support for basic presentation equipment (e.g. projectors, computers, monitors, loudspeakers, microphones), for solution presentations from the top-ranked teams.

TASK 1: Cephalometric Landmark Detection in Lateral X-ray Images

SUMMARY

Abstract

Provide a summary of the challenge purpose. This should include a general introduction in the topic from both a biomedical as well as from a technical point of view and clearly state the envisioned technical and/or biomedical impact of the challenge.

Cephalometric analysis is a fundamental examination which is routinely used in fields of orthodontics and orthognathics. The key operation in the analysis is marking craniofacial landmarks from lateral cephalograms. These landmarks serve as the datum of the succeeding qualitative assessment of angles and distances, which provide diagnosis information of the craniofacial condition of a patient and affect treatment planning decision. However, reliable landmark annotations often require experienced doctors, and even for seasoned orthodontists, manually identifying these landmarks can be a time-consuming and labor-intensive process. Hence, fully automatic and accurate landmark localization has been a long-standing area with a great deal of need.

In this challenge, we aim to promote the development of universal cephalometric landmark detection in lateral X-ray images. This is an extension of CL-Detection2023 challenge. In CL-Detection2023, the challenge task is to detect 38 landmarks. In CL-Detection2024, we will add the cervical vertebrae and ruler landmark detection task, which can use to examine the skeletal growth of a patient and assess the magnification errors during the scanning process, respectively. Specifically, the detection algorithm is expected to accurately locate 53 landmarks (13 soft tissue-related landmarks, 6 tooth-related landmarks, 19 skull-related landmarks, 13 cervical spine-related landmarks, 2 calibration ruler landmarks) in lateral X-ray images. Our challenge aims to provide a comprehensive benchmark for cephalometric landmark detection methods.

Compared to the dataset in CL-Detection2023 (600 lateral X-ray images), in CL-Detection2024, we further increase the dataset with an inclusion of data from a new medical center, resulting in a total of to 700 lateral X-ray images. The all data from this new center will serve as a test set to explore the algorithm's domain generalization. To the best knowledge, this will be the most diverse and most landmark annotated public dataset for cephalometric landmark detection.

In addition, based on the results in CL-Detection2023, we found that detection models cannot achieve a good tradeoff between segmentation accuracy and efficiency. Thus, motivated by the FLARE 2022&2023 challenge, in CL-Detection2024, the challenge evaluation criteria are not limited to detection accuracy, but also include two new evaluation metrics: runtime and GPU memory consumption, which provide a comprehensive evaluation of detection accuracy and efficiency.

In summary, the CL-Detection2024 challenge has three main features:
(1) Task: this is the first challenge for cervical vertebrae landmark detection in panoramic X-ray images.
(2) Dataset: we provide the most diverse and most meticulous lateral X-ray dataset, including 700 2D X-ray images from 4 medical centers.
(3) Evaluation: we not only focus on detection accuracy but also detection efficiency, which are in concordance with real clinical practice and requirements.

Keywords

List the primary keywords that characterize the task.

Cephalometric, Landmark Detection, Lateral X-ray Images, Efficiency

ORGANIZATION

Organizers

a) Provide information on the organizing team (names and affiliations).

The challenge is organized as a joint effort of different institutions and researchers in Shenzhen University, China; Shenzhen University Hospital, China; the Fifth Affiliated Hospital of Xinjiang Medical University, China. We are mainly divided into two corresponding groups for our challenge.

Medical Group:
- Prof. Jun Cao and Ms. Juan Dai, Department of Stomatology, Shenzhen University General Hospital, Shenzhen University, China.
- Dr. Tao Guo, Dr. Juncheng Chen and Dr. Sen Zuo, Department of Stomatology, the Fifth Affiliated Hospital of Xinjiang Medical University, China.

Technical Group:
- Prof. Bingsheng Huang, Dr. Hongyuan Zhang, Ms. Haoyu Xie and Dr. Haipeng, Wang, Medical AI Lab, School of Biomedical Engineering, Shenzhen University Medical School, Shenzhen University, China.

b) Provide information on the primary contact person.

