Endoscopic Vision Challenge: Structured description of the challenge design

CHALLENGE ORGANIZATION

Title

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

Endoscopic Vision Challenge

Challenge acronym

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

EndoVis

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.

Minimally invasive surgery using cameras to observe the internal anatomy is the preferred approach to many surgical procedures. Furthermore, other surgical disciplines rely on microscopic images. As a result, endoscopic and microscopic image processing as well as surgical vision are evolving as techniques needed to facilitate computer assisted interventions (CAI). Algorithms that have been reported for such images include 3D surface reconstruction, salient feature motion tracking, instrument detection or activity recognition. However, what is missing so far are common datasets for consistent evaluation and benchmarking of algorithms against each other. As a vision CAI challenge at MICCAI, our aim is to provide a formal framework for evaluating the current state of the art, gather researchers in the field and provide high quality data with protocols for validating endoscopic vision algorithms.

Challenge keywords

List the primary keywords that characterize the challenge.

Surgical Vision, Endoscopy, Classification, Segmentation

Year

The challenge will take place in ...

2020

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.

50 (based on numbers from previous EndoVis challenges).

Publication and future plans

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

The joint publications will be coordinated by the particular sub-challenge organizers.

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.

Depends on the specific sub-challenges, last year two sub-challenges used the DREAM/synapse platform for example.

TASK: CATARACTS - Surgical Workflow Analysis

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.

Surgical microscopes or endoscopes are commonly used to observe the anatomy of the organs in surgeries. Analyzing the video signals issued from these tools are evolving as techniques needed to empower computerassisted interventions (CAI). A fundamental building block to such capabilities is the ability to automatically understand what the surgeons are performing throughout the surgery. In other words, recognizing the surgical activities being performed by the surgeon and segmenting videos into semantic labels, that differentiates and localizes tissue types and different instruments, can be deemed as an essential steps toward CAI. The main motivation for these tasks is to design efficient solutions for surgical workflow analysis, with potential applications in post-operative analysis of the surgical intervention, surgical training and real-time decision support. Our application domain is cataract surgery. As a challenge, our aim is to provide a formal framework for evaluating new and current state-of-the-art methods and gather researchers in the field of surgical workflow analysis.

Analyzing the surgical workflow is a prerequisite for many applications in computer assisted interventions (CAI), such as real-time decision support, surgeon skill evaluation and report generation. To do so, one crucial step is to recognize the activities being performed by the surgeon throughout the surgery. Visual features have proven their efficiency in such tasks in the recent years, thus, a dataset of cataract surgery videos is used for this task. We have defined twenty surgical activities for cataract procedures. This task consists of identifying the activity at time t using solely visual information from the cataract videos. In particular, it focuses on the online workflow analysis of the cataract surgery, where the algorithm estimates the surgical phase at time t without seeing any future information.

Keywords

List the primary keywords that characterize the task.

surgical activity recognition, cataract surgery

ORGANIZATION

Organizers

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

Hassan Al Hajj, Ph.D., Inserm UMR 1101, Brest France
Gwenolé Quellec, Ph.D., Inserm UMR 1101, Brest France
Pierre-Henri Conze, Ph.D., IMT Atlantique, LaTIM UMR 1101, UBL, Brest France
Mathieu Lamard, Ph.D., Univ Bretagne Occidentale, Inserm UMR 1101, Brest France
Béatrice Cochener, M.D., Ph.D. Univ Bretagne Occidentale, Inserm UMR 1101, Service d'Ophtalmologie, CHRU Brest, Brest France

b) Provide information on the primary contact person.

Hassan ALHAJJ, hasalhajj@gmail.com

Life cycle type

Define the intended submission cycle of the challenge. Include information on whether/how the challenge will be continued after the challenge has taken place.

Examples:

- One-time event with fixed submission deadline
- Open call
- Repeated event with annual fixed submission deadline

Open call.

Challenge venue and platform

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

MICCAI.

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

grand-challenge.org

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

Part of https://endovis.grand-challenge.org/, Sub-challenge: https://cataracts2020.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.

Publicly available 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.

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.

The winner will receive an award. We will provide details on our sponsorship two months before the conference.

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.

Participants waive the rights to decide whether to show their results on the results page. Every time new scores are submitted for an algorithm, the last and the best scores for that algorithm will be displayed on the results page.

