Pubic Symphysis-Fetal Head Segmentation from Transperineal Ultrasound Images: Structured description of the challenge design

CHALLENGE ORGANIZATION

Title

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

Pubic Symphysis-Fetal Head Segmentation from Transperineal Ultrasound Images

Challenge acronym

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

PSFHS

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.

The high risk of maternal and perinatal morbidity is associated with longer labor duration due to the slow progression of fetal descent, but accurate assessment of fetal descent by monitoring the fetal head (FH) station remains a clinical challenge in guiding obstetric management. Based on clinical findings, the transvaginal digital examination is the most commonly used clinical estimation method of fetal station. However, this traditional approach is very subjective, often difficult, and unreliable. The need of an objective diagnosis found its solution in the use of transperineal ultrasound (TPU) able to assess FH station by measuring the angle of progression (AoP) that is the extension the FH goes through in its descent.

Manual segmentation of symphysis pubis (SP)-fetal head from ITU images for clinical radiologists is considered as the most reliable but extremely time-consuming procedure prone to subjectivity and large inter-observer variability. With the rapid development of artificial intelligence in medical images, automatic measurement algorithms based on ITU images are expected to solve the above problems.

Challenge keywords

List the primary keywords that characterize the challenge.

Angle of progression; Transperineal ultrasound image; Pubic symphysis; Fetal head; Image segmentation

Year

The challenge will take place in ...

2023

FURTHER INFORMATION FOR MICCAI ORGANIZERS

Workshop

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

none

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.

Based on the number (747) of participants of the 2018 challenge for automated measurement of the fetal head circumference, the number (>200 hospitals in China) of clients using intrapartum ultrasound of our sponsorand the
impact of Prof. Dong Ni in fetal ultrasound research, we optimistically estimate that 50 teams will participate in this
challenge, and expect the number of teams that provide a final submission to be 10~15.

Publication and future plans

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

We aim to summarize the design, proposed methods and results of the challenge in a manuscript to be submitted to a peer-reviewed scientific journal in the field of medical image analysis. To this end, we aim to invite the best performing participants to contribute by describing their methods and experiences. Furthermore, we aim to make the code of the best performing methods publicly available for the purpose of reproduction of results and further research with the license(CC-BY-NC-ND).

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.

For algorithm implementation and training, the participants will use their own resources. For testing as part of the challenge we would use the platform grand-challenge.org.

TASK: Segmentation of PS-FH

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.

The high risk of maternal and perinatal morbidity is associated with longer labor duration due to the slow progression of fetal descent, but accurate assessment of fetal descent by monitoring the fetal head (FH) station remains a clinical challenge in guiding obstetric management. Based on clinical findings, the transvaginal digital examination is the most commonly used clinical estimation method of fetal station. However, this traditional approach is very subjective, often difficult, and unreliable. The need of an objective diagnosis found its solution in the use of transperineal ultrasound (TPU) able to assess FH station by measuring the angle of progression (AoP) that is the extension the FH goes through in its descent.

Manual segmentation of symphysis pubis (SP)-fetal head from ITU images for clinical radiologists is considered as the most reliable but extremely time-consuming procedure prone to subjectivity and large inter-observer variability. With the rapid development of artificial intelligence in medical images, automatic measurement algorithms based on ITU images are expected to solve the above problems.

Keywords

List the primary keywords that characterize the task.

Angle of progression; Transperineal ultrasound image; Pubic symphysis; Fetal head; Image segmentation

ORGANIZATION

Organizers

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

Organizing team:
Jieyun Bai, Guangdong Provincial Key Laboratory of Traditional Chinese Medicine Information Technology, Jinan University, China
Zhanhong Ou, College of Information Science and Technology, Jinan University, China
Yaosheng Lu, Guangdong Provincial Key Laboratory of Traditional Chinese Medicine Information Technology, Jinan University, China
Dong Ni, Shenzhen University, China
Gaowen Chen, Obstetrics and Gynecology Center, Zhujiang Hospital, Southern Medical University, China

Clinical Evaluators and Annotation Approvers:
Gaowen Chen, Obstetrics and Gynecology Center, Zhujiang Hospital, Southern Medical University, China
Zhanhong Ou, College of Information Science and Technology, Jinan University, China

Technical Group:
Zhanhong Ou, College of Information Science and Technology, Jinan University, China

Data Contributor:
Gaowen Chen, Obstetrics and Gynecology Center, Zhujiang Hospital, Southern Medical University, China
Yaosheng Lu, Guangdong Provincial Key Laboratory of Traditional Chinese Medicine Information Technology, Jinan University, China

b) Provide information on the primary contact person.