Bingsheng Huang, Medical AI Lab, School of Biomedical Engineering, Shenzhen University Medical School, Shenzhen University, China.
Email: huangb@szu.edu.cn

Hongyuan Zhang, Medical AI Lab, School of Biomedical Engineering, Shenzhen University Medical School, Shenzhen University, China.
Email: zhanghongyuan2017@email.szu.edu.cn

Life cycle type

Define the intended submission cycle of the challenge. Include information on whether/how the challenge will be continued after the challenge has taken place.Not every challenge closes after the submission deadline (one-time event). Sometimes it is possible to submit results after the deadline (open call) or the challenge is repeated with some modifications (repeated event).

Examples:

- One-time event with fixed conference submission deadline
- Open call (challenge opens for new submissions after conference deadline)
- Repeated event with annual fixed conference submission deadline

One-time event with fixed conference submission deadline.

Challenge venue and platform

a) Report the event (e.g. conference) that is associated with the challenge (if any).

MICCAI

b) Report the platform (e.g. grand-challenge.org) used to run the challenge.

https://competitions.codalab.org/competitions/

c) Provide the URL for the challenge website (if any).

To be announced.

Participation policies

a) Define the allowed user interaction of the algorithms assessed (e.g. only (semi-) automatic methods allowed).

Fully automatic.

b) Define the policy on the usage of training data. The data used to train algorithms may, for example, be restricted to the data provided by the challenge or to publicly available data including (open) pre-trained nets.

No additional data and pre-trained models are allowed.

c) Define the participation policy for members of the organizers' institutes. For example, members of the organizers' institutes may participate in the challenge but are not eligible for awards.

Members of the organizers' institutes that are associated with the challenge can participate but are not eligible for awards.

d) Define the award policy. In particular, provide details with respect to challenge prizes.

We will provide a prize (cash and certificate) for the top three teams. The exact money of cash prize is not known yet. In addition, we are actively seeking sponsorship. The details of this will be announced on our challenge homepage at a later date.

e) Define the policy for result announcement.

Examples:

- Top 3 performing methods will be announced publicly.
- Participating teams can choose whether the performance results will be made public.

1) All results will be announced publicly through a public leaderboard and each participating team will receive the model validation results after the challenge. But only the team that submits the method description paper will be eligible for the final ranking.

2) Open source code is required. The organizers encourage publicly available code submissions.

f) Define the publication policy. In particular, provide details on ...

- ... who of the participating teams/the participating teams' members qualifies as author
- ... whether the participating teams may publish their own results separately, and (if so)
- ... whether an embargo time is defined (so that challenge organizers can publish a challenge paper first).

An overview paper will be written with the organizing team's members. Each participating team who presented their method at the challenge session is allowed two co-authorships. The participating teams are encouraged to publish their results separately elsewhere when citing the overview paper, and (if so) no embargo will be applied.

Submission method

a) Describe the method used for result submission. Preferably, provide a link to the submission instructions.

Examples:

- Docker container on the Synapse platform. Link to submission instructions: <URL>
- Algorithm output was sent to organizers via e-mail. Submission instructions were sent by e-mail.

Submissions will be made via Docker container. The participating teams should upload their algorithm to the online submission system. The score will be immediately computed by the server and shown on result leaderboard of challenge website.

b) Provide information on the possibility for participating teams to evaluate their algorithms before submitting final results. For example, many challenges allow submission of multiple results, and only the last run is officially counted to compute challenge results.

Participants will be allowed to validate their algorithm on validation set three times before a pre-defined deadline and only once on the testing set.

