Learn2Reg - The Challenge (2023): Structured description of the challenge design

CHALLENGE ORGANIZATION

Title

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

Learn2Reg - The Challenge (2023)

Challenge acronym

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

Learn2Reg

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.

Medical image registration plays a very important role in improving clinical workflows, computer-assisted interventions and diagnosis, and research studies involving e.g. morphological analysis. Besides ongoing research into new concepts for optimisation, similarity metrics, domain adaptation and deformation models, deep learning for medical registration is currently starting to show promising advances that could improve the robustness, generalisation, computational speed and accuracy of conventional algorithms to enable better practical translation. Nevertheless, before Learn2Reg there was no commonly used benchmark dataset to compare stateof-the-art learning-based registration among another and with their conventional (not trained) counterparts. Since last year, the Learn2Reg challenge is a type 2 challenge meaning that only submissions containing the algorithm (e.g. Dockers) are allowed. This facilitates reproducibility and further use of the algorithms in the research community.

The most prevailing unmet requirements for clinical adoption of medical image registration are two-fold: 1) limited generalisation of one registration model to an unseen task, 2) limited clinical focus of evaluation metrics. We strongly believe that the extensions we propose for our challenge design in 2023, can tackle both aspects and further bridge the gap between research and practical use.

We could demonstrate initial successes in the area of developing self-configuring registration algorithms that can automatically be trained and optimised on a variety of unseen datasets. This step is crucial, since challenge datasets can never comprehensively mirror all aspects of a certain clinical application and for registration to become a universal tool the hyper-parameter choices should become part of an automatic pipeline. Several methods were submitted that produced robustly good results on the three hidden datasets. However, a number of algorithms - especially deep-learning-based ones - still had problems working equally well on all datasets.

In terms of evaluation metrics, our previous designs had either limited clinical relevance (e.g. anatomical

landmarks in lung scans are only a proxy for relevant tumour lesions) or limited coverage, i.e. using either target registration error or Dice overlap exclusively cannot necessarily detect deterioration in quality outside the supervised structures.

For this reason, we extend the tasks from 2022 in certain aspects of Learn2Reg 2023 with a particular focus on expanding the evaluation metrics, adding more hidden datasets and simplifying the setup of self-configuring registration algorithms.

Innovations in Task 1 and 3.
In Task 1, the challenge is divided into two phases: In phase 1, participants train/tune their algorithms locally and submit the algorithms via grand-challenge. The best teams of this phase are invited to participate in phase 2. In phase 2, the participants submit a training docker that will be run by the organizers on a larger dataset that includes additional annotations that are not publically available. The trained networks will be made available via grand-challenge. Specifically the expanded datasets for lung / throrax registration will comprise: anatomical landmarks, geometric keypoint correspondences, therapeutically relevant target structures and semantic anatomical segmentations.

We improve Task 3 from L2R 2022 by introducing changes to facilitate implementing and testing a selfconfiguration registration docker. Innovations include a set of baseline algorithms to guide implementation, an additional dummy dataset with supervision and evaluation configuration files to fully comprehend our evaluation method, as well as predicted semantic label information at test time. We further introduce measures to facilitate a successful submission, including the possibility of sanity checks and a improved timeline for implementation and refinement.

Challenge keywords

List the primary keywords that characterize the challenge.

registration, deformable, thorax, abdomen, brain, oncology, multimodal, realtime

Year

The challenge will take place in ...

2023

FURTHER INFORMATION FOR MICCAI ORGANIZERS

Workshop

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

none

Duration

How long does the challenge take?

Full 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.

We expect 40-70 attendees (based on our 2019 MICCAI tutorial and Learn2Reg challenge at MICCAI 2020,2021,2022), with approx. 20 entries

Publication and future plans

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

We will encourage all participants to submit a short (4-8 pages) LNCS paper of their contribution in particular of any novel approaches to medical image registration. Furthermore, we aim to submit two separate publications for task 1 and task 3 to summarize the results.

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. Docker containers for training will be either run on grand-challenge or on local hardware for Task 1 and local hardware for Task 3. No onsite challenge is planned, but participants will get the opportunity to upload their methods at least twice to be evaluated on a small subset of the test data in order to avoid any pitfalls (wrong orientation, etc.).