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).

Participants are allowed to publish their own results separately only after publication of a challenge paper (expected mid 2021).

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.

Docker containers with specific input/output protocols will be submitted by participants and benchmarked internally by us. Test data wont be available to the public. Submission instructions to be defined.

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.

Validation set will be provided and erroneous docker submissions will be reported to participants after benchmarking the docker containers. To that end, submitted docker containers will be evaluated on a very small subset of the testing set (e.g. 50 images) and reported back for sanity check. Multiple submissions are allowed for the validation phase. Only one submission per algorithm is allowed for the test set.

Challenge schedule

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

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

Release date of the training and validation data: April 1, 2020
Registration date: March 1, 2020
Release date of test data: September 21 2020
Submission deadline: September 27 2020

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).

Study approved by the IRB of Brest University Hospital on 28 janvier 2013. Informed consent was obtained from all patients.

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: Additional comments: We will be sending a document, describing the terms and conditions of the challenge, to the participants which must sign it before their registration is validated.

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.

We will provide Python code for the evaluation metrics used to benchmark the docker instances so that it can be used for local validation.

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

Participants decide whether to publish their code or not.

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.

If the community plan to collect share new annotations for the dataset, they must do it through the CATARACTS website. We are not going to release the test case labels. When the conference is finished, we will have a webpage dedicated for evaluating the test set in order to help the community to evaluate new methods.

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

Research, Assistance, Intervention planning, Surgery, Training.

Task category(ies)

State the task category(ies).

Examples:

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

Temporal Segmentation.

Cohorts

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

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

The final application should be using video streams of cataract surgeries performed in hospitals.

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

The challenge is based on video streams data of cataract surgeries performed in Brest University Hospital.

Imaging modality(ies)

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

Surgical microscope videos.

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 videos show the surgical field of the cataract surgeries where only the anterior segment of the eye and the tool tips are present.

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

Patients were 61 years old on average (minimum: 23, maximum: 83, standard deviation: 10). There were 38 females and 12 males.

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.

Eye shown in cataract surgery videos.

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.

Surgeon activities.

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)).

Specificity, Sensitivity, Precision.

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).

Surgeries were performed under an OPMI Lumera T microscope (Carl Zeiss Meditec, Jena, Germany). Videos were recorded with a 180I camera (Toshiba, Tokyo, Japan) and a MediCap USB200 recorder (MediCapture, Plymouth Meeting, USA).

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

When the surgeon starts the intervention, a simple push to a button starts the recording of the surgical field.

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.

Brest University Hospital, France.

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).

Surgeries were performed by three surgeons: a renowned expert, a one-year experienced surgeon and an intern.

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 are videos of cataract surgery. Twenty surgical activities have been identified. Each frame of the videos has a list of twenty classes, indicating in which activities the surgeon is at.

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

50 cataract surgeries have been recorded.

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

The dataset was divided into a training set (25 videos), a validation set (5 videos) and a test set (20 videos). Dataset was divided in a way to have enough videos for training, thus the 25 videos for it. To select the best algorithms, we provide a validation dataset of 5 videos. The participants can check the performance of their algorithms using the validation set. 20 test cases to show the complexity of the problem, thus better evaluation accuracy of the algorithms.

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.

Division was made in such a way that 1) each activity appears in the same number of videos from the training and test sets (plus or minus one/two).

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.

Surgery experts have annotated key events throughout the surgery. We deduced from these events the start/end timestamps of the activities.

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

The annotation is performed by a web application build for this task. All annotators are trained by an experienced ophthalmologist to recognize properly the key events we want to annotate.

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.

Annotators were two non-M.D. cataract surgery experts trained by an experienced ophthalmologist.

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.

Annotations from both experts were adjudicated. Disagreements on the events were automatically detected. In case of disagreement, experts watched the video together and jointly determine the actual events happening.

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.

No pre-processing steps.

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 most relevant error sources might be related to the start/end timestamps of an activity because it depends on the definition of the activity and the images quality at the moment of the start/end the activity.

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

N/A

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)

F1-score.

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

It is widely used in the field when the problem is defined as multi-class classification and segmentation.

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.

Metric-based aggregation: for each participant, we compute a score for each case, then, a mean score over all cases. Participants will be ranked by decreasing order of the mean score. In case of a tie, participants will be assigned the same rank.