Challenge schedule

Provide a timetable for the challenge. Preferably, this should include

- the release date(s) of the training cases (if any)
- the registration date/period
- the release date(s) of the test cases and validation cases (if any)
- the submission date(s)
- associated workshop days (if any)
- the release date(s) of the results

- 15 March, 2024: Announcement of challenge's opening and Open for registration.
- 20 March, 2024: Training data release.
- 15 June, 2024: Deadline for validation submission.
- 15 July, 2024: Deadline for testing submission.
- 1 September, 2024: Invite top teams to prepare presentations and participate in the MICCAI2024 Satellite Event.
- 6/12 October, 2024: Announce final results.

Ethics approval

Indicate whether ethics approval is necessary for the data. If yes, provide details on the ethics approval, preferably institutional review board, location, date and number of the ethics approval (if applicable). Add the URL or a reference to the document of the ethics approval (if available).

- Two already public ISBI challenge and PKU cephalogram dataset with a total of 502 dental X-ray images were used. Not required the ethics approval.
- The collection and use of the 98 dental X-ray images have been approved by the Research Ethics Committee of Shenzhen University General Hospital.
- The first public 100 dental X-ray images collection and use has been approved by the Research Ethics Committee of the Fifth Affiliated Hospital of Xinjiang Medical University.
All datasets are fully anonymized.

Data usage agreement

Clarify how the data can be used and distributed by the teams that participate in the challenge and by others during and after the challenge. This should include the explicit listing of the license applied.

Examples:

- CC BY (Attribution)
- CC BY-SA (Attribution-ShareAlike)
- CC BY-ND (Attribution-NoDerivs)
- CC BY-NC (Attribution-NonCommercial)
- CC BY-NC-SA (Attribution-NonCommercial-ShareAlike)
- CC BY-NC-ND (Attribution-NonCommercial-NoDerivs)

CC BY NC ND.

Code availability

a) Provide information on the accessibility of the organizers' evaluation software (e.g. code to produce rankings). Preferably, provide a link to the code and add information on the supported platforms.

The information on the accessibility of the organizers' evaluation software can be found on the challenge website.

b) In an analogous manner, provide information on the accessibility of the participating teams' code.

Top 10 teams should make their code publicly available. The remaining teams are encouraged to make their code publicly available.

Conflicts of interest

Provide information related to conflicts of interest. In particular provide information related to sponsoring/funding of the challenge. Also, state explicitly who had/will have access to the test case labels and when.

There is no explicit sponsoring/funding of the challenge, but we aim to contact technology companies which are willing to sponsor the challenge in the form of awards and computational resources. Only organizing members of the challenge has access to the test case labels during and after the CL-Detection 2024 challenge.

MISSION OF THE CHALLENGE

Field(s) of application

State the main field(s) of application that the participating algorithms target.

Examples:

- Diagnosis
- Education
- Intervention assistance
- Intervention follow-up
- Intervention planning
- Prognosis
- Research
- Screening
- Training
- Cross-phase

Diagnosis, Treatment planning, Intervention planning, Research.

Task category(ies)

State the task category(ies)

Examples:

- Classification
- Detection
- Localization
- Modeling
- Prediction
- Reconstruction
- Registration
- Retrieval
- Segmentation
- Tracking

Localization.

Cohorts

We distinguish between the target cohort and the challenge cohort. For example, a challenge could be designed around the task of medical instrument tracking in robotic kidney surgery. While the challenge could be based on ex vivo data obtained from a laparoscopic training environment with porcine organs (challenge cohort), the final biomedical application (i.e. robotic kidney surgery) would be targeted on real patients with certain characteristics

defined by inclusion criteria such as restrictions regarding sex or age (target cohort).

a) Describe the target cohort, i.e. the subjects/objects from whom/which the data would be acquired in the final biomedical application.

The target cohort are the patients with cephalometric analysis from any possible hospital or medical centers diagnosed by X-ray. The clinical goals are two fold; the automatically detected landmarks can be used for (i) providing diagnosis information of the craniofacial condition of a patient, (ii) further serving as the datum of the succeeding qualitative assessment of angles and distances which would affect treatment planning decision.

b) Describe the challenge cohort, i.e. the subject(s)/object(s) from whom/which the challenge data was acquired.