Jieyun Bai bai_jieyun@126.com or baijieyun@jnu.edu.cn
Zhanhong Ou Zhanhong_Ou@foxmail.com;

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

Repeated event with fixed 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.

grand-challenge.org for publicity and own online submission site for final model submission

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

None at this moment

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.

Additional data 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.

May participate but not eligible for awards and not listed in leaderboard.

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

We foresee the following awards for the best submission:
First prize: 3,000 RMB
Second prize: 2,000 RMB
Third prize: 1,000 RMB
4rd-7th: 500 RMB

At the moment, we are in the process of coordinating support for this challenge by the Lianying Medical Technology Co., Ltd and by potential sponsors. Depending on the outcome of this process this award policy may be adapted.

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.

All submissions will be reported in the leaderboad.
Participating teams can opt out of publication of their results in the leaderboard.
Top 7 performing methods will be announced publicly as part of a scientific session at the MICCAI annual meeting.

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).

The top 7 performing teams will be invited to contribute to draft and submit a manuscript describing the methods and results of the challenge in a peer-reviewed journal. The first and last author of the top 7 submissions will be also listed as authors of this planned manuscript. The participating teams may publish their own results separately after coordination to avoid siginficant overlap with the challenge paper.

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.

Algorithm container submission (type 2) on Grand Challenge

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.

Provide the large size of the training set.
The test data will be released three weeks before the submission.
Multiple submissions are allowed. However, only the last submission will be considered for the challenge results.
Also, the number of Docker submissions will be limited to one per day.

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

Challenge announcement and release of training cases: 03/2023
Registration: starting 03/2023
Submission deadline: 09/2023

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).

All data is anonymized.
Ethics approval by the Medical Ethics Committee of NanFang Hospital of Southern Medical University (NFCE-2019–024) and Jinan University(JNUKY-2022-019)

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.

Evaluation code will be made public before the system is open for submission

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

Publication of algorithm code will be a prerequisite for award eligibility. To this end code will need to be published on a publicly accessible repository of the teams' choice after within a week after submission deadline.

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 are no conflicts of interest.
Lianying Medical Technology Co., Ltd is the principal sponsor of the challenge by collecting and providing clinical data.
Only the organisers and technical group of the challenge have access to test case labels.

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

Decision support, Research, Treatment planning, CAD, Diagnosis, Prognosis.

Task category(ies)

State the task category(ies).

Examples:

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

Segmentation and Prediction

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 patients requiring to assess fetal head station in labor

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 patients with different fetal head station in labor in two large medical centers.

Imaging modality(ies)

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

standard B-mode ultrasound

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 B mode ultrasound images that were collected in different sites around China using different ultrasound machines

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

The images are acquired from pregnant woman with a variety of age (from 18-year old to 46-year old)

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 is pubic symphysis-fetal head in the 2D B mode ultrasound image

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 algorithm target is pubic symphysis-fetal head shown in the 2D B mode ultrasound image. Each image has three points used to calculate AoP.

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)).

Dice coefficient, mean-surface-distance, Hausdorff distance, AoP difference between predicted and measured AoP

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).

Transperineal ultrasound examinations were performed in standard B-mode ultrasound using systems of different vendors such as Philiph-cx50, Toshiba Aplio300, Voluson P8, Esaote Mylab, Lian-med ObEye and Youkey Q7

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

In order to obtain high-quality images, the transducer was prepped by covering it with a surgical latex glove filled with coupling gel, then the prepped transducer, after applying gel, was placed between labia below the pubic symphysis to obtain a sagittal plane, small adjustments in the form of lateral movements of the probe were made until an image obtained showed clear maternal pelvic (pubic symphysis) and fetal (fetal skull) landmarks that did not show any shadows from the pubic rami

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.

All images have been already been acquired at NanFang Hospital of Southern Medical University/Jinan University and were not previously released in any challenge

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 data was collected from different sites by local doctors, while the label were provided by 7 annotators, including 2 advisors and 5 graduate/undergraduate students

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 a 2D ultrasound image of the pubic symphysis (PS)-fetal head (FH) from a patient. The training cases include the corresponding annotations of target segmentation and measured AoP.

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

4000 training cases, 401 Test cases in the first stage 700 Test cases in the second stage

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

time for the very laborious ground truth annotation process.
1101 test cases: ~20% is a reasonable proportion. We will use 1101 new test cases that include a wide range of pregnant woman pelvis.
4000 training cases: We will release a single training dataset, instead of separate training and validation datasets. since the total dataset size is moderate at 5101 cases. Participants can split the 4000 cases into separate training and validation datasets if they choose, or use the entire training dataset for cross-validation

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.