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

As the submissions will be done through docker containers, we do not expect missing results. We should also note that participants will get evaluation results on the validation set after submitting their docker image that can serve as a sanity check to make sure the docker image works correctly.

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

As mentioned, this metric summarises model performance by taking into account true, false and missed detections.

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.

t-test on the scores (a score per test case) is used to assess the variability of rankings.

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

Paired sample t-test is the standard statistical test to determine whether the mean difference between two sets of observations (one observation per test case in our case) is zero. It should be noted, however, that two assumptions on the distribution of differences may not be met: normality and absence of outliers. These assumptions should be studied with care.

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.

N/A

TASK: CATARACTS - Semantic Segmentation

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.

Video processing and understanding can be used to empower computer assisted interventions (CAI) as well as the development of detailed post-operative analysis of the surgical intervention. A fundamental building block to such capabilities is the ability to understand and segment video frames into semantic labels that differentiate and localize tissue types and different instruments. Deep learning has advanced semantic segmentation techniques dramatically in recent years. Different papers have proposed and studied deep learning models for the task of segmenting color images into body organs and instruments. These studies are however performed different dataset and different level of granualirities, like instrument vs. background, instrument category vs background and instrument category vs body organs. In this challenge, we create a fine-grained annotated dataset that all anatomical structures and instruments are labelled to allow for a standard evaluation of models using the same data at different granularities. We introduce a high quality dataset for semantic segmentation in Cataract surgery. We generated this dataset from the CATARACTS challenge dataset, which is publicly available. To the best of our knowledge, this dataset has the highest quality annotation in surgical data to date.

Keywords

List the primary keywords that characterize the task.

cataract surgery; semantic segmentation;

ORGANIZATION

Organizers

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

Danail Stoyanov, Prof., Wellcome/EPSRC Centre for Interventional and Surgical Sciences, University College London, London UK; Digital Surgery Ltd. London UK
Imanol Luengo, Ph.D., Digital Surgery Ltd. London UK
Abdolrahim Kadkhodamohammadi, Ph.D., Digital Surgery Ltd. London UK

b) Provide information on the primary contact person.

Imanol Luengo, imanol.luengo@digitalsurgery.com

Life cycle type

Define the intended submission cycle of the challenge. Include information on whether/how the challenge will be continued after the challenge has taken place.

Examples:

- One-time event with fixed submission deadline
- Open call
- Repeated event with annual fixed submission deadline

One time event.

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).

Part of https://endovis.grand-challenge.org/, Sub-challenge: https://cataracts-semantic-segmentation2020.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.

Publicly available 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.

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.

The winner will receive an award (to be decided).

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.

Participants waive the rights to decide whether to show their results on the results page. For each algorithm, we will only report the best score on the result page.

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

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

We will publish a challenge paper. Upto the top N teams will be invited to join the organisers and contribute to a paper to describe and summarize the submitted methods, results and the challenge findings. Participants are allowed to publish their own results separately only after publication of the challenge paper (expected mid 2021).

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.

Docker containers with specific input/output protocols will be submitted by participants and benchmarked internally by us. Test data wont be available to the public. Submission instructions to be defined.

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.

Validation set will be provided and erroneous docker submissions will be reported to participants after benchmarking the docker containers. To that end, submitted docker containers will be evaluated on a very small subset of the testing set (e.g. 50 images) and reported back for sanity check. We allow each team to submit upto 20 working docker images. Broken docker images will not be counted.

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

Release date of the training and validation data: April 1, 2020
Registration date: March 1, 2020
Release date of test data: September 21 2020
Submission deadline: September 27 2020

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).

Study approved by the IRB of Brest University Hospital on 28 janvier 2013. Informed consent was obtained from all patients.

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: Additional comments: We will be sending a document, describing the terms and conditions of the challenge, to the participants which must sign it before their registration is validated.

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.

We will provide Python code for the evaluation metrics used to benchmark the docker instances so that it can be used for local validation.

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

Participants decide whether to publish their code or not.

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.

If the community plan to collect share new annotations for the dataset, they must do it through the CATARACTS website. We are not going to release the test case labels. There will be only a webpage to evaluate the results on the test cases. We will continue to maintain the webpage to benchmark new models on the test set after the conference.

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

Research, Education, Training, Decision support, Prevention.

Task category(ies)

State the task category(ies).

