Trackerless 3D Freehand Ultrasound Reconstruction Challenge: Structured description of the challenge design

CHALLENGE ORGANIZATION

Title

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

Trackerless 3D Freehand Ultrasound Reconstruction Challenge

Challenge acronym

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

TUS-REC

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.

Reconstructing 2D Ultrasound (US) images into a 3D volume enables 3D representations of anatomy to be generated which are beneficial to a wide range of downstream tasks such as quantitative biometric measurement, multimodal registration, 3D visualisation and interventional guidance. Although substantive progress has been made recently through non-deep-learning- and deep-learning-based approaches, this application is still challenging due to 1) inherent accumulated error - frame-to-frame transformation error will be accumulated through time when reconstructing long sequence of US frames, and 2) a lack of publicly-accessible data with synchronised spatial location, often obtained from tracking devices, for benchmarking the performance and for training learning-based methods.

This field has witnessed the development of 3D US reconstruction from previous non-deep learning approaches such as speckle decorrelation (Chen et al. 1997) and linear regression (Prager et al. 2003) to current deep learning-based methods such as convolutional and Long Short-Term Memory (LSTM) neural networks (Guo et al. 2020, Mikaeili et al. 2022, Miura et al. 2021). One of the first deep learning based approach was proposed in (Prevost et al. 2018), in which the proposed CNN model is compared with speckle decorrelation. Since, various network models were adapted into this application, such as ConvLSTM (Luo et al. 2021) and transformers (Ning et al. 2022). Additional information (for instance signals from inertial measurement units) and sequential models have been investigated to improve the reconstruction performance (Prevost et al. 2018, Luo et al. 2022, Guo et al. 2022). However, none of these studies used the same dataset, there is a clear need for performance benchmarking to establish a standardized basis for comparison and evaluation. Moreover, several learning-based methods have relied on data from only 12-40 subjects, highlighting the necessity for additional open training data.

The proposed TUS-REC challenge aims to provide in vivo US data, acquired from both left and right forearms of

one hundred volunteers (1200 scans, approximately 1,206,900 frames in total), tracked by a time-synchronised optical tracker, to provide a ground-truth for comparing different 3D reconstruction methods and more importantly to advance the discovery of new methods for freehand US reconstruction. The volunteer study of forearms is widely used in literature (Prevost et al. 2018, Luo et al. 2022) which is ethically and practically feasible and thus is the first step towards other specific, potentially more complex clinical applications. The outcome of the challenge includes 1) open-sourcing the first largest tracked US datasets with accurate positional information; 2) establishing one of the first benchmarks for 3D US reconstruction, suitable for modern learning-based data-driven approaches. While the challenge cannot currently transition into clinical applications, it paves the way from experimental volunteer studies to potential clinical applications for this challenging task.

Challenge keywords

List the primary keywords that characterize the challenge.challenge_

Trackerless, Freehand, Ultrasound, 3D Reconstruction, Spatial transformation estimation

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.

Agreed to be held in conjunction with the 5th ASMUS workshop.

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.

From the experience of the MICCAI challenge mu-RegPro, organised by us, and other data challenge events, e.g., Learn2Reg and BRATS-Reg, we expected 20-30 attendees, with 10-15 unique entries, before the main MICCAI 2024 deadline. We have at least three international groups that have already expressed interest in participating. However, we expect long-lasting lifetime for this challenge with many more future participants, after the conference.

Publication and future plans

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

We were planning to submit a challenge paper including the analysis of the dataset and the results. Members of the top five teams will be invited as co-authors. Additionally, we encourage participant teams submit papers about their methodologies for freehand US reconstruction.

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.

The challenge will take place online. No on-site computing resources are required at the conference. We anticipate the in-person event will summarise the results and report the overall experience, as part of the ASMUS 2024 workshop, within a two-hour slot, with the same technical equipment support, including projectors, monitors, loud speakers, and microphones.

TASK 1: Trackerless Freehand Ultrasound Reconstruction

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.

This challenge aims to provide a benchmark for freehand US reconstruction with publicly available data from forearms of one hundred volunteers, using multiple predefined scanning protocols, targeted at improving the reconstruction performance in this challenging task.