The challenge cohort consists of subjects with cephalometric analysis. A total of 700 X-ray images are included, which includes two already public ISBI challenge and PKU cephalogram dataset with a total of 502 dental X-ray images, 98dental X-ray images collected from Shenzhen General Hosiptal, and the first public 100 dental X-ray images collection. We annotated a total of 53 craniofacial landmarks for all cases, establishing a comprehensive benchmark for cephalometric landmark detection methods.

Imaging modality(ies)

Specify the imaging technique(s) applied in the challenge.

Dental X-ray.

Context information

Provide additional information given along with the images. The information may correspond ...

a) ... directly to the image data (e.g. tumor volume).

The images are 2D dental X-ray images (.bmp format with 8-bit depth) that were collected in different sites around China using different imaging machines (e.g. Sordex CRANEXr Excel Ceph, Sordex Cranex D Ceph and Planmeca ProMax).

b) ... to the patient in general (e.g. sex, medical history).

No additional context information will be given.

Target entity(ies)

a) Describe the data origin, i.e. the region(s)/part(s) of subject(s)/object(s) from whom/which the image data would be acquired in the final biomedical application (e.g. brain shown in computed tomography (CT) data, abdomen shown in laparoscopic video data, operating room shown in video data, thorax shown in fluoroscopy video). If necessary, differentiate between target and challenge cohort.

The data origin are dental X-ray images of head from the clinical routine.

b) Describe the algorithm target, i.e. the structure(s)/subject(s)/object(s)/component(s) that the participating algorithms have been designed to focus on (e.g. tumor in the brain, tip of a medical instrument, nurse in an operating theater, catheter in a fluoroscopy scan). If necessary, differentiate between target and challenge cohort.

The participating algorithms will be designed to localize the cephalometric landmarks in lateral X-ray images.

Assessment aim(s)

Identify the property(ies) of the algorithms to be optimized to perform well in the challenge. If multiple properties are assessed, prioritize them (if appropriate). The properties should then be reflected in the metrics applied (see below, parameter metric(s)), and the priorities should be reflected in the ranking when combining multiple metrics that assess different properties.

- Example 1: Find highly accurate liver segmentation algorithm for CT images.
- Example 2: Find lung tumor detection algorithm with high sensitivity and specificity for mammography images.

Corresponding metrics are listed below (parameter metric(s)).

Accuracy, Precision, Runtime, Robustness, Usability.

DATA SETS

Data source(s)

a) Specify the device(s) used to acquire the challenge data. This includes details on the device(s) used to acquire the imaging data (e.g. manufacturer) as well as information on additional devices used for performance assessment (e.g. tracking system used in a surgical setting).

X-ray images were acquired from systems of different vendors such as Sordex CRANEXr Excel Ceph, Sordex Cranex D Ceph and Planmeca ProMax. An inhouse developed software was used to develop the labels of the landmarks from the X-ray images.

b) Describe relevant details on the imaging process/data acquisition for each acquisition device (e.g. image acquisition protocol(s)).

Regarding the imaging process and data acquisition for each acquisition device, here are the relevant details:

Sorder CRANEXr Excel Ceph: All cephalograms were acquired in TIFF format with a Soredex CRANEXr Excel Ceph machine (Tuusula, Finland) using Soredex SorCom software (3.1.5, version 2.0). The image resolution was 1935×2400 pixels with a pixel spacing of 0.1mm/pixel.

Sordex Cranex D Ceph: The X-ray images were acquired in PNG format with a Soredex Cranex D Ceph machine (Tuusula, Finland) using CLINIVIEW software (version 10.2.6.4) or Digora software. The image resolution was 2880×2304 pixels with a pixel spacing of 0.096/pixel.

Planmeca ProMax: These X-ray images were acquired by Planmeca ProMax 3D machine (Finland) and Planmeca Romexis software. The average resolution size of these images from this machine is 2089 × 1937 pixels, while the pixel spacing is about 0.125 mm/pixel.

c) Specify the center(s)/institute(s) in which the data was acquired and/or the data providing platform/source (e.g. previous challenge). If this information is not provided (e.g. for anonymization reasons), specify why.