Examples:

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

Segmentation.

Cohorts

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

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

The final application should be using video streams of cataract surgeries performed in hospitals.

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

The challenge is based on video streams data of cataract surgeries performed in Brest University Hospital.

Imaging modality(ies)

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

Surgical microscope videos.

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 videos show the surgical field of the cataract surgeries where only the anterior segment of the eye and the tool tips are present.

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

Patients were 61 years old on average (minimum: 23, maximum: 83, standard deviation: 10). There were 38 females and 12 males.

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.

Eye shown in cataract surgery videos.

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.

Surgeon activities.

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)).

Specificity, Sensitivity, Precision, Accuracy.

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).

Surgeries were performed under an OPMI Lumera T microscope (Carl Zeiss Meditec, Jena, Germany). Videos were recorded with a 180I camera (Toshiba, Tokyo, Japan) and a MediCap USB200 recorder (MediCapture, Plymouth Meeting, USA).

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

When the surgeon starts the intervention, a simple push to a button starts the recording of the surgical field.

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.

Brest University Hospital, France.

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).

Surgeries were performed by three surgeons: a renowned expert, a one-year experienced surgeon and an intern.

Training and test case characteristics

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

Examples:

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

Training and validation sets contain color images extracted from 25 videos, each video corresponding to a different patient. Test set will be composed of another 10 videos from different patients, captured with the same microscope and in the same conditions. The training and validation dataset consist of approximately 4.5K color images (with their corresponding segmentation masks associated), extracted uniformly across different phases of the procedures. We have 20 different semantic classes. The class labels are Pupil, Surgical Tape, Hand, Eye Retractors, Iris, Skin, Cornea, Hydrosdissection, Cannula, Viscoelastic Cannula, Capsulorhexis Cystotome, Rycroft Cannula, Bonn Forceps, Primary Knife, Phacoemulsifier Handpiece, Lens Injector, Irrigation/Aspiration Handpiece, Secondary Knife, Micromanipulator and Capsulorhexis Forceps.

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

Frames from 35 cataract surgery videos have been annotated for scene segmentation.

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

The dataset was divided into a training set and validation set to be splitted by the participants at their convenience (25 videos), and a test set (10 videos).

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.

Division was made in such a way that each instrument is proportionally distributed in both the training and test sets.

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 data is manually annotated to provide pixel-level label for the selected frames. To ensure the quality of the annotations, each frame has been annotated by at least two different annotators. Annotated frames have also been QAed by another annotator. If the annotators have difficulties in determining class labels or do not agree on a label, a medical liaison officer specialised in cataract surgery helps to clarify the situation.

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

The annotation is performed by roto artists. All annotators are trained by our medical liaison officer to get familiar with phacoemulsification cataract surgery and different instruments used at each step. As both the roto artist and the medical officer were inhouse, the annotator had continuous access to the medical officer and could use her help when needed.

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.

- 2x Roto artists, responsible for creating the fine grained segmentation masks, had more than two years of experience.
- Medical Liaison officer has MSc on Medical Visualisation & Human Anatomy and more than two years of experience working on cataract video analysis - gave annotation guidelines and training to roto artists, as well as validated our annotations with KOLs and QA'd the rotoscoped annotations to ensure robustness.

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.

Each frame has been annotated by at least two annotators and is QA'd by another annotator. If after QA'ing the generated segmentation masks for a frame does not match, a third annotator will annotate the frame by making corrections in both annotations provided by previous annotators. If there is a disagreement in class labels, the medical officer's opinion is sought. The medical officer will also check the segmentation masks to verify the correctness of the class labels.

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.

No preprocessing is used. The frame are extracted from videos and used to generate the dataset.

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 most relevant sources of error could be:
1) inaccurate delineation of the semantic classes. To mitigate this, we annotate and QA each frame twice by different annotators. If the segmentation masks for a frame does not match, we ask a third annotator to modify the mask so that they both generate the same segmentation masks;
2) difference in class labels, as some of the instruments are very similar, the annotator might have difficulty in determining the right class labels. To mitigate these cases, we ask the medical officer to determine or verify class labels.

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

N/A

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)

We use Intersection over union (IoU).

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

The IoU, which is also known as Jaccard Index, is one of the commonly used metrics to evaluate scene segmentation. This metric summarise not only the correct detection but also false detection and false negative. We will compute IoU per class and compute the final metric as the average of IoU.