TASK: Thoracic Image Registration for Lung Cancer

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 new task focusses on the highly clinically relevant task of lung nodules tracking in longitudinal CT series and the alignment of further therapeutic structures. For lung screening and in more general follow-up assessment of metastasis, the findings of the baseline image (nodules and metastasis) have to be found on the follow-up image in order to detect changes in size, etc.. Finding matching nodules/lesions completely manually is time-consuming and prone to error, especially in the lungs, because structures look very similar. For this reason, registration can simplify the reporting process. When considering radiotherapy the propagation of planning scans to intraoperative cone-beam CT becomes relevant and here not only the lung parenchyma but also surrounding organsat-
risk have to be accurately aligned, e.g. trachea, esophagus, spinal cord and heart.

In this task, the challenge is divided into two phases: In phase 1, participants train/tune their algorithms locally and submit the algorithms via grand-challenge. The best teams (at least 5 teams) of this phase are invited to participate in phase 2. In phase 2, the participants submit a training docker that will be run by the organizers on a larger dataset that includes additional annotations that are not publically available. The trained networks will be made available via grand-challenge.

Keywords

List the primary keywords that characterize the task.

intra-patient, CT, lung, cancer, deformable registration

ORGANIZATION

Organizers

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

Mattias Heinrich (Uni Lübeck), Malte Sieren (University Clinic Schleswig-Holstein) Christoph Großbröhmer (Uni Lübeck), Alessa Hering (Radboudumc and Fraunhofer MEVIS)

b) Provide information on the primary contact person.

Alessa Hering (Radboudumc and Fraunhofer MEVIS)

Life cycle type

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

Examples:

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

Repeated event as open call challenge.

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).

https://learn2reg.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.

Phase 1: no additional training data is allowed
Phase 2: Docker submission

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.

Organisers and team members of organisers may participate and are ranked but cannot win prizes.

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

(tba)

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 an obvious error was made in data processing. Participating teams may choose to submit multiple methods, given there are sufficiently distinct and not simply other hyperparameters and provided that each algorithm is described to clarify differences. Organisers reserve the right to remove lower scoring duplicate submissions from the same team of algorithms that are deemed too similar.

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).

Up to two team members qualify as author for joint publication. Participants are free to publish their own results separately, but can only make reference to overall results once the challenge paper is published (submission to arXiv is considered as a sufficient waiting period).

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.

Validation of phase 1: Docker container containing the trained network.
Phase 2: Docker container containing the training scripts.

We encourage participants to only submit containers that can be run in a rootless mode to increase the usability of the submitted Docker containers on computational servers.

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

Participants will get the opportunity to upload their dockers to be evaluated on the validation data of phase 1 in order to avoid any pitfalls (wrong orientation, etc.). We also provide our evaluation scripts to test them on the training data.

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

February 2023: registration to challenge open
Early March 2023: release of training data of phase 1, release of evaluation metrics and instructions for docker submission
July 2023: evaluation of phase 1 as part of online meet-up with Q&A; invitation to best five teams to train their algorithm on larger dataset
August 2023: final evaluation
MICCAI 2022: presentation of results

Ethics approval

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

If necessary (e.g. not the case for already open sourced data) ethics approval will be requested prior to releasing any new data.

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: to be confirmed

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 software, code to evaluate accuracy, complexity etc will be made publicly available.

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 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.

Data providers who annotate test cases have access to test labels (of specific sub-task), organisers who implement evaluation metrics and scripts will have partial access to test 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

Screening, Longitudinal study, Treatment planning, Assistance.

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 included in lung screening programs, lung tumour follow-up patients and patients undergoing concurrent radiochemotherapy

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

Subjects of a lung screening trial (NLST). NLST enrolled 53,454 current or former heavy smokers ages 55 to 74. Participants were randomly assigned to receive three annual screens with either low-dose helical CT or standard chest X-ray. In addition we include up to 20 NSCLC patients that received pre-treatment fan-beam 4DCT for

planning and intra-operative cone-beam 4DCT.

Imaging modality(ies)

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

Computed Tomography (CT), 4DCT

Context information

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

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

Automatic keypoint correspondences and lung masks will be provided for all training data. For the additional radiotherapy data, segmentations of organs-at-risk (heart, esophagus, spinal cord, trachea and other structures) will be provided.

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

no further information are given

Target entity(ies)

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

lungs and thorax shown in CT data

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.

Highly accurate alignment of inner and outer lung structures: lobes, trachea, fissures, vessels and airways as well as plausible deformations with spatial smoothness (low standard deviation of Jacobian determinants).

Main goal is the accurate tracking/propagation of the lung nodules to the follow-up scan.

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, Complexity, Reliability.

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 data was aquired in 32 medical centers across the US. Therefore, a variety of device types were used. (See https://cdas.cancer.gov/nlst/) 4DCT: 16-slice, helical CT scanner (Brilliance Big Bore, Philips Medical Systems) with a slice thickness of 3 mm and 512×512 axial resolution

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

Low-dose CT scan of the thorax during inspiration and extreme phases of 4DCT.