Some of the datasets (400/700) were from the ISBI challenge, some datasets (102/700) were acquired from the PKU cephalogram dataset, and some dataset (98/700) were collected from Shenzhen University General Hospital. We further extend another dataset (100/700) supplied by the Fifth Affiliated Hospital of Xinjiang Medical University.

d) Describe relevant characteristics (e.g. level of expertise) of the subjects (e.g. surgeon)/objects (e.g. robot) involved in the data acquisition process (if any).

The labels were provided by Dr. Jun Cao with over 20 years of experience at Department of Dental, Shenzhen University General Hospital, School of Medicine, Shenzhen University, China.

Training and test case characteristics

a) State what is meant by one case in this challenge. A case encompasses all data that is processed to produce one result that is compared to the corresponding reference result (i.e. the desired algorithm output).

Examples:

- Training and test cases both represent a CT image of a human brain. Training cases have a weak annotation (tumor present or not and tumor volume (if any)) while the test cases are annotated with the tumor contour (if any).
- A case refers to all information that is available for one particular patient in a specific study. This information always includes the image information as specified in data source(s) (see above) and may include context information (see above). Both training and test cases are annotated with survival (binary) 5 years after (first) image was taken.

Training and test cases both represent cephalometric X-ray image of each patient. The training cases include the corresponding annotations of landmark. A case refers to a cephalometric patient.

b) State the total number of training, validation and test cases.

The total number of cases is 700. The dataset to-be-released contains a training set of 400 cases, a validation set of 100 cases and a test set of 200 cases.

c) Explain why a total number of cases and the specific proportion of training, validation and test cases was chosen.

The 600 cases in the CL-Detection2024 challenge dataset follow the same data division as CL-Detection2023 (400 for training, 100 for validation, and 100 for testing). An additional 100 cases collected from the Fifth Affiliated Hospital of Xinjiang Medical University have been included in the test set to evaluate the algorithm's generalization.

d) Mention further important characteristics of the training, validation and test cases (e.g. class distribution in classification tasks chosen according to real-world distribution vs. equal class distribution) and justify the choice.

None.

Annotation characteristics

a) Describe the method for determining the reference annotation, i.e. the desired algorithm output. Provide the information separately for the training, validation and test cases if necessary. Possible methods include manual image annotation, in silico ground truth generation and annotation by automatic methods.

If human annotation was involved, state the number of annotators.

Manual landmark annotations were made by using in-house developed software. The annotations were performed by two experienced doctors. Both doctors were given training on the annotation software by the software developers.

b) Provide the instructions given to the annotators (if any) prior to the annotation. This may include description of a training phase with the software. Provide the information separately for the training, validation and test cases if necessary. Preferably, provide a link to the annotation protocol.

The annotators, in accordance with the guidelines outlined in dental X-ray image reporting and Contemporary Orthodontics [10], additionally classified each landmark into distinct categories, including soft tissue, teeth-related, and spine-related.

c) Provide details on the subject(s)/algorithm(s) that annotated the cases (e.g. information on level of expertise such as number of years of professional experience, medically-trained or not). Provide the information separately for the training, validation and test cases if necessary.

The annotation process was carried out by 2 senior doctors with more than 5 years of experience, under the supervision of Dr. Jun Cao, who possesses over 20 years of experience in cephalometric X-ray imaging.

The doctors extent the CL-Detection2023 challenge dataset with more landmark annotations, and annotated landmarks from scratch on the first public dataset of 100 images. Next, they conducted a double-check on the quality of all landmarks to ensure precision. Finally, our dataset, includes 700 X-ray images from 4 medical centers with multi-center, multi-vendor and more-landmarks.

d) Describe the method(s) used to merge multiple annotations for one case (if any). Provide the information separately for the training, validation and test cases if necessary.

No Aggregation.