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 the IoU per class and use the average IoU across all the classes to rank participants. We will also report the IoU metric per class as supplementary material.

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

As the submissions will be done through docker containers, we do not expect missing results. We should also note that the participant will get evaluation results on the validation after submitting their docker image that can serve a sanity check to make sure the docker image works correctly.

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

As mentioned, this metric summarises model performance by taking into account true, false and missed detections.

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.

t-test on the scores (a score per test case) is used to assess the variability of rankings.

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

Paired sample t-test is the standard statistical test to determine whether the mean difference between two sets of observations (one observation per test case in our case) is zero. It should be noted, however, that two assumptions on the distribution of differences may not be met: normality and absence of outliers. These assumptions should be studied with care.

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.

N/A

TASK: MIcro-Surgical Anastomose Workflow recognition on training sessions

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.

Automatic and online recognition of surgical workflows is mandatory to bring computer-assisted surgery (CAS) applications inside the operating room. According to the type of surgery, different modalities could be used for workflow recognition. In the case where the addition of multiple sensors is not possible, the information available for manual surgery is generally restricted to video-only. In the case of robotic-assisted surgery, kinematic information is also available. It is expected that multimodal data would make easier automatic recognition methods.
The "MIcro-Surgical Anastomose Workflow recognition" (MISAW) sub-challenge provides a unique dataset for online automatic recognition of surgical workflow by using both kinematic and stereoscopic video information on a micro-anastomosis training task. Participants are challenged to recognize online surgical workflow at different granularity levels (phases, steps, and activities) by taking advantage of both modalities available. Participants can submit results for the recognition of one or several granularity levels. In the case of several granularities, participants are encouraged (but not required) to submit the result of a multi-granularity workflow recognition, i.e. recognize different granularity levels thanks to a unique model.

Keywords

List the primary keywords that characterize the task.

Surgical Process Model, Workflow recognition, Multi-modality, OR of the future

ORGANIZATION

Organizers

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

Arnaud Huaulmé, Univ Rennes, INSERM, LTSI - UMR 1099, F35000, Rennes, France
• Duygu Sarikaya, Univ Rennes, INSERM, LTSI - UMR 1099, F35000, Rennes, France
• Kevin Le Mut, Univ Rennes, INSERM, LTSI - UMR 1099, F35000, Rennes, France
• Pierre Jannin, Univ Rennes, INSERM, LTSI - UMR 1099, F35000, Rennes, France
• Kanako Harada, Department of Mechanical Engineering, the University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan

b) Provide information on the primary contact person.

Arnaud Huaulmé, arnaud.huaulme@univ-rennes1.fr

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.

Examples:

- One-time event with fixed submission deadline
- Open call
- Repeated event with annual fixed submission deadline

One time event.

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.

www.synapse.org

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

Part of https://endovis.grand-challenge.org/, Sub-challenge: https://www.synapse.org/MISAW

Participation policies

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

Fully automatic.

Additional points: Automatic methods using kinematic and video data only.

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.

Data used to train algorithms is limited to data provided and publicly available datasets, including pre-trained networks. Publicly available datasets only cover data that have been available to everyone at the beginning of the challenge.

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.

The award policy will be dependent of the ranking on each task (see item 27). Challengs prizes depending on the availability of sponsorship.

e) Define the policy for result announcement.

Examples:

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

The top three performing methods will be announced publicly during the challenge day. The remaining teams could decide whether or not their identity should be publicly revealed (e.g. in the challenge publication).

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).

All participating teams that reveal their identity can nominate two members of their team as co-authors for the challenge publication.
The method description submitted by the participant will be used in challenge publication. Personal data of the participant will include their names, affiliation and contact addresses. References used in the method's description may be published in the challenge results as well.
Participating teams may publish their own results separately with an explicit allowance from the challenge organizers once the challenge publication has been accepted for publication.

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.

Docker container on the Synapse platform. Detailed instructions for submission will be provided with the training data. Algorithm output must provide for each timestamp the recognized value separated by a tabulation. Only one value for phase and step recognition. For activity recognition, results must provide the result for the left hand and right hand (separate by a tabulation). For multi-granularity, results must provide phase, step, left-hand and right-hand activity.

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.

