MR to Ultrasound Registration for Prostate Challenge: Structured description of the challenge design

CHALLENGE ORGANIZATION

Title

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

MR to Ultrasound Registration for Prostate Challenge

Challenge acronym

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

mu-RegPro (to be shown as greek letter Mu)

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.

Multimodal image registration between pre-operative and intra-operative imaging enables the fusion of clinically important information during many surgical and interventional tasks. The registration of magnetic resonance imaging (MRI) and transrectal ultrasound (TRUS) images assists prostate biopsy and focal therapy, arguably having transformed prostate cancer patient care to a less invasive and more localized diagnostic, monitoring and treatment pathway. Though, even with great progress having been made by the community in the past two decades, challenges remain in this application. First, paired MRI and TRUS data from a sizable patient cohort are not routinely stored in clinical practice, and publicly-accessible data is scarce and low-quality. Second, annotating anatomical and pathological landmarks on both images - critical in representing corresponding locations for validation - requires expert domain knowledge and experience from multiple disciplines including urology, radiology and pathology.

In addition to its prevalence-warranted clinical importance, this is also a unique application that saw a wide range of registration algorithms proposed and housed intriguing debates such as rigid-versus-nonrigid and FLE-versus-TRE. Both feature- and intensity-based classical methods and unsupervised or segmentation-driven learning methods have been innovated with some most technically interesting approaches in the field such as biomechanical regularisation and statistical motion modelling.

The mu-Reg challenge aims to provide well-curated, yet real-world clinical data, with more than a hundred paired MR and TRUS images, annotated carefully by researchers and clinicians with more than 15 years of experience working with this application. The outcome of the challenge includes one of the first multimodal imaging data, facilitated with expert annotations for validation, for benchmarking advancement in registration methodology, as well as for future research in managing the most common non-skin cancer in men.

Challenge keywords

List the primary keywords that characterize the challenge.

registration, multimodal, deformable, prostate, MRI, ultrasound

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.

In discussions to be held in conjunction with ASMUS.

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.

20-30 attendees, with 10-15 entries (based on previous registration challenges at MICCAI and other venues, e.g. Learn2Reg, BRATS-Reg)

Publication and future plans

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

We aim to submit a publication which summarizes the dataset and results from the challenge. Members of the top five teams will be invited as co-authors.

We will additionally encourage participants to submit papers of their contributions and/or any novel methodologies for medical image registration.

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.

We will make use of the grand-challenge.org website to host the challenge, evaluation engine and leaderboards.

Training of algorithms will be run on the participants' computing infrastructure.

Docker containers for testing (Type 2 challenge) will be run on grand-challenge.org.

No onsite computing challenge is planned. However, for the in-person portion (results, presentations, etc.) we will require AV equipment (projector(s), microphone(s), and speakers.

TASK: Prostate MRI to TRUS Registration

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.

Same as Challenge abstract.

Keywords

List the primary keywords that characterize the task.

registration, multimodal, deformable, prostate, MRI, ultrasound

ORGANIZATION

Organizers

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

Zachary M. C. Baum, MSc, University College London
Shaheer U. Saeed, BEng, University College London
Zhe Min, PhD, University College London
Yipeng Hu, PhD, University College London
Dean C. Barratt, PhD, University College London

b) Provide information on the primary contact person.

Zachary Baum, zachary.baum.19@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

One time 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

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

Will be provided should the challenge be accepted and placed on grand-challenge.org.

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 and not listed in leaderboard.

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

Successful participants will receive certificates of participation.

The first-place and runner-up participants will receive additional certificates.

We have one agreed sponsor, and we are actively seeking additional sponsorship so that we may be able to provide additional prizes as well.

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 results will be announced publicly unless errors were made in the data processing. Teams will be permitted to make multiple distinct submissions (must not be a simple change of hyperparameters). The leaderboards will be available to view publicly.

The top 5 teams will be invited to present their work at the challenge event in a 10-15 minute presentation.

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).

As mentioned previously; we aim to submit a publication which summarizes the dataset and results from the challenge. Members of the top five teams will be invited as co-authors.

We encourage participants to submit their contributions and/or any novel methodologies for medical image registration. However, participants should only refer to the specific challenge results (e.g. if published before. they

should only discuss their methodology, but not the dataset and results used/obtained in the challenge) once the challenge paper is published (submission to arXiv is considered as a sufficient waiting period). Participants should cite the challenge paper once it is published if their work has not been published already.