Various reconstruction approaches from non-deep leaning to deep learning can be proposed and tested with the same public dataset including various scan trajectories and number of frames, which enables the investigation of potentially long sequence of ultrasound frames, with variable anatomy and protocol. The accuracy will be measured with four types of carefully-designed evaluation metrics, representing two levels of local and global reconstruction performances (details are discussed in Section "Metric"), so will be the computational speed. Established standard algorithms, such as Prevost et al. 2018 and Li et al 2023, will be provided as the baselines. Participant teams are expected to make use of the sequential data and potentially make knowledge transfer from other domains such as computer vision and computer-assisted intervention. The participant teams are expected to take US scan as input and output two sets of pixel displacement vectors, indicating the transformation to reference frame (details are clarified in Section "Metric"). The evaluation process will take the generated displacement vectors from their dockerized models, and produce the final accuracy score to represent the reconstruction performance, at local and global levels, representing different clinical application of the reconstruction methods.

This challenge has the potential to provide high-quality evidence on the accuracy and speed of trackerless reconstruction algorithms, which can be used for the purposes of comparison and inform further technical developments.

Keywords

List the primary keywords that characterize the task.

Trackerless, Freehand, Ultrasound, Three-dimensional, Reconstruction, Transformation

ORGANIZATION

Organizers

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

Qi Li, University College London
Shaheer U. Saeed, University College London
Dean C. Barratt, University College London
Matthew J. Clarkson, University College London

Tom Vercauteren, King's College London
Yipeng Hu, University College London

b) Provide information on the primary contact person.

Qi Li, qi.li.21@ucl.ac.uk

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

Open call

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://github.com/UCL

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

We will use https://github.com/UCL to run the challenge if the challenge is accepted.

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.

Public and private data are permitted, however their use must be disclosed by participants.

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.

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

Participants who successfully participated the challenge will be awarded certificates of participation. The first-place and runner-up achievers will receive additional certificates. Currently, we have two sponsors expressed interest to fund these awards and, as an official event of the MICCAI Special Interest Group on Medical Ultrasound (SIGMU), the SIG budget will be available for covering any offset. We are also active in seeking further potential sponsorships to enhance our ability to offer extra prizes.

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.

The results from all participants will be made publicly available on leaderboards unless the submitted codes incurred errors during the evaluation process. Teams are allowed to make multiple distinct submissions (but must ensure they are not merely simple variations in hyperparameter values). The top five teams will be invited to showcase their work during the challenge event through 10-15 minutes presentations as part of the live demonstration session held with ASMUS 2024.

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).

We are planning to submit a paper including challenge dataset summary and results analysis. Members of the top five participating teams will be invited as co-authors.

The participating teams can only publish their novel methodology without discussion of data and obtained results if the above mentioned paper is not published (submission to arXiv is considered as a sufficient waiting period). Once the challenge paper from the organizing team is published, the participants should cite this challenge paper if their work has not been published.

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.

We will prepare a Docker container to provide a consistent submission environment. The detailed information of the release and usage of the Docker container will be announced in our website mentioned above, once the challenge is accepted. Participants will be asked to Dockerize and submit their trained model/algorithm to us, via our secured server hosted at our university.

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.

We are planning to provide a small validation set, which allows participants to tune their models using these unseen data. The participants are allowed to submit multiple submissions (as mentioned in Section "Result announcement policy"), and the best result will be selected for competing. The number of submissions for each team is limited to 5 and the total number of submissions is limited to 100, to preserve variations in hyperparameters.

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 website/registration open: Monday, April 1 2024
Training dataset release: Monday, May 13 2024
Validation dataset release: Monday, July 29 2024
Submissions begins: Monday, Aug 12 2024
Submissions Closes: Monday, Aug 19 2024
Winners Announcement and Speaker Invitations: Monday, Sep 16 2024
Challenge Event @ MICCAI 2024: Monday, Oct. 6 2024

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).

The ethics for the dataset has already been approved by the Ethics Committee of local institution (UCL Department of Medical Physics and Biomedical Engineering) on 20th Jan. 2023 with reference number [24055/001].

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 SA

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 evaluation code, together with the baseline models, will be publicly available once this section on our website is open to public.

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

Although not required, we encourage participants to make their code publicly available. We will provide links to the available code from participants on our website.

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.

No conflicts of interest associated with this challenge.