Data pre-processing method(s)

Describe the method(s) used for pre-processing the raw training data before it is provided to the participating teams. Provide the information separately for the training, validation and test cases if necessary.

The X-ray images were anonymised by first removing any patient related information on each image. Then all the images were renamed by assigning an ID number to each patient and converted to the same format (.bmp). No information concerning the relations between the identity and the ID number of each patient will be released to the public.

Sources of error

a) Describe the most relevant possible error sources related to the image annotation. If possible, estimate the magnitude (range) of these errors, using inter-and intra-annotator variability, for example. Provide the information separately for the training, validation and test cases, if necessary.

Due to the relatively low brightness of soft tissue regions, some landmarks are illegible, which would likely result in inconsistent annotation that affects the inference performance of models and the selection of models. Majority of the landmarks are clear enough to allow accurate annotation. For landmarks on soft tissue, the two senior doctors manually adjust the image contrast using in-house developed software. This process enhances the discernibility of soft tissue structures and ensures the accurate annotation of corresponding landmarks.

In addition, we have calculated the inter-observer variability of the two senior doctors to assess human performance for each landmark. Current findings suggest that Interclass Correlation Coefficients (ICCs) are excellent for all landmarks (ICC over 0.90 for each landmark). Moreover, our double-check and the quality control

approach will largely reduce this source of error.

b) In an analogous manner, describe and quantify other relevant sources of error.

None.

ASSESSMENT METHODS

Metric(s)

a) Define the metric(s) to assess a property of an algorithm. These metrics should reflect the desired algorithm properties described in assessment aim(s) (see above). State which metric(s) were used to compute the ranking(s) (if any).

- Example 1: Dice Similarity Coefficient (DSC)
- Example 2: Area under curve (AUC)

The detection inference results are evaluated using the Mean Radial Error (MRE), Success Detection Rate (SDR) for 2.0 mm precision range, Running time (RT) and Area under GPU memory-time curve (GPU-area).

All metrics will be used to compute the ranking.

b) Justify why the metric(s) was/were chosen, preferably with reference to the biomedical application.

The four metrics are complementary.

Specifically, the Mean Radial Error (MRE) measures the difference between two landmarks. The error between the ground truth implant and the automatic results from participants' algorithm can be effectively measured using this metric. MRE serves as a distance error metric, with a lower value indicating better performance, formulated as follows:

MER = sum(sqrt(delta_xi^2+delta_yi^2))/N, i from 1 to N

where delta_xi is the absolute distance in x-direction between the predicted and the reference landmark, delta_yi is the absolute distance in y-direction between the predicted and the reference landmark, and N represents the number of detection landmark.

Success Detection Rate (SDR) measures the accuracy between the ground truth implant and the automatic results. It has been shown to provide a good measure of localization quality and is widely used in landmark detection applications. SDR represents the proportion of successfully detected landmarks relative to the total number of landmarks and serves as an accuracy metric, with a higher value indicating better performance as more points are closer to the reference landmarks. SDR with precision less than 2.0mm is formulated as follows:

SDR = the cardinality of {j:L2 norm(Ld(j)-Lr(j))<z}/N*100%

where Ld and Lr, represent the locations of the predicted landmark and the reference landmark, z denotes 2.0mm precision range used in the evaluation, and N represents the number of landmarks.

Running time (RT) measures the duration from initiating the model to obtaining its output. RT refers to the average processing time for each sample's complete execution, encompassing the entire workflow from data read-in, through pre-processing, model inference, post-processing, to results write-out, evaluated under the native Torch environment. Moreover, we also include the area under the GPU utilization-time curve to measure the GPU consumption.

These evaluation metrics will guide the algorithm to strike a balance between performance and efficiency.

Ranking method(s)

a) Describe the method used to compute a performance rank for all submitted algorithms based on the generated metric results on the test cases. Typically the text will describe how results obtained per case and metric are aggregated to arrive at a final score/ranking.