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.

Participants will be asked to Dockerize their trained network/algorithm/method and will submit these to grandchallenges.org for evaluation and ranking on the validation and test sets at the appropriate times in the 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.

We plan to release access to evalua tion on a small validation set ahead of the release of the test set. This evaluation process will allow participants to tune their methods on unseen data and to ensure they avoid any errors with image orientation or preprocessing.

Validation data will be released publicly instead of through an evaluation period on grand-challenges.

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

Registration Opens: Monday, April 3 2023

Training Data Release: Monday, April 3 2023

Validation Data Release: Monday, June 5 2023

Test Data Evaluation Period Begins: Monday, July 3 2023

Baseline Performance Release: Monday, July 3 2023

Final Evaluation Deadline: Monday, July 17 2023

Winners Announced & Speaker Invitations: Monday, July 24 2023

Presentation of Results, Challenge Talks, Prizes: MICCAI 2023

(Subject to change based on MICCAI 2023 deadlines)

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).

We have received approval from London-Dulwich Research Ethics Committee to use these images for research purposes. The approval reference number is REF 14/LO/0830.

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.

Additional comments: The data shall be available for use exclusively for research purposes, due to the restrictions from original ethics approval and patient consent (https://clinicaltrials.gov/ct2/show/record/NCT02341677?view=record).

The training and validation data will remain publicly available after the completion of the challenge. The training and validation data may be used within the research remit of this challenge, and in further research-related publications. The training and validation data is not to be used commercially. However, if the desired use is unclear, the organizers ask that those accessing the data refrain from further use or distribution outside of this challenge.

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 to evaluate the registration algorithms (validation and test sets) will be made publicly at the same time that that section of the challenge becomes available to enter on the grand-challenges platform.

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

Participating teams' are encouraged, but not required to make their code publicly available. We will provide links to any available source code on the challenge 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.

We have no conflicts of interest to declare.

All challenge and task organizers will have access to the 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, 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

Registration.

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.

Patients with mpMRI who will undergo an MRI-targeted TRUS-guided prostate biopsy procedure to assist in the diagnosis and staging of their prostate cancer.

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

Subjects of SmartTarget Biopsy Clinical Trial (Hamid et al., 2019). 141 men who had undergone a prior (positive or negative) transrectal ultrasound biopsy and had a discrete lesion on mpMRI (Likert score 3–5) requiring targeted transperineal biopsy were enrolled at University College London Hospital, London, UK; 129 underwent both biopsy strategies (visual registration and image fusion) and completed the study.

Of the 141 patients, 108 pairs had US images acquired in the transverse plane during their biopsy, and 33 in the sagittal plane. The 108 pairs which were acquired in the transverse plane, as well as their paired MRIs, are included in the challenge.

Imaging modality(ies)

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

T2 Weighted MRI

Transrectal Ultrasound (TRUS)

Context information

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

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

Anatomical landmarks which were identified in each pair of MRI and TRUS images, such as lesions, zonal structures, water-filled cysts, and calcifications, as well as prostate gland masks will be provided in the training data.

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

No further information is given on a per-patient basis.

However; summary statistics, such as age, PSA, lesion volume, and Gleason scores for all patients are available in (Hamid et al., 2019). Notably, all patients had prior MRI imaging scored by experienced radiologists ahead of their targeted biopsy procedure.

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 prostate gland and surrounding structures/tissues are shown in both the MRI and TRUS images for each patient.

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.

Accurate alignment of the prostate gland and other anatomical structures, such as those provided as landmarks for localizing relevant anatomical and potentially pathological targets during guided prostate biopsies.

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)).

Runtime, Accuracy, Robustness.

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).

MRI: As specified in the clinical trial protocols (https://clinicaltrials.gov/ct2/show/record/NCT02341677?view=record), a mixture of 1.5T and 3T MRI scanners at University College London Hospital were used for the T2-weighted MR imaging, given that the mpMRI data were read by the radiologists with any discrete lesion visible scoring 3, 4 or 5 on PI-RADs scale.

TRUS: a standard clinical ultrasound machine (HI-VISION Preirus, Hitachi Medical Systems Europe) equipped with a bi-plane (C41L47RP) transperineal probe, at University College London Hospital.

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

MRI: Patients underwent mpMRI as the standard of care in accordance with the British Society of Urogenital Radiology and European Society of Urogenital Radiology standards in sequences as described in (Ahmed et al., 2017). Only the T2-weighted sequences are provided for the challenge. All MRI volumes are of size 120 x 128 x

128.