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.

The used data is selected out of the NLST data that is aquired at 33 U.S. medical centers.

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).

not applicable

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.

One case is a pair of baseline and follow-up thorax CT scans or two phases of a 4DCT sequence of the same patient respectively.

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

Phase 1:
training: 110
validation: 22

Phase 2:
training: 1300
validation: >100
test: >200

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

The participants get a small but sufficient number of training cases to set up a training procedure. The trained networks will be evaluated on the phase 1 validation cases to select the best 5 algorithms.

Those algorithms will be invited to submit a docker container for training their network, which will be run by the challenge organisers on a hidden larger dataset that includes data and annotations that cannot be made publicly available.

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.

We aim to select a diverse dataset regarding scanner, gender, age, location of nodules in the lung, etc.

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.

The annotators segmented the lung nodules using CIRRUS lung screening (https://www.diagnijmegen.nl/software/cirruslungs/). The centers of gravity used as landmarks in this study was derived from the segmentation. See details at https://wiki.cancerimagingarchive.net/pages/viewpage.action?pageId=21267414

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.

--

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.

A large number of subjects annotated the scans in the NLST study (see https://cdas.cancer.gov/nlst/)

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.

not applicable each case was annotated by one/same rater

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.

Common pre-processing to same voxel resolutions and spatial dimensions as well as affine pre-registration will be provided to ease the use of learning-based algorithms for participants with little prior experience in image registration. For NLST cases lung masks will be provided to enable the exclusion of outer lung-structures, which increase the problem complexity (e.g. by sliding motion). For 4DCT thorax registration both inner- and outer-lung

structures are of relevance and a larger mask is provided. Here the provided RT struct slice ROIs are firstly converted into volumetric segmentation masks and mapped into the correct scan resolution. The missing order of structures are brought into a new set of labels that are usable for deep learning algorithms.

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 landmarks used for evaluation are derived from the center of gravity of the nodule segmentation mask. If a nodule grows/shrinks irregularly, the CoG can change its position accordingly. Therefore, an error of 1mm are expected. The manual segmentation masks can contain annotation errors of also approx. 1mm.

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

not applicable

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) TRE of landmarks located at the center of gravity of the lung nodules
2) Robustness: 30% highest TRE of all cases
3) TRE of manually annotated landmarks in the lung
4) Dice Score of lung lobe / organ-at-risk segmentations
5) SD (standard deviation) of log Jacobian determinant
6) Run-time computation time (only awarded when evaluation server is used with provided docker container)

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

DSC or TRE respectively measure accuracy; HD95 measures reliability; Outliers are penalised with the robustness score (30% of lowest mean DSC or 30% of highest mean TRE) The smoothness of transformations (SD of log Jacobian determinant) are important in registration, see references of Kabus and Leow

Run-time computation time is relevant for clinical applications.

After accuracy, time and smoothness metrics are converted into significance ranks for each team and task, we employ a geometric average across those scores. This way outlier solutions that may e.g. provide good accuracy at the cost of very little smoothness will be penalised more severely as compared to a simple arithmetic mean (see Ranking methods)

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.

All metrics but 4) (robustness) use mean rank per case (ranks are normalised to between 0.1 and 1, higher being better). For multi-label tasks the ranks are computed per structure and later averaged. As done in the Medical Segmentation Decathlon we will employ "significant ranks" http://medicaldecathlon.com/files/MSD-Rankingscheme.pdf

Across all metrics an overall score is aggregated using the geometric mean. This encourages consistency across criteria. The time ranks are only considered with 50% weight (since not all participants are able to use docker containers for evaluation).

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

Missing results will be awarded the lowest rank (potentially shared and averaged across teams).

With this approach, we want to keep the entry barrier for new participants as low as possible so that they can submit their methods to only some tasks. Nevertheless, we will additionally perform an evaluation of complete submissions for the gain of information.

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

The geometric mean encourages consistency across criteria. A ten-fold difference between highest and lowest score is fixed to be independent of number of participants.

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/submission will result in lowest rank for this case.
Ties will result in average rank among all equal participants.

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

The geometric mean is more robust against outliers, hence methods that perform well on all metrics are encouraged.

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 several baseline algorithms to compare new methods against, there are:
- PDD-Net (MICCAI '19) unsupervised training
- corrField
- Voxelmorph (CVPR'18) with and without label supervision
- Deeds
- Elastix, NiftyReg, and/or ANTs (where applicable)

TASK: Multi Task 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.