This ranking scheme aim to fairly evaluate the performances of the participating team algorithms on cephalometric landmark detection task. The main benefit of this ranking scheme is that it can aggregate the metrics with different dimensions. A similar ranking scheme was also employed in MICCAI BraTS Challenge.

b) Describe the method(s) used to manage submissions with missing results on test cases.

We will exclude the participants who fail to report on the whole testing set.

c) Justify why the described ranking scheme(s) was/were used.

The ranking scheme includes the following three steps:
Step 1. Compute the four metircs for each testing case between the ground truth and the results from the participants.
Step 2. Take the mean of the four metrics over the test cases, and rank separately among the teams. MRE, RT, GPU-area are ranked in descending order and SDR is ranked in ascending order.
Step 3. A final overall rank is given by taking the average of the four ranks. In the case of equal average rankings for two teams, they are considered tied.

Statistical analyses

a) Provide details for the statistical methods used in the scope of the challenge analysis. This may include

- description of the missing data handling,
- details about the assessment of variability of rankings,
- description of any method used to assess whether the data met the assumptions, required for the particular statistical approach, or
- indication of any software product that was used for all data analysis methods.

We will exclude the participants who fail to report on the whole testing set. Besides the statistical values such as mean, standard deviation of the four evaluation metrics, we use the p-value in Wilcoxon test to assess whether the top performing/ranking algorithms are significantly better than the rest of algorithms. To measure the variability, ranking variability will be characterized using the bootstrap.

b) Justify why the described statistical method(s) was/were used.

The Wilcoxon test is chosen because it is non-parametric and allows us to perform the analysis with minimal hypotheses, and the Bootstrap is a simple nonparametric method that relies on minimal assumptions.

Further analyses

Present further analyses to be performed (if applicable), e.g. related to

- combining algorithms via ensembling,
- inter-algorithm variability,
- common problems/biases of the submitted methods, or
- ranking variability.

The further analyses will be discussed in a further publication after the challenge.

ADDITIONAL POINTS

References

Please include any reference important for the challenge design, for example publications on the data, the annotation process or the chosen metrics as well as DOIs referring to data or code.

[1] Proffit W R, Fields Jr H W, Sarver D M. Contemporary orthodontics[M]. Elsevier Health Sciences, 2006.
[2] Ricketts R M. Orthodontic Diagnosis and Planning: --Their roles in preventive and rehabilitative dentistry[M]. Rocky Mountain/Orthodontics, 1982.
[3] Downs W B. Variations in facial relationships: their significance in treatment and prognosis[J]. American journal of orthodontics, 1948, 34(10): 812-840.
[4] Steiner C C. Cephalometrics for you and me[J]. American journal of orthodontics, 1953, 39(10): 729-755.
[5] Durão A P R, Morosolli A, Pittayapat P, et al. Cephalometric landmark variability among orthodontists and dentomaxillofacial radiologists: a comparative study[J]. Imaging science in dentistry, 2015, 45(4): 213-220.
[6] Wang C W, Huang C T, Lee J H, et al. A benchmark for comparison of dental radiography analysis algorithms[J]. Medical image analysis, 2016, 31: 63-76.
[7] Zeng M, Yan Z, Liu S, et al. Cascaded convolutional networks for automatic cephalometric landmark detection[J]. Medical Image Analysis, 2021, 68: 101904.
[8] Li W, Lu Y, Zheng K, et al. Structured landmark detection via topology-adapting deep graph learning[C]//European Conference on Computer Vision. Springer, Cham, 2020: 266-283.
[9] Zhong Z, Li J, Zhang Z, et al. An attention-guided deep regression model for landmark detection in cephalograms[C]//International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, Cham, 2019: 540-548.
[10] Proffit W R, Fields Jr H W, Sarver D M. Contemporary orthodontics[M]. Elsevier Health Sciences, 2006.
[11] CL-Detection 2023 Challenge Homepage: https://cl-detection2023.grand-challenge.com.

Further comments

Further comments from the organizers.

N/A