TRUS: A range of 57–112 TRUS frames were acquired in each case by rotating a digital transperineal stepper (D&K Technologies GmbH, Barum, Germany) with recorded relative angles covering the majority of the prostate gland, using a standard clinical ultrasound machine (HI-VISION Preirus, Hitachi Medical Systems Europe) equipped with a bi-plane (C41L47RP) transperineal probe at a depth of 65 mm and a frequency of 9.0 MHz. These parasagittal slices were then used to reconstruct a 3D volume. Software was written in C++ to retrieve and reconstruct data stored by the TRUS scanner that specify the pixel intensity values for each radial B-mode scanline, the physical length of the scanline (in millimeters), the in-plane angle of the field-of-view of the B-mode images, the angular positions of each B-mode image plane, the coordinates of the origins of the scanlines, and the center of rotation of the transducer. Using these data, the position vector of any point identified in a reconstructed TRUS volume can be determined with respect to a local image coordinate system. A 3D TRUS image of the scanned volume is reconstructed by interpolating the TRUS intensity values at measured locations to calculate the intensity values across a rectangular grid. All reconstructed TRUS volumes are of size 81 x 118 x 88.

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 MRI and TRUS images were acquired at the University College London Hospital, London, UK.

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).

MRI: Patients underwent mpMRI as the standard of care in accordance with the British Society of Urogenital Radiology and European Society of Urogenital Radiology standards in sequences as described in (Ahmed et al., 2017).

TRUS: All the surgeons who participated in the Smart Target BIOPSY study were Urology Fellows with at least 6 months of training in transperineal targeted biopsies. All had been assessed and approved as independently competent. Each had performed approximately 50 or more procedures and most 100 to 200 procedures. Additionally, each surgeon used a rotating digital transperineal stepper (D&K Technologies GmbH, Barum, Germany) to acquire the TRUS images using a standard clinical ultrasound machine (HI-VISION Preirus, Hitachi Medical Systems Europe) equipped with a bi-plane (C41L47RP) transperineal probe.

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, validation, and test cases comprise paired MRI and TRUS volumes. Each will be accompanied by a

complete prostate gland segmentation, as well as at least three additional anatomical landmark annotations (as described in 23).
Using intensity-based methods, feature-based methods, or some combination of the two, participants' should provide a function which accepts:

Input: a moving image, a fixed image, and a moving label.

And produces:

Output: a warped moving label, and a dense deformation field (DDF).

Participants are permitted to deform or transform the images in any manner (e.g. parametric or non-parametric transformation) which they should choose, however; they must provide an equivalent DDF for purposes of metric computation on the test data.

While the test data will not have segmentation labels publicly visible, participants may derive segmentations as part of their inference process.

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

Training Set - 65

Validation Set - 8

Test Set - 35

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

The total is based on the availability of the data for the dataset. Training, validation, and test sets are chosen to approximate a 60% / 5% / 35% split. This will ensure that the validation set can be used to help ensure the correct functioning of the submitted methods and that the testing data set can remain sufficiently large.

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.

Data distributions in the three datasets are randomly assigned in order to represent a distribution of a real-world dataset which ensures consistency between all three datasets.

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.

Given that this task encompasses unsupervised transform prediction, there is no ground truth transformation which may be used to compute a loss function value. However, methods based on similarity measures from the images or provided segmentations, or other approaches, may be used to drive the learning process.

For the provided annotations, the prostate gland segmentations on the MRI images were acquired as part of the Smart Target clinical trial protocols (Donaldson et al., 2017).

The gland segmentations on the TRUS images were manually edited by two medical imaging research fellows based on automatically contoured prostate glands on original TRUS slices obtained using the method proposed in (Ghavami et al., 2018).

Besides full gland segmentations for all cases, the landmarks include apex, base, urethra, visible lesions, junctions between the gland, gland zonal separations, vas deferens and the seminal vesicles, and other patient-specific point landmarks such as calcifications and fluid-filled cysts were manually defined. In total, there are 417 landmark pairs, where each patient has a minimum of 3 landmarks.

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.

Experienced radiologists and medical imaging researchers were asked to locate candidate landmarks in the image pairs, where each landmark was reviewed and agreed upon by at least one other annotator.