Only challenge organizers will have access to the test case labels. However, for maximising the transparency and reproducibility, all test labels will be made publicly available after a one-year embargo, when we plan to provide additional test data for continuation of the 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
-

Decision support, Intervention planning, Research, Treatment planning, Assistance, Surgery.

Task category(ies)

State the task category(ies)

Examples:

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

Reconstruction, Tracking

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.

A wide range of standard US scanning that can benefit from volume reconstruction but without external tracker can be regarded as the clinical applications of the challenge. Examples include abdominal US, fetal examination, head-and-neck scanning, vascular US, echocardiography, prostate and liver intervention guidance. One of the target cohorts specific to the data is US examination for vascularity and orthopaedic conditions.

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

The data in this challenge is acquired from both left and right forearms of 100 volunteers. No specific exclusion criteria as long as the participants do not have allergies or skin conditions which may be exacerbated by US gel.

Imaging modality(ies)

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

B-mode US

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 position information recorded by the optical tracker (NDI Polaris Vicra, Northern Digital Inc., Canada) will be provided along with the images, which indicates the position of the US probe for each frame in the camera coordinate system, described as homogeneous transformation matrix with respect to reference frame. The US images were acquired on an Ultrasonix machine (BK, Europe) with a curvilinear probe (4DC7-3/40). A calibration matrix will also be provided, denoting the transformation between US image coordinate system and US probe coordinate system while these data were acquired. Although temporal-calibrated data will be provided, the timestamps for both transformation from the optical tracker and ultrasound frames will also be provided, for reference purpose.

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

All scanned forearms are in good health. No further information is given on a per-subject basis.

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.

While this challenge focuses on US data from forearms, the proposed methodology could be adapted to other applications which involves standard US scanning such as liver, abdomen and cardiac images etc.

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 is expected to take the entire scan as input and output two different sets of transformation-representing displacement vectors as results, a set of displacement vectors on individual pixels and a set of displacement vectors on provided landmarks. There is no requirement on how the algorithm is designed internally, for example, whether it is learning-based method; frame-, sequence- or scan-based processing; or, rigid-, affine- or nonrigid transformation assumptions. Details are explained further in "Metric" section.

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, Runtime

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).

The 2D US images were acquired using an Ultrasonix machine (BK, Europe) with a curvilinear probe (4DC7-3/40). The associated position information of each frame was recorded by an optical tracker (NDI Polaris Vicra, Northern Digital Inc., Canada).

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

The acquired US frames were recorded at 20 fps, with an image size of 480×640, without speckle reduction. The frequency was set at 6MHz with a dynamic range of 83 dB, an overall gain of 48% and a depth of 9 cm.

Both left and right forearms of volunteers were scanned. For each forearm, the US probe moves in three different trajectories (straight line shape, "C" shape, and "S" shape), in a distal-to-proximal direction followed by a proximal-to-distal direction, with the US plane perpendicular of and parallel to the scanning direction. The dataset contains 1200 scans in total, 12 scans associated with each subject.

(Luo et al. 2023) designed four typical types of scans with diverse scanning velocities and poses: linear, loop, sector, and fast-slow. In our designed experiment, we also use three typical scan trajectories, and demonstrate the inherent dependency from anatomy and protocol (Li et al. 2023), in a controlled experiment where the scanning speed is kept relatively constant during individual scans.

The information recorded by the optical tracker includes the timestamp and probe to tracker homogeneous transformation matrix for each US frame.

The software used to acquire the data is Plus Toolkit (Lasso et al. 2014), which also provides a temporal calibration method. The spatial calibration is based on a pinhead-based method (Hu et al. 2017).

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 US image data with associated transformation data were acquired at University College London, London, U.K, with a racial-, gender-, age-diverse subject cohort.

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).

Both the US machine and optical tracker are calibrated and reviewed by a senior researcher with over 20 years' experience in tracked US field. The calibration process is based on a pinhead-based method (Hu et al. 2017). Most of the data was acquired by multiple biomedical imaging researchers with at least three years' experience, with

the help of the senior researcher.

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.

For each case in training, validation, and test sets, the data that can be processed include raw 2D images with time stamps from a freehand US scan, and the desired algorithm output is two different types of transformation-representing results, a set of displacement vectors on individual pixels and a set of displacement vectors on provided landmarks. The ground-truth transformation matrix will be provided for the training and validation datasets.