Submission of multiple results possible. Only the last run is officially counted to compute challenge results. No leaderboard or evaluation results will be provided before the end of the challenge.

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

• June 1st: Release of the training data
• Until August 23th: registration
• August 24th: Release of the test data
• September 20th (23:59 PST): Submission deadline
• October 4-8: challenge day

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).

Not applicable. Data does not include patient information.

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: Publicly available for non-commercial use after the challenge and the challenge paper submission.

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.

Organizers' evaluation scripts will be publicly available (open access) after the challenge.

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

Participating teams are encouraged (but not required), to provide their code as open access.

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.

This work was funded by ImPACT Program of Council for Science, Technology and Innovation, Cabinet Office, Government of Japan.
All challenge organizers and some members of their institute had access to training and test cases. Therefore, there are not allowed to participate in the challenge.

MISSION OF THE CHALLENGE

Field(s) of application

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

Examples:

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

Research, Assistance, Education, Training, Decision support.

Task category(ies)

State the task category(ies).

Examples:

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

Uni-granularity surgical workflow recognition (3 different granularity levels)
• Multi-granularity surgical workflow recognition

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.

Robotic micro-surgical suturing of dura mater during endonasal brain tumor surgery.

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

Robotic micro-surgical anastomosis on artificial blood vessel.

Imaging modality(ies)

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

Stereoscopic video.

Context information

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

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

No context information is given along with the images.

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

No patients are involved, data are acquired on artificial blood vessel.

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.

On final application, data would be acquired on endonasal brain tumor surgery.

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

The algorithm target is the automatic recognition of one or several granularity levels of the micro-surgical anastomosis task.

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)).

Robustness, Accuracy

Additional points: The algorithms must be applicable for online application.
• The recognition has to be accurate, robust and reproduceable

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 Kinematic and video data was synchronously acquired at 30 Hz thanks to a master-slave robotic platform (Mitsuishi et al. 2013).
Workflow annotation was acquired manually thanks to the video and the software "Surgery Workflow Toolbox [annotated]" provide by the IRT b<>com (Garraud et al. 2014)

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

Video and kinematic data are recorded synchronously at 30 Hz.

Kinematic data are recorded from encoders mounted on the two robotic arms. Kinematic data consists of x, y, z, alpha, beta, gamma).

Video data was captured with a high-definition stereo-microscope (960x540 px). Due to the not fixed boundary between the left image and the right image, i.e. the position of the centerline is a little different between the trials, we have removing 40 px on the center of the stereoscopic image to have two images of 460x540 px. Final video resolution is 920x540 px.

Worflow annotation details were provide at item 23.

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.

- Kinematic and video data were acquired by the department of mechanical engineering of the University of Tokyo.
- Workflow data were acquired by the MediCis team, of LTSI Laboratory from University of Rennes.

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).

- 3 expert surgeons (good medical expertise, poor robotic manipulation expertise.
- 3 engineering students (poor medical expertise, good robotic manipulation expertise).

Training and test case characteristics

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

Examples:

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

Training and test cases are composed of:
• Kinematic data (x, y, z, alpha, beta, gamma at 30hz)
• Stereoscopic video data (Left/right image 460x540 px. Final video resolution is 920x540 px. 30Hz)
• Workflow annotation (not provide for test cases), containing for each timestamp the label for phase, steps, left-hand activity and right-hand activity.

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

Training: 17 cases
Testing: 10 cases
No validation cases are provided. It is up to the participants to split the training dataset into training and validation data.

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

Two users are removed from the training dataset. The training dataset represent 37% of the total dataset.

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.

Training dataset (17 cases; 4 users):
Surgeon 2: 3 cases
Surgeon 3: 4 cases
Engineering student 1: 6 cases
Engineering student 2: 4 cases

Test dataset (10 cases; 2 user)
Surgeon 1: 4 cases
Engineering student 3: 6 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.

Workflow annotation was performed manually by two observers.

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

The observers have been asked to annotate the surgical workflow at the following granularity levels: phase, step, and activity. The annotation protocol is available at: https://ged.univ-rennes1.fr/nuxeo/site/esupversions/fde75c3e-b37c-4ac7-8ac2-4ec5c21fba88

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.

Two non-medical observers with expertise on annotation process.

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.

To create the gold-standard the differences will be discussed between both annotators to obtain a consensus.

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.