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 original observers are three biomedical imaging researchers with more than 5 years experience in this application, for manual segmentation of both image types and identifying corresponding anatomical and pathological landmarks (as detailed in Hu et al. 2018). One observer with more than 10 years experience performed a final additional quality control for all the cases, for released challenge data. All observers went through a two-day advanced course for radiologist, hosted by the urology department and radiology department at University College London Hospitals, training for this specific multimodal imaging application.

The method proposed in (Ghavami et al., 2018) was used to annotate the TRUS prostate gland boundaries.

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.

All annotations were agreed upon by at least two annotators as opposed to any merging or averaging of the annotations.

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 image volumes (MRI and TRUS) were resampled to 0.8mm^3 isotropic voxel size, which is sufficent for developing and validating registration algorithms.

The TRUS volumes were centre cropped to a field of view which includes the required anatomical structures.

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.

While inter- and intra-observer error and variability is possible, it is minimized by the use of consensus labeling techniques during the annotation process.

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

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)

1) Dice Similarity Coefficient (DSC): computed between the warped mask of the TRUS image and the mask of the MRI image over the prostate gland boundary; averaged over all cases in the test set; in the range [0, 1] where higher indicates better registration

2) Robustness of DSC (RDSC): DSC averaged over 68% of the highest-DSC cases in the test set

3) Target Registration Error (TRE): registration error based on the l1 norm for anatomical landmarks which are indicated by regions identifiable in both sets of imaging, such as zonal structures, waterfilled cysts, and calcifications. A lower TRE between the TRUS and MRI images indicates better registration. TRE is computed per-case, where the TRE for each case if the root-mean-square (RMS) TRE over all landmarks for that case. The mean of the TRE over all cases is then normalised to the range [0, 1] by assuming unregistered images have maximum TRE. This maximum TRE is obtained by determining the maximum individual landmark pre-registration TRE in the test set. If any submitted aggregate TREs are higher than the maximum individual landmark pre-registration TRE, the value will be clipped to 1.

4) Robustness of TRE (RTRE): TRE averaged over 68% of lowest-TRE cases in the test set

5) Robustness of Targets (RTs): registration error based on the l1 norm for 3 lowest-error landmarks out of 5 total landmarks between the TRUS and MRI images; averaged over all cases in the test set

6) 95th Percentile Hausdorff Distance (95%HD): 95th percentile of the distances between boundary points in one set to the nearest point in the other set where sets are based on organ boundary points from TRUS and MRI image segmentations; averaged over all cases in the test set; normalised to the range [0, 1] by assuming unregistered images have maximum 95%HD (if any cases submitted with worse performance than unregistered then values are clipped to the range [0, 1]).

7) Standard Deviation of log Jacobian Determinant (StDJD): computed using the jacobian over the deformation

field; averaged over all cases in the test set

8) Runtime: the amount of time it takes to compute the warped image using the algorithm; computed over all cases and then averaged to obtain the average per-case runtime.

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

1) Dice Similarity Coefficient (DSC): DSC reflects the volumetric overlap between the registered image pair which is desirable in use cases such as intra-operative augmented reality where preoperative images (or features from these images) may be registered and overlaid onto the intraoperative images (Bianchi et al., 2021); furthermore, DSC is capable of reflecting both size and localization agreement between the registered image pair which is useful for medical imaging applications such as intraoperative lesion localization using registration of intra-operative and preoperative images (Zijdenbos et al., 1994) (Bertels et al. 2019)

2) Robustness of DSC (RDSC): we chose to compute robustness metrics to ignore the effect of outliers if they are present; the value of 68% was chosen as we assumed computed metric values to be normally distributed so it then follows that approximately 68% of the errors must lie within one standard deviation of the mean error, which we deem to be sufficient to ignore the effect of outliers

3) Target Registration Error (TRE): TRE specifically allows us to evaluate registration performance for the specific features of interest in the biomedical application (Datteri et al. 2012); in this challenge, these include apex, base, urethra, visible lesions, junctions between the gland, gland zonal separations, vas deferens and the seminal vesicles, and other patient-specific point landmarks such as calcifications and fluid-filled cysts. Furthermore, TRE may also indicate physical plausibility e.g. DSC may be very high but points at the features of interest may not be properly aligned, which is why TRE supplements DSC as a metric for evaluating registration performance

4) Robustness of TRE (RTRE): see point 2

5) Robustness of Targets (RTs): see point 2; additionally, accounting for robustness over targets allows us to ignore some labelling inconsistencies in the data that may have occurred due to errors and variability in human labels

6) 95th Percentile Hausdorff Distance (95%HD): 95%HD indicates the 95th percentile of distances between the boundaries of the organs of interest, which is important in clinical applications such as for intraoperative guidance (Bianchi et al., 2021)

7) Standard Deviation of log Jacobian Determinant (StDJD): StDJD indicates the smoothness of deformation which is may indicate the physical plausibility of the deformations (Kabus et al., 2009) (Leow et al., 2007)

8) Runtime: the runtime is important in our application of interest, which is preoperative to intraoperative registration, where ideally real-time registration would be desirable to generate any overlays of the preoperative images onto intraoperative images (Bianchi et al., 2021)

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.