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

Training Set - 720 scans
Validation Set - 96 scans
Test Set - 384 scans

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

The total number of cases is determined by all the available data currently. We split the dataset into training, validation, and test cases with an approximate ratio of 60%, 8%, and 32% to ensure the representativeness of the test data from powering statistical testing requirement, which is illustrated in Section "Details for the statistical methods".

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 subjects were randomly split into training, validation and test datasets. The random process ensures the consistency of demographic characteristics across the three datasets. US scans from the same subject will be assigned to the same set which avoids the information leak.

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.

No manual annotation is needed as the label (transformation matrix of US probe for each US frame) is generated by the optical tracker and the accompanying software.

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 PLUS software was used to fetch US image data from US machine, along with the associated position information (i.e., label) simultaneously. All researchers who acquired the data went through a detail instruction of the software in the link (http://perk-software.cs.queensu.ca/plus/doc/nightly/user/DataAcquisitionProcessing.html), and were asked to pay attention to whether the subjects were within the field-of-view of the optical tracker during acquisition.

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.

One senior researcher fellow with over 20 years' experience in this application and multiple junior researcher with at least 3 years' experience acquired the dataset, with each scan reviewed by and agreed with each other.

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 multiple annotation merge is required since only one ground-truth transformation matrix per frame is generated by the PLUS software in this application.

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 frames in a scan with invalid transformation matrices were removed and the remaining raw images along with their associated transformation matrices were stacked in time order and stored as "key-value" record in a .h5 file, with additional information such as scan protocol and number of frames of this scan. The Python code snippets that reads the data from the h5 files will be provided.

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 major source of errors comes from the precision of the optical tracker. All the labels are acquired by an optical tracker with the 3D root-mean-square (RMS) volumetric accuracy acceptance criterion being less than or equal to 0.25 mm and the 3D RMS repeatability acceptance criterion being less than or equal to 0.20 mm. This may be considered insignificant compared to state-of-the-art reconstruction errors in this application, which has been widely reported in the range from several to tens of several millimetres.

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

The forearm may move slightly during the scanning. This type of errors is assumed random over different cases and the resultant variance will be taken into account during the statistical testing results when reporting and

summarising the results.

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)

For each scan, the submitted method is tasked to output two types of transformation-representing displacement vectors (hereinafter is also referred to as the prediction), at global and local levels.
- The "global displacement vectors" will be used to reconstruct individual frames (without the first frame), with respect to the first frame in the scan as the "global reference frame".
- The "local displacement vectors" will be used to reconstruct individual frames (without the first frame), with respect to the immediately previous frame in the scan as the "local reference frame".

For each scan, the performance of a submitted method will be evaluated using two reconstruction errors, "landmark reconstruction error" and "pixel reconstruction error".
- The "landmark reconstruction error" is defined as average Euclidean distance, between the ground-truth-reconstructed frame and the prediction-reconstructed frame, averaged over a set of predefined anatomical landmarks in each scan.
- The "pixel reconstruction error" is defined as the same average Euclidean distance, between the ground-truth-reconstructed frame and the prediction-reconstructed frame, averaged over all pixels of all but the first frames in each scan.

Thus, each submitted method should output four sets of displacement vectors:
- Global displacement vectors for pixel reconstruction (number of the "GP" vectors = number of pixels in the entire scan)
- Global displacement vectors for landmark reconstruction (number of the "GL" vectors = number of defined landmarks in the scan)
- Local displacement vectors for pixel reconstruction (number of the "LP" vectors = number of pixels in the entire scan)
- Local displacement vectors for landmark reconstruction (number of the "LL" vectors = number of defined landmarks in the scan)

Based on these four output vector sets, four evaluation metrics will be used in this challenge:
- Global pixel reconstruction error (GPE), the pixel reconstruction error using GP vectors.
- Global landmark reconstruction error (GLE), the landmark reconstruction error using GL vectors.
- Local pixel reconstruction error (LPE), the pixel reconstruction error using LP vectors.
- Local landmark reconstruction error (LLE), the landmark reconstruction error using LL vectors.

The landmark is defined based on anatomical structures, such as vessel branches, bony structures and other ad

hoc landmarks. It is estimated that between 10-20 landmarks will be available for each scan, but this is subject to further verification. Further details and summary statistics of the landmarks will be made available by the challenge commence. The final score on the four evaluation metrics will be averaged over all scans in the test set.