This task replicates the challenge of L2R 2020+2021 and serves as a continuation of the benchmark on intrapatient abdominal MR-CT alignment, inter-patient abdominal CT mapping, lung CT registration and inter-patient whole brain mapping (OASIS MRI). Detailed information about these applications can also be found in our recent comparison journal paper: https://ieeexplore.ieee.org/document/9925717. To lower the entry barrier for researchers that are new to the field, this task also allows for submission of displacement fields and single-task solutions (trained offline). However, to win this challenge task a docker submission is obligatory.

Keywords

List the primary keywords that characterize the task.

intra-patient, CT, MR, multimodal, lung, inter-patient, brain

ORGANIZATION

Organizers

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

Mattias Heinrich (Uni Lübeck), Christoph Großbröhmer (Uni Lübeck), Alessa Hering (Radboudumc and Fraunhofer MEVIS)

b) Provide information on the primary contact person.

Christoph Großbröhmer (Uni Lübeck)

Life cycle type

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

Examples:

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

Repeated event as open call challenge.

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).

https://learn2reg.grand-challenge.org/

Participation policies

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

Fully automatic.

Additional points: for each task, a different set of hyperparameters and algorithmic choice is possible and can be user defined

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.

Private data is allowed.

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

Organisers and team members of organisers may participate and are ranked but cannot win prizes.

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

(tba)

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 an obvious error was made in data processing. Participating teams may choose to submit multiple methods, given there are sufficiently distinct and not simply other hyperparameters and provided that each algorithm is described to clarify differences. Organisers reserve the right to remove lower scoring duplicate submissions from the same team of algorithms that are deemed too similar.

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).

Up to two team members qualify as author for joint publication. Participants are free to publish their own results separately, but can only make reference to overall results once the challenge paper is published (submission to arXiv is considered as a sufficient waiting period).

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.

Either submission of displacement fields through Grand-Challenge or submission of docker through grandchallenge (required for time bonus)

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

Participants will get the opportunity to upload their dockers to be evaluated on the validation data of phase 1 in order to avoid any pitfalls (wrong orientation, etc.). We also provide our evaluation scripts to test them on the training data.

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

September 2021: all training and validation data for multi-task registration (Abdomen MR-CT, Lung CT and OASIS) is already available
February 2023: registration to challenge open
July 2023: optional evaluation of validation results and discussion at WBIR 2022
August 2023: final evaluation
MICCAI 2023: presentation of results

Ethics approval

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

all data is already open source with appropriate ethics approval provided

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: to be confirmed

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 software, code to evaluate accuracy, complexity etc will be made publicly available.

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 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.

Data providers who annotate test cases have access to test labels (of specific sub-task), organisers who implement evaluation metrics and scripts will have partial access to test 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, Diagnosis.

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.

Variety of patient cohorts that can benefit from improved medical image registration, e.g. (multimodal) imageguided interventions, image guided-radiotherapy planning, lung diagnostics, etc.

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

Abdomen CT-MR: Patients from an colorectal cancer chemotherapy trial, the baseline sessions of the abdominal CT scans were randomly selected from metastatic liver cancer patients; the remaining scans were acquired from a retrospective post-operative cohort with suspected ventral hernias. Additional hidden dataset of general population study with whole-body MRI.

Lung CT: Patients with clinical routine follow-up CT lung examinations (inspiration-inspiration), that may have been part of lung screening cancer trials (e.g. Nelson study and National Lung Screening Trial) as well as inspiration-expiration CT scan pairs from patients with breathing disorders (COPD, etc.) The data will be collected from Radboud University Medical Centre, Nijmegen

OASIS: The dataset consists of T1w MRIs acquired as part of the OASIS dataset project (https://www.oasisbrains.org/), which contains hundreds of neuroimaging datasets, with subjects ranging from 42-95 years old and covering normals as well as subjects in various stages of cognitive decline.

Imaging modality(ies)

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

Computed Tomography (CT), Magnetic Resonance Imaging (MRI)

Context information

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

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

Abdomen CT-MR: 8 training pairs with four (each) manually annotated abdominal organs, additional 30 unpaired CTs and 30 unpaired MRIs with same structural annotations
Lung CT: 20 training pairs with automatic keypoint correspondences (>2000 per case), binary lung masks OASIS MRI: 416 3D MR scans with 35 anatomical labels

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

no further information are given

Target entity(ies)

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

abdominal organs in MRI and CT, lungs in CT, brain in MRI

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.