We compute an overall score for each team using the formula below:

0.2*(DSC) + 0.1*(RDSC) + 0.3*(1-TRE) + 0.1*(1-RTRE) + 0.1*(1-RTs) + 0.2*(1-95%HD)

The overall scores will be in the range [0, 1] where higher is better. We will report these to 3 decimal places. Rankings will be generated based on these scores. In the event of the same scores, we will award a higher rank based on the StDJD and award. In the case of a further tie, we will award a higher rank to the submission with a smaller runtime.

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

Since we plan to run a Type 2 challenge, where we run the participants' code in a docker container, we do not expect this scenario to arise. In the unlikely event of such a scenario, the minimum score (0) will be given for any case which does not run correctly or for which metrics are unable to be calculated.

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

In our overall score computation we weight TRE the highest, with DSC and 95%HD having lower weights. TRE is weighted the highest since it is of most relevance in the clinical application of intraoperative guidance indicating error in alignment between areas of interest and registration of landmarks of interest (Bianchi et al. 2021). We include 95%HD in our weighting since it ensures registration of boundaries which is of interest for intraoperative guidance applications where overlays of pre-operative images may be placed onto intraoperative images (Bianchi et al. 2021). The robustness measures are weighted least since we want to ensure good performance over the entire test set while at the same time minimising effects of outliers e.g. nearly perfectly registered cases or cases left completely unregistered.

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.

Given the Type 2 nature of the challenge, and that all image resolutions are identical between the training, validation, and test sets, this is unlikely to occur in this challenge. However, for missing data, the minimum score (0) will be given for any case which does not run correctly or for which metrics are unable to be calculated.

To ascertain the number of samples to be placed in the test set we conduct a simple statistical power analysis. We assume that the computed metric values are normally distributed (see below). Our null hypothesis is that there is no difference between the mean metric value computed between TRUS and MRI images when unregistered vs when registered. This null hypothesis allows us to investigate whether the registration algorithm can register images such that it changes the DSC to reflect the registration. We use Cohen's D value with the desired effect size of 0.8 for the comparison, which is generally accepted (Cohen, 1988), to perform our power analysis (Lakens, 2013). We use a p-value threshold of 0.05 for the t-test as the threshold for statistical significance and we want our study to be 90% powered. Using these parameters, we conduct a statistical power analysis for a t-test to compute the estimated sample size. Using these parameters, we estimate the sample size to be 33.8 which is why we include 35 samples in our test set. The statistical power of 90% would lead to only a 10% probability of encountering a

Type II error i.e. not rejecting the null hypothesis when there is a significant effect. And due to the p-value threshold being set to 0.05 this can be interpreted as a 5% probability of making a Type I error i.e. rejecting the null hypothesis if it were true. We thus conclude that the number of test cases is sufficient for our use case.

The assumption of normality of the computed metrics should also be tested since that is the basis for the metric design and the above statistical power analysis. We conduct the Shapiro-Wilk test for normality on the metrics for a small subset (24 randomly sampled images) of our data (un-registered) and find p-values in the range of 0.05 to 0.07. We then compute the same metrics for TRUS images registered to MRI images using a baseline algorithm LocalNet (Hu et al., 2018) and get p-values in the range of 0.05 to 0.06. Thus the null hypothesis that the computed metrics are from a normally distributed population, cannot be rejected.

For analysis, we will also conduct statistical tests (t-tests) between the top five submissions to compare them with each other and with the baselines provided, however, these tests will not impact the overall rankings of submissions.

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

For missing data handling, we suggest ignoring missing values so as not to skew the computed means of the performance metrics.

For the statistical power analysis, we use Cohen's D value which is a commonly used effect size measure in statistical analyses that involve comparisons made using t-tests (Cohen, 1988) (Lakens, 2013).