Only data formatting will be done before provided it on workflow annotation and video:

Workflows annotation was modified to switch from continuous sequence to a discrete sequence at 30Hz (synchronize to kinematic and video data). For each timestamp, the annotation would like as follow: timestamp_number, phase_value, step_value, activity_Left_Hand_value, activity_Right_Hand_value. With a tabulation as separator.

The videos were cropped to remove 40 px from the centerline due to the non-fixed boundary between the left image and the right image. Final video resolution is 920x540 px.

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 boundary of workflow annotation is observer-dependent, so the transition between two components of the same granularity (e.g. phase) is not framed-perfect. These variabilities could be similar to the intra-observer variability describe in (Huaulmé et al. 2019). The observer who performed the annotation for the challenge is the one who performs annotation for the intra-observer study of this publication.

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

N/A

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)

Frame by frame scores: balanced accuracy, precision, recall, f1 for all class.
Application-dependent scores: balanced accuracy, precision, recall, f1 for all class

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

Balanced accuracy: global result of the recognition by taking into account unbalanced class.

Precision, Recall, f1: Local results of the recognition to take into account under-represented classes

Frame by Frame scores: classic metric used for accuracy precision recall and f1.

Application-dependent (AD) scores: The frame by frame scores does not take into account the most relevant source of error on workflow annotation (item25.a) and the needs of the application. AD-scores re-estimate classic scores by using acceptable delay thresholds for a transitional window (Dergachyova et al. 2016).

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.

To perform the ranking, we will use a metric-based aggregation on the balanced Application-Dependent accuracy. For one participant, we will aggregate metric values over all test cases and aggregate overall metrics to obtain a final score.
The score s_uni for an uni-granularity recognition algorithm (a_i) will be:
Linear equation: s_uni(a_i)=sum(balanced_AD_accuracy_trial_t)/T ; with sum from t=0 until T.

The score s_multi for a multi-granularity recognition algorithm will be the mean of the score for each uni-granularity score:
Linear equation: s_multi (a_i)= (s_phase(a_i)+s_step(a_i)+s_activity(a_i))/3 .

The different ranking will be made according to the task made by the participant. One ranking for each uni-granularity recognition (e.g. phase only), another for the multi-granularity recognition.
All algorithms performing a multi-granularity recognition will be included in corresponding uni-granularity recognition ranking.

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

Our challenge is multi-class recognition. In the case of missing results, we will consider results as good as a total random recognition. For example, in 3 classes problem, the missing result will be set to 1/3. For a 12 classes problem, to 1/12.

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

We decided to use a metric-based aggregation according to the conclusion of (Maier-Hein et al. 2018) reporting this type of aggregation as one of the most robust.

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.

Stability will be investigated via bootstrapping and hypothesis testing.

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

Bootstrapping was identified by (Maier-Hein et al. 2018) as an appropriate approach to investigate ranking variability.

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.

N/A

TASK: SurgVisDom - Surgical visual domain adaptation: from virtual reality to real, clinical environments

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.

Surgical data science is revolutionizing minimally invasive surgery. By developing algorithms to be context-aware, exciting applications to augment surgeons are becoming possible. However, there exist many sensitivities around surgical data (or health data more generally) needed to develop context-aware models. This challenge seeks to explore the potential for visual domain adaptation in surgery to overcome data privacy concerns. In particular, we propose to use video from virtual reality simulation data from clinical-like tasks to develop algorithms to recognize activities and then to test these algorithms on videos of the same task in a clinical setting (i.e., porcine model).

Keywords

List the primary keywords that characterize the task.

visual domain adaptation; surgery; virtual reality; endoscope video; activity recognition

ORGANIZATION

Organizers

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

Aneeq Zia (Intuitive Surgical),
Kiran Bhattacharyya (Intuitive Surgical),
Xi Liu (Intuitive Surgical),
Ziheng Wang (Intuitive Surgical),
Anthony Jarc (Intuitive Surgical)

b) Provide information on the primary contact person.

Aneeq Zia (aneeq.zia@intusurg.com)

Life cycle type

Define the intended submission cycle of the challenge. Include information on whether/how the challenge will be continued after the challenge has taken place.

Examples:

- One-time event with fixed submission deadline
- Open call
- Repeated event with annual fixed submission deadline

One time event.

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).