Abdomen MR-CT: Alignment of abdominal organs (liver, kidneys and spleen) and plausible transformations (low standard deviation of Jacobian determinants). Secondary goal alignment of smaller anatomies (pancreas, aorta, etc).
Lung CT: Alignment of inner lung structures: lobes, fissures, vessels and airways, determined by 100 manual landmark pairs for each of the 10 test cases and plausible transformations (low standard deviation of Jacobian determinants).
OASIS MRI: Inter-subject alignment of anatomical 35 brain structures and plausible transformations (low standard deviation of Jacobian determinants).

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)).

Applicability, Runtime, Accuracy, Robustness, Complexity.

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).

Abdomen MR-CT: CT scanners across different centres including additional CT data from the Vanderbilt University Medical Center (VUMC) and different clinical MRI scanner with additional scans from population study
Lung CT: Philips Brilliance 16P or Philips Mx8000 IDT 16
OASIS MRI: MRI-T1w volumes involving 1.5T or 3T Siemens scanners.
More information can be found here: https://www.oasis-brains.org/files/oasis_cross-sectional_facts.pdf

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

Abdomen MR-CT: data collected from TCGA-KIRC, TCGA-KIRP and TCGA-LIHC (see further details below) varying CT and MR imaging protocols, usually CT around ~1mm resolution, MRI anisotropic with larger slice thickness.
Lung CT: The inspiration scans will be acquired using a low-dose protocol (30 mAs) while the expiration scan with ultra-low-dose (20 mAs). The scanner could e.g. be a Philips Brilliance 16P with slice thickness of 1.00 mm and slice spacing of 0.70 mm. Pixel spacing in the X and Y directions may vary from 0.63 to 0.77 mm with an average value of 0.70 mm.
OASIS MRI: Data is either acquired or reshaped to 1x1x1 mm isotropic resolution of T1w-weighted MRI whole brain scans. Sequence details: MP-RAGE TR (ms) 9.7, TE (ms) 4.0, Flip angle 10°, TI (ms) 20, TD (ms) 200, Sagittal Orientation, Thickness 1.25 mm, gap 0 mm.

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.

Abdomen MR-CT: The Cancer Genome Atlas Kidney Renal Clear Cell Carcinoma [TCGA-KIRC], The Cancer Genome Atlas Cervical Kidney renal papillary cell carcinoma [KIRP], The Cancer Genome Atlas Liver Hepatocellular Carcinoma [TCGA-LIHC]
Lung CT: The data will be provided by Radboud University Medical Centre Nijmegen,
OASIS MRI: The data is acquired at WUSTL Knight ADRC, the processed data will be provided by MIT / MGH (Boston).

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).

not applicable

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.

Abdomen MR-CT one case is a pair of MRI and CT scan of the same patient with manual annotations, potentially before/after surgery, each additional case is an unpaired MRI or CT scan with manual annotations.
Lung CT: one case is a pair of inspiration and expiration CT with manual annotations, additional training data is provided by preprocessing the public DIR-Lab to same dimensions which comprise 10 4DCT sequences with at least 300 landmark pairs each.
OASIS MRI: one case is a single MRI scan with segmentation information of 35 anatomical brain structures

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

Abdomen MR-CT: training: 8 + 60, test: 8
Lung CT: training 20+10, test 10
OASIS: training 416, test 38

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

Abdomen MR-CT: Collecting paired multimodal data and manually annotating them is difficult and timeconsuming, therefore the use of additional unpaired (labelled) MR-CT data can be explored by participants. The number of test cases (8) is sufficient to establish a ranking of substantially different accurate methods, but will likely lead to equal ranks (due to limited significance) for similar performant algorithms.
Lung CT: 60 training scans (30 pairs) is still a fairly low number, but by providing very dense supervision (keypoint correspondences) we are confident that registration networks can be trained successfully.
OASIS MRI: for inter-subject registration 416 training pairs enable 172'640 registration pairs which is an ideal case for supervised learning. The number of test subjects was restricted to limit the required transfer of displacement fields but still provides suitable variability for robust significance testing.
Data augmentation can help to overcome limitations of small datasets (cf. K Eppenhoff and J Pluim "Pulmonary ct registration through supervised learning with convolutional neural networks" TMI 2018).

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.

Abdomen MR/CT: In this intra-subject registration task, a number of paired MR/CT scans can be directly employed for supervised training. Cross-domain learning can be an integral part of this challenge task with the provision of unpaired CT and MRI scans.
OASIS MR: In this inter-subject registration task, all potential pairs can be employed for training. To limit the amount of test transformations to be processed, we will announce around 100 randomly selected pairs of test subjects that should be registered for evaluation. To measure inverse consistency of algorithms, we will include a subset of bi- directive cases).