For the assumption of normality, we use the Shapiro-Wilk test (Shapiro and Wilk, 1965) since it is deemed to have the highest statistical power compared to other tests for normality with the same sample size based on Monte-Carlo simulations (Nornadiah et al. 2011). We only use a limited subset of samples since for larger sample sizes the Shapiro-Wilk test can be sensitive to even trivial deviations from normality (Field, 2009). We use 24 samples based on a statistical power analysis with desired effect size = 0.8, p-value threshold = 0.05, and desired statistical power = 0.75.

The statistical tests between submissions and to compare them to baselines, allow us to further evaluate if any performance improvements are meaningful over the entire test set.

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 the following weakly-supervised baseline algorithms, trained using the training data:

VoxelMorph (Balakrishnan et al., 2019)

LocalNet (Hu et al., 2018)

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.

Ahmed et al., Diagnostic accuracy of multi-parametric MRI and TRUS biopsy in prostate cancer (PROMIS): a paired validating and confirmatory study. The Lancet 389(10071), pp. 815-822. 2017.

Balakrishnan et al., VoxelMorph: a learning framework for deformable medical image registration. IEEE Transactions on Medical Imaging 38(8), pp. 1788-1800. 2019.

Bertels et al., Optimizing the dice score and jaccard index for medical image segmentation: Theory and practice. International conference on medical image computing and computer-assisted intervention. Springer, Cham. 2019.

Bianchi et al., The Use of Augmented Reality to Guide the Intraoperative Frozen Section During Robot-assisted Radical Prostatectomy. European Urology 80.4, pp. 480-488. 2021.

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

Datteri et al. Estimation and reduction of target registration error. International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, Berlin, Heidelberg. 2012.

Donaldson et al., The SmartTarget Biopsy Trial: a prospective paired blinded trial with randomisation to compare visual-estimation and image-fusion targeted prostate biopsies. The Journal of Urology 197(4), pp. E425. 2017. Field. Discovering statistics using SPSS (3rd ed.). SAGE Publications. 2009.

Ghavami et al., Automatic slice segmentation of intraoperative transrectal ultrasound images using convolutional neural networks. SPIE Medical Imaging: Image-Guided Procedures, Robotic Interventions, and Modeling 10576. 2018.

Hamid et al., The SmartTarget Biopsy Trial: A Prospective, Within-person Randomised, Blinded Trial Comparing the Accuracy of Visual-registration and Magnetic Resonance Imaging/Ultrasound Image-fusion Targeted Biopsies for Prostate Cancer Risk Stratification. European Urology 75(5), pp. 733-740. 2019.

Hu et al., Weakly-supervised convolutional neural networks for multimodal image registration. Medical Image Analysis 49, pp. 1-13. 2018.

Kabus et al., Evaluation of 4D-CT lung registration. Med. Image Comput. Assist. Interv. Springer, pp. 747–754. 2009.

Lakens et al., Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. Front. Psychol Sec. Cognition 4, pp. 863. 2013.

Leow et al., Statistical properties of Jacobian maps and the realization of unbiased large-deformation nonlinear image registration. IEEE transactions on medical imaging 26.6, pp. 822-832. 2007.

Nornadiah et al., Power comparisons of Shapiro–Wilk, Kolmogorov–Smirnov, Lilliefors and Anderson–Darling tests. Journal of Statistical Modeling and Analytics. 2 (1), pp. 21–33. 2011.

Shapiro and Wilk, An analysis of variance test for normality (complete samples). Biometrika. 52 (3–4), pp. 591–611. 1965.

Zijdenbos, et al., Morphometric analysis of white matter lesions in MR images: Method and validation. IEEE Trans. Medical Imaging 13(4), pp. 716–724. 1994.

Further comments

Further comments from the organizers.

All members of the organizing team have a strong background in intensity-based and feature-based registration, between 3 and 20 years, especially as it relates to prostate cancer-related image analysis problems. Additionally, the organizers have distinct experience in the clinical translation of learning-based methods, computer vision, and machine learning for medical imaging.

Several members of the organizing team have extensive experience with event coordination, especially within MICCAI-based settings (e.g. 2020-2022 MICCAI ASMUS Workshop, MICCAI SIG-MUS)

It is also of note that the overarching goal of this challenge is to provide high-quality, open research data for multimodal registration problems to the community. While we list this challenge as a one-time with a fixed end date, we foresee this data providing value long after the challenge has been completed, with the possibility of additional challenges and data being released in the future as the community and our access to high-quality data grows.