Runtime will be considered as additional evaluation metric which is the consumed time of predicting positions for all frames in a scan, averaged over all scans in the test set.

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

<Rationale in using Euclidean distance-based error metrics, as opposed to errors defined in transformation parameter space> Direct measuring the accuracy of parameters of transformation matrix is difficult since the contribution and weighting between rotation and translation components can be sensitive to experimental and imaging configurations, such as reference coordinates and definition of rotation axis, and dependent on application. In this challenge, metrics based on distance error is preferred, which is arguably considered a more practical type of measure to reflect the difference between ground truth and prediction in Euclidean space (Luo et al. 2023).

<Justification of using displacement-based representation of transformation, as opposed to rigid- or affine matrices, as the algorithm output> Although the ground-truth are available in the form of rigid transformation, we argue that, from experience in developing similar numerical algorithms, enforcing the algorithm output to be homogeneous transformation is not only unnecessary, but sometimes misleadingly encourages a more numerically challenging solution due to issues such as gimbal lock in using rotation matrix, local minima in numerical optimisation. In addition, the displacement-based representation allows flexibility for a quantitatively more accurate reconstruction, with a near-rigid transformation, which could be sufficient for clinical use. However, there is no requirement in the internal methodology adopted, for example, the submitted algorithm can convert a single estimated rigid transformation matrix to all these four types of required displacement vectors as output.

<Justification of the local and global reconstruction errors> The reconstructed scan from either local level or global level displacement are capable of representing different types of reconstruction performance, such as frame-level reconstruction error and accumulated error of the algorithm (Li. et al. 2023, Wein et al. 2020). To streamline the evaluation, other monotonic metrics, such as final drift and Dice overlap, albeit commonly reported in literature (e.g. Li et al 2023), are not included. However, in practical applications, one might choose to reconstruct a sequence of ultrasound frames (as opposed to the entire scan or two adjacent frames, which are represented by local and global errors, respectively), using a pre-optimised sequence length that is most suitable to the downstream application. Without specifying a single target clinical application, these two adopted local and global errors shall represent the range of accuracy between the choices of the reconstruction length.

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.

First, for each individual metric, the values for all submissions will be normalised to the range [0, 1] by the value of baseline method, and then the final score will be generated using the formula below:

Overall score = 0.25*(1-GPE*) + 0.25*(1-GLE*) + 0.25*(1-LPE*) + 0.25*(1-LLE*)

where * indicates the normalised reconstruction errors. The "overall score" within the range of [0,1] will be used to produce the final rank for all the submitted algorithms. The final score will be reported with 3 decimal places and the higher the better. For the submissions with the same final score, the rank will be generated based on the runtime. A smaller runtime will be awarded a higher rank. A maximum runtime will be imposed for challenge submissions, benchmarked as the speed of our baseline methods, to encourage usability in the clinical applications. All the raw values for the defined metrics will also be made available.

We also report four other categories of scores, global reconstruction score = 0.5*(1-GPE*) + 0.5*(1-GLE*), local reconstruction score = 0.5*(1-LPE*) + 0.5*(1-LLE*), landmark reconstruction score = 0.5*(1-GLE*) + 0.5*(1-LLE*) and pixel reconstruction score = 0.5*(1-GPE*) + 0.5*(1-LPE*). These are provided for reference and research interest without formal ranking.

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

The submitted algorithms will be run on a docker container such that this case is not likely to happen. In case such event occurs, the minimum score (0) will be given to any case where the code cannot run or the metric cannot be computed successfully.

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

All the metrics are normalized to a common scale so that the metrics with larger values will not dominate. The two levels of measurement with two types of displacement are assumed equally important to make a desirable reconstruction performance. Therefore, we use equal weights to generate a benchmark for the challenge.

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.

Missing data is unlikely to happen as the docker will be used during test phase (Type 2 challenge). In case this event occurs, the missing data raised either from incorrectly running code or unable calculated metrics would be given the minimum score (0).