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.

Abdomen MR/CT: MRI and paired MRI/CT: manual 3D voxel segmentation of at least four abdominal organs: spleen, right kidney, left kidney, liver from experienced graduate student with 3+ years experience in medical imaging.

Additional CT: Thirteen abdominal organs were considered regions of interest (ROI), including spleen, right kidney, left kidney, gall bladder, esophagus, liver, stomach, aorta, inferior vena cava, portal and splenic vein, pancreas, left adrenal gland, and right adrenal gland. The organ selection was essentially based on [Shimizu A, Ohno R, Ikegami T, Kobatake H, Nawano S, Smutek D. Segmentation of multiple organs in non-contrast 3D abdominal CT images. International Journal of Computer Assisted Radiology and Surgery. 2007;2:135–142.]. As suggested by a radiologist, the heart was excluded for lack of full appearance in the datasets, and instead the adrenal glands were included for clinical interest. These ROIs were manually labeled by two experienced undergraduate students with 6 months of training on anatomy identification and labeling, and then verified by a radiologist on a volumetric basis using the MIPAV software.Same raters produced the anatomical landmarks for both the training and testing data.

Lung CT: 50-100 manual selected corresponding landmark pairs (within the same subject, primarily at vessel/airway bifurcations) have been annotated by the experienced graduate students from medical informatics degrees (following in principle: K. Murphy et al: "Semi-automatic Reference Standard Construction .." MICCAI 2008), additional automatic keypoint correspondences will be provided.

OASIS MRI: Automatic Freesurfer segmentations of 40 small structures (subcortical and deep brain) with manual verification. The software employed were FreeSurfer 7.1 and SAMSEG. We fused these probabilistic label maps in both subject space and atlas space, and manually verify each of the 400 scans. For Learn2Reg, we will distribute the original images, the processed images, and the fused segmentation label maps for the 350 training and validation subjects.

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.

see above

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.

Abdominal MR/CT: 3 experienced medical imaging researcher,
Lung CT: manual landmarks: two experienced graduate students, for keypoints: corrField
https://grandchallenge.org/algorithms/corrfield/
Hansen, Lasse, and Mattias P. Heinrich. "GraphRegNet: Deep Graph Regularisation Networks on Sparse Keypoints for Dense Registration of 3D Lung CTs." IEEE Transactions on Medical Imaging (2021).
OASIS: FreeSurfer with manual QC

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.

not applicable each case was annotated by one/same rater

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.

Abdomen MR/CT: The scans have been modified for direct usage in registration and deep learning algorithms: We have reorientated the data, resampled it to an isotropic resolution of 2 mm, and used cropping, padding and affine pre-alignment to achieve voxel dimensions of 192x160x192.
Lung CT: Common pre-processing to same voxel resolutions and spatial dimensions as well as affine preregistration will be provided to ease the use of learning-based algorithms for participants with little prior experience in image registration. Lung masks will be provided to enable the exclusion of outer lung-structures, which increase the problem complexity (e.g. by sliding motion)
OASIS MRI: Common pre-processing to same voxel resolutions, intensity normalization, and spatial dimensions as well as affine pre-registration will be provided to ease the use of learning-based algorithms for participants with little prior experience in image registration.

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.

Abdomen CT/MR (train/test): Dice overlap of >90% is expected for inter-rater agreement of smaller structures and >95% for large organs. Lung CT (train/test): based on previous studies inter-rater variabilities of 0.5-1.0mm are expected for manual landmarks and 1.0-1.5mm for automatic keypoint correspondences OASIS MRI (train/test) an agreement of 75-85% with manual annotation is expected. Previous studies indicate inter-rater agreement of more than 75% Dice overlap across all structures, but these vary dramatically. Certain deep structures (like the Thalamus) lacks significant contrast, whereas cortex lacks sufficient resolution.