The split of the train and test is balanced in term of the manufacturer devices. The split of training and test data is to
make sure the test set is large enough to yield statistically meaningful results and is representative of the dataset. In other words, our goal is to make sure that the model generalizes well to new data.

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.

All segmentations were performed manually using the software Pair.There were 7 annotators, including 2 advisors and 5 graduate and undergraduate students.

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.

According to suggestions of radiologists, the following points in the image annotation were abided: (1) In an ideal situation, the pubic symphysis (PS) and fetal head are elliptical in the two-dimensional image; (2) In grayscale ultrasound images, the outer borders of the SP and fetal head mainly appear bright white. When there is no obvious white border, the boundary is determined according to the difference of the local gray value; and (3) The

lower right corner of the pubic symphysis is adjacent to the bladder, and the boundary of the pubic symphysis should be determined with the bottom edge of the white area.

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.

For target segmentation, annotations were supervised by D.Q. and G.C., under very close advice. D.Q. and G.C. are currently obstetricians with more than 7 years of experience in medical image analysis. The trained raters who manually traced the interfaces were graduate and undergraduate students in science/engineering or pre-med, whose work was closely reviewed by D.Q. and then G.C., and corrected where necessary. G.C. performed the final segmentation cleanup step. Before the challenge, all segmentations will be reviewed by a clinician and adjusted as necessary.

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.

N/A

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 US images were anonymised by first removing any patient related information on each image. Then all images were renamed and converted to the same format

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.

The delineation of structures in ultrasound images is a challenging task as some of the boundaries is less welldefined.
We create multiple segmentations for our test set by same raters and by different raters and give an estimate of intra- and inter-subject variability.

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

A further potential error source are image artifacts that may result in altered image properties.

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)

Dice score, Hausdorff distance and Mean Surface Distance for target segmentation;

The AoP difference (AoP) between predicted and manually measured AoP for prediction.

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

Dice is more sensitive to the inner filling of the mask, while Hausdorff distance and Mean Surface Distance are more
sensitive to the segmented boundary. The combination of these two scores is sensible for our application.
AoP is calculated the difference between the prediction and the ground truth. It is a good indicator of whether the prediction is consistent with the label.

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.

For each case, we will compute the 2 Dice scores, 2 Hausdorff distance measures , 2 Mean Surface Distance and 1 AoP, which will be averaged across all cases to form a mean Dice score (mDSC), a mean Hausdorff distance (mHD) ,
a Mean Surface Distance and a mean (1-AoP/AoPg)(larger is better). The mDSC is in [0,1] (higher is better), and the mHD and MSD are typically in [0, 1] (lower is better). Our proposed metric aggregation is:
S=0.25mDSC+0.25[0.5(1-mHD)+0.5(1-MSD)]+0.5mean(1-AoP/AoPg), where AoPg is measured AoP.

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

If there are the missing results(1)incomplet test cases and 2)not a three-category result) on the test cases, the submission will fail and the error notification will appear. The participants need to re-submit their
results.

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

This ranking scheme aim to fairly evaluate the performances of the participating team algorithms on the individual and combination of the segmentation task and AoP prediction task. From the ranking scheme, we can compare the performances of them and define the top winners.

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.

For unsegmented data, DSC=0, HD=1, SD=1 and mean(1-AoP/AoPg)=1. We create multiple
segmentations for our test set by same raters and by different raters and give an estimate of intra- and intersubject variability. Further, we compute average segmentations via aggregation methods.

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

Ultrasound annotation is difficult, even for human raters. This would be possible to say whether a segmentation model is close to the average human rater opinion, and/or whether it is within a certain standard deviation from a

multi-rater segmentation.

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 challenge organizers will analyze the submitted algorithms in a journal publication. Depending upon the submitted methods, this may include comparisons of different classes of algorithms (e.g., deep learning vs others), typical failure modes, etc

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.

Progression[J]. Computational and mathematical methods in medicine, 2022, 2022: 5192338.
Lu Y, Zhou M, Zhi D, et al. The JNU-IFM dataset for segmenting pubic symphysis-fetal head[J]. Data in brief, 2022, 41: 107904.
Bai J, Yu S, Lu Y, et al. A Framework for Computing Angle of Progression from Transperineal Ultrasound Images for Evaluating Fetal Head Descent Using a Novel Double Branch Network[J]. Frontiers in Physiology, 2565.
Zhou M, Yuan C, Chen Z, et al. Automatic angle of progress measurement of intrapartum transperineal ultrasound image with deep learning[C]//International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, Cham, 2020: 406-414.

Further comments

Further comments from the organizers.

This challenge is an initiative of Jinan University, Southern Medical University and Shenzhen University.