Statistical power analysis is used to decide test data size to decrease the possibility of making a Type I & II errors in hypothesis testing. The Cohen's D value is used to calculate the effect size, in which the system error from optical tracker (0.25 mm) is regarded as the difference between two means from two groups, and the standard deviation (0.46 mm) comes from the baseline results in Li et al. 2023. We conducted a statistical power analysis for a t-test with statistical significance of 0.05 and statistical power of 0.9. The estimated test data size is 31 such that the test data size is set at 32 (384 scans in total). This ensures that only 10% probability of encountering a Type II error and 5% probability of making a Type I error.

For analysis, we will conduct statistical tests on results from the top five teams and compare them with baseline methods, while these tests will not impact the overall ranking of these submissions, but will be reported.

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

We will ignore the missing data to avoid making skewness to the computed means of the performance metrics.

The commonly used Cohen's D value is used to conduct statistical analysis, which is a popular effect size measure in statistical analyses (Cohen, 1988).

The statistical tests between submissions and baselines indicate if the results are statistically significant.

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.

We will provide two baseline algorithms based on Prevost et al. 2018 and Li et al. 2023, trained using the training data.

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.

Chen et al., "Determination of scan-plane motion using speckle decorrelation: Theoretical considerations and initial test," International Journal of Imaging Systems and Technology, vol. 8, no. 1, pp. 38-44, 1997.

Cohen, "Statistical power analysis for the behavioral sciences (2nd ed.)," Hillside, NJ: Lawrence Erlbaum Associates, 1988.

Guo et al., "Sensorless freehand 3D ultrasound reconstruction via deep contextual learning," In Medical Image Computing and Computer Assisted Intervention, pp. 463-472, 2020.

Guo et al., "Ultrasound volume reconstruction from freehand scans without tracking," IEEE Transactions on Biomedical Engineering, vol. 70, no. 3, pp. 970-979, 2022.

Hu et al., "Freehand ultrasound image simulation with spatially-conditioned generative adversarial networks," in Molecular imaging, reconstruction and analysis of moving body organs, and stroke imaging and treatment, pp. 105-115, 2017.

Lasso et al., "Plus: open-source toolkit for ultrasound-guided intervention systems," IEEE Trans. Biomed. Eng., vol. 61, no. 10, pp. 2527-2537, 2014.

Li et al., "Trackerless freehand ultrasound with sequence modelling and auxiliary transformation over past and future frames," in 2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI), pp. 1-5, 2023.

Li et al., "Long-term dependency for 3D reconstruction of freehand ultrasound without external tracker," IEEE Transactions on Biomedical Engineering, 2023.

Luo et al., "Self context and shape prior for sensorless freehand 3D ultrasound reconstruction," In International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 201-210, 2021.

Luo et al., "Deep motion network for freehand 3D ultrasound reconstruction," In International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 290-299, 2022.

Luo et al., "RecON: Online learning for sensorless freehand 3D ultrasound reconstruction," Medical Image Analysis, vol. 87, pp. 102810, 2023.

Mikaeili et al., "Trajectory estimation of ultrasound images based on convolutional neural network," Biomedical Signal Processing and Control, vol. 78, pp. 103965, 2022.

Miura et al., "Pose estimation of 2D ultrasound probe from ultrasound image sequences using CNN and RNN," In Simplifying Medical Ultrasound: Second International Workshop, ASMUS, pp. 96-105, 2021.

Ning et al., "Spatial position estimation method for 3D ultrasound reconstruction based on hybrid transfomers," In 2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI), pp. 1-5, 2022.

Prager et al., "Sensorless freehand 3-D ultrasound using regression of the echo intensity," Ultrasound in medicine & biology, vol. 29, no. 3, pp.437-446, 2003.

Prevost et al., "3D freehand ultrasound without external tracking using deep learning," Medical Image Analysis, vol. 48, pp. 187-202, 2018.

Wein et al., "Three-dimensional thyroid assessment from untracked 2D ultrasound clips," In International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 514-523, 2020.

Further comments

Further comments from the organizers.

All organizing members have a strong background in trackerless freehand US reconstruction, ranging from 3 to 20 years. Several members have experience of event coordination demonstrated in the previous MICCAI events, such as 2020-2023 MICCAI ASMUS Workshop, MICCAI SIG-MUS, and MICCAI mu-RegPro Challenge.

Our goal is to provide high quality, open-source data for freehand US reconstruction to the community. As this challenge is set as an open call, the publicly available data and code will still contribute to the community for future research.