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

not applicable

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)

Lung CT:
1) TRE of landmarks located at the center of gravity of the lung nodules
2) Robustness: 30% highest TRE of all cases
3) TRE of manually annotated landmarks in the lung

4) Dice Score of lung lobe / organ-at-risk segmentations
5) SD (standard deviation) of log Jacobian determinant
6) Run-time computation time (only awarded when evaluation server is used with provided docker container)

Abdomen CT/MR and OASIS MR:
1) DSC (Dice similarity coefficient) of segmentations
2) HD95 (95% percentile of Haussdorff distance) of segmentations
3) Robustness: 30% lowest DSC of all cases
4) SD (standard deviation) of log Jacobian determinant
5) Run-time computation time (only awarded when evaluation server is used with provided docker container)

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

see above

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.

see above

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

see above

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

see above

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.

see above

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

see above

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.

see above

TASK: Self Configuring Meta Learning

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 task expands on Task3 from L2R 2022 and deals with the importance of developing fully automatic selfconfiguring methods that learn their hyperparameters based on training and validation data and require no user interaction. Any algorithm suitable for medical image registration which can be inferred on all Task 3 datasets is considered a valid submission and by our terminology "self-configuring". The "self-configuration" might be pursued in a rule-based (i.e. CT or MRI preprocessing) or empiric (i.e. hyperparameter selection by evaluation of metrics) fashion.

Medical image registration covers multiple complementary anatomical sights, modalities and deformation types. One-fit-all models rarely reach top performance across multiple tasks. Many state-of-the-art algorithms require multiple hyper-parameters that are specifically chosen for each challenge. We aim to find the best self-configuring registration algorithm that can automatically be trained and optimised on a variety of hidden datasets. Possible solutions include modular architectural designs that enable adaptive changes in cost function and optimisation strategy or hyper-networks that can be efficiently fine-tuned with few parameters after initial training. Participants are encouraged to upload a single docker container that automatically learns a registration model on any given training/validation dataset (e.g. but not limited to multimodal abdominal fusion, intra-patient lung CT and followup as well as inter-subject alignment).

We provide a standardised data format and folder structure inspired by the MedicalDecathlon and the nnUNet framework and samples of closely related tasks and a detailed description of the dimensions, spacings, expected range of deformation and available type of supervision (keypoints, segmentations and/or none) for all hidden tasks.

For this year, we improve this task by introducing changes to facilitate further implementing and testing a selfconfiguring registration docker and adopting to the standardised data format: In the first place, we provide multiple ready-to-docker baseline algorithms for training and inference as an example for implementation and replication of validation results. Secondly, we have already published an additional Abdomen-MR registration dataset which includes test data (including labels) and evaluation configurations to fully replicate our docker training, testing and evaluation pipeline. Since many image registration solutions make use of some kind of label/keypoint supervision, we further offer these for our hidden datasets (including corrfield-generated keypoints and nnUNet-generated segmentations for keypoint and label supervision respectively). This allows participants to focus on the image registration technique itself, lowers the complexity of implementation, might give additional insight for the comparison of registration algorithms independent of the quality supervision used, and reduces repetitive energy-consumption heavy computations.

Furthermore, we will enforce and release a detailed information policy regarding preprocessing of our hidden datasets (e.g. modality-dependent normalisation techniques) as well as relevant information to estimate training and testing expenditure (e.g. maximum number of labels, possibility of large datasets).

We also adopted changes in scheduling this task by extending the developement timeframe and will encourage participants to submit the docker for a sanity check before the final deadline. We will offer at least one round of feedback to improve their training docker (e.g. to improve on certain tasks that underperformed). Information shared with particular participants will be published for all teams, ensuring information equity among participants.

Keywords

List the primary keywords that characterize the task.

automatic design choices, hyper-parameter search, multi-task

ORGANIZATION

Organizers

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

Mattias Heinrich (Uni Lübeck), Christoph Großbröhmer (Uni Lübeck), Alessa Hering (Radboudumc and Fraunhofer MEVIS)

b) Provide information on the primary contact person.

Mattias Heinrich (Uni Lübeck)

Life cycle type

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

Examples:

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

Repeated event as open call challenge.

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).

https://learn2reg.grand-challenge.org/

Participation policies

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

Fully automatic.

Additional points: self configuring methods for hidden registration tasks

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.

Docker containers.

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.

Organisers and team members of organisers may participate and are ranked but cannot win prizes.

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

(tba)

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 an obvious error was made in data processing. Participating teams may choose to submit multiple methods, given there are sufficiently distinct and not simply other hyperparameters and provided that each algorithm is described to clarify differences. Organisers reserve the right to remove lower scoring duplicate submissions from the same team of algorithms that are deemed too similar.

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).

Up to two team members qualify as author for joint publication. Participants are free to publish their own results separately, but can only make reference to overall results once the challenge paper is published (submission to arXiv is considered as a sufficient waiting period).

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.

Intend to submit and pre-evaluation is requested before July 2023 to ensure enough processing hardware for hyper-parameter search.

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

Participants of Task 3 should submit results for the snapshot evaluation of Task2 at WBIR 2023 in early July. A

compulsory submission for the validation of results on Task 2 is mandatory before July 2023 with a note to the organisers that participation in the docker training of Task 2 is intended. The best teams on this pre-evaluation (at least five teams) will be invited to submit dockers for hidden multitask evaluation. The pre-selection is to ensure that enough processing hardware for hyper-parameter search is available.
We further offer a sanity check prior to the final docker submission to reduce pitfalls (e.g. incorrect hardware specifications).

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

February 2023: registration to challenge open important no training data will be released (hidden learning)
July 2023: compulsory evaluation of validation results on Task 2, invitation to best five teams to submit dockers for hidden multitask evaluation
Late August 2023: submission of docker for sanity checks
September 2023: submission of docker
MICCAI 2023: evaluation and presentation of results

Ethics approval

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

ethics approval exist but no data will be shared

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)

Additional comments: no data has to be shared

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 software, code to evaluate accuracy, complexity etc is already publicly available.

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 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.

Data providers who annotate test cases have access to test labels (of specific sub-task), organisers who implement evaluation metrics and scripts will have partial access to test 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

Data reduction, Education, Research, Training.

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.

By developing self-configuring registration tools the outcome could potentially applied to any 3D image registration task.

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

Challenge participants do not know what 3D image data is used to learn self-configuring methods or use other meta-learning strategies for fully automatic hyper-parameter setting. The data always contains a set of 3D training and validation scans of same dimensions and sizes, it may contain an ROI mask, it may contain pointcorrespondences, it may contain segmentation labels.

Imaging modality(ies)

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

unknown to challenge participants (US, MRI, CT)

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 exact challenge will be unknown to participants, however, the dimensions of datasets and modality types for each task will be encoded in a dataset.json file.

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

no further information are given

Target entity(ies)

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

unknown to challenge participants

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.

Ability to generalise across unknown 3D registration tasks only based on the hidden validation set. It potentially has to deal with inter-subject and multimodal registration. That means the algorithm has to be able to

automatically find suitable hyper-parameters within a reasonable training time (as a rule of thump: training different network architectures for 24 hours per task is considered okay, but early stopping is encouraged). Ideally methods are used that enable hyper-parameter tuning after the main network is trained (e.g. conditional or hyper networks)

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)).

Applicability, Usability, Integration in workflow, Robustness, Complexity, Reliability.

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 data may comprise different MRI sequences, CT scans as well as 3D ultrasound

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

hidden from participants, the data will be pre-processed to same dimensions

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.

hidden from participants

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).

not applicable

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.

Each case will consists of a scan, potentially an ROI mask (also available for test/validation), potentially multi-label segmentation (only available for training/validation), potentially keypoint/landmark locations. The whole dataset will contain information about voxel dimensions and the case IDs to be registered. We will also provide computergenerated supervision in terms of nnUNet-segmentations and corrfield-keypoints for test data to facilitate implementation.

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

the number of training scans will range from 20 to 100 with a varying number of anatomical labels

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

Some variability of training sizes (for individual sub-tasks) will be included in the hidden datasets to evaluate the ability of self-configuring methods to avoid over-fitting and reach generalisation in different scenarios

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.

Within our hidden training, validation and test sets, a comparable variability of scanner types and sequences is given. To enable the automatic hyper-parameter tuning the same folder structure and naming conventions are used for all datasets. This will be directly compatible with the nnUNet framework (imagesTr/img0001_0000.nii.gz labelsTr/img0001.nii.gz dataset.json etc). Masks and supervision that are available at test time are provided additionally (masksTs/predictedlabelsTs/keypointsTs).

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.

some exact details have to be hidden to avoid informing participants about the hidden tasks, but annotation guidelines are in general comparable to the six previous Learn2Reg challenge tasks.

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.

see above

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.

depending on dataset, either one or two annotators with experience in medical image analysis

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.

not applicable each case was annotated by one/same rater

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 scans have been modified for direct usage in registration and deep learning algorithms: We have reorientated the data, resampled it to an isotropic resolution, and used cropping, padding and affine pre-alignment to achieve same voxel dimensions (within each task)

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.

similar annotation errors as described for the previous six Learn2Reg challenge tasks are expected

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

not applicable

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)

see Task 1 and 2

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

see above

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.

see above

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

see above

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

see above

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.

see above

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

see above

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.

In addition to the above-mentioned baseline algortihms, we provide ready-to-docker baseline algorithms and will report their results.

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.

Learn2Reg 2020 and 2021 challenge overview paper in TMI (https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9925717)
Newly added 4DCT dataset: https://doi.org/10.1016%2Fj.ijrobp.2012.12.023