Part of https://endovis.grand-challenge.org/, Sub-challenge: https://survisdom.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.

Open to publicly available data including pre-trained nets. Any privately prepared annotations on public data will need to be released at time of 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.

May not participate.

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

3 monetary prizes for 1st, 2nd, and 3rd place. $500, $300 and $200, for 1st,2nd and 3rd, respectively.

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.

Top three performing methods will be announced publicly and posted on the website.

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

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

The organizers will publish a challenge paper within six months after the challenge. Following which, the participating teams can publish their own results from the challenge citing the challenge paper. Possibility of a combined publication amongst the participating teams/organization team will also be discussed after the challenge.

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.

Submission instructions will be posted to the website and sent via email. Results will be submitted via a docker container. The docker containers will need to be sent directly to the organizers with instructions on how to run.

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.

The participants will not be allowed to evaluate their algorithms before submission - only one final submission per team.

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

Release of training cases in June 2020; registration ongoing; submission date September 2020; MICCAI 2020; release of results at MICCAI 2020; release of a subset of test set after MICCAI 2020

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).

An existing Western IRB will be used.

Data usage agreement

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

Examples:

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

CC BY NC ND.

Additional comments: (Attribution-NonCommercial-NoDerivs).

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.

Open source on challenge site.

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

Open and private code submission will be accepted.

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.

Sposorship/funding will be done by Intuitive Surgical. The organizers who are affiliated with Intuitive will perform testing.

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

Research, Assistance, Training.

Task category(ies)

State the task category(ies).

Examples:

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

Classification.

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.

Robotic clinical-like activities on a porcine model used during training courses; subjects will be surgeons of varying experiences.

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

Virtual reality tasks of the same clinical-like activities as the target cohort; subjects will be surgeons and non-surgeons (but with extensive experience operating the robot).

Imaging modality(ies)

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

One channel of endscopic video.

Context information

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

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

Each video clip will include a label of the activity being performed.

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

Not applicable in VR or porcine.

Target entity(ies)

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

The data will be acquired from VR and training courses on porcine models. The clinical-like activity is performed in a standardized manner and is also represented in the VR simulator.

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.

Surgical activity recognition during clinical-like task. There will be 3 activities to recognize.

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)).

Sensitivity, Precision, Accuracy.

Additional points: The assessment aim will be multi-class classification measured by the average f1-score across different classes.

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 Intuitive Data Recorder (IDR) will be used to capture video at 720p and 60fps from one channel of the endoscope on da Vinci surgical systems.

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

N/A

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.

Data will be collected at Intuitive Surgical training labs.

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).

Experience ranges from 0 robotic cases to over 1000 robotic cases.

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.

Our dataset will consists of 3 commonly performed surgical tasks (needle-driving, knot tying and dissection) in VR and porcine. Training cases will represent individual video clips from VR based skills tasks performed on the da Vinci SimNow platform along with analogous surgical tasks performed on a porcine model. Test cases will include the same surgical tasks performed on a porcine model. Each training video clip will have a single task label, whereas the video clips in the test set can contain more than one surgical task (non-overlapping). The participants will need to produce a frame-by-frame prediction for each video clip in the test set.

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

22 (VR) + 3 (porcine) training cases will be made available for the 3 tasks resulting in a total set of 75 training videos. The test set will contain a total of 30 video clips with a balance representation of the 3 surgical tasks.

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

The numbers indicated were kept keeping in mind data collection technicalities and to provide enough data to the participants for developing meaningful models.

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 will try to ensure class balance within training and testing sets across VR and porcine models. The 3 chosen tasks are equally important and prevalent in surgical training.

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.

1-2 human annotators will be used to temporally annotate porcine and VR video data where required.

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.

Annotators will only specify start and stop times of individual tasks for porcine data - no other special instruction would be given.

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.

Annotators will be humans who are well-versed with clinical knowledge required for such annotations.

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.

One case will be annotated by only one annotator. However, each annotation will be validated and adjusted by an expert as needed.

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.

Raw videos will be provided to the participants.

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 surgical tasks chose are relatively simple to annotate and the annotations will be validated by experts. Therefore, we expect the range of annotation error to be very low.

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

To the best of our knowledge, we don't expect any other 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)

Frame level average F1-score across the 3 classes will be used for assessment.

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

Many surgical/non-surgical activity recognition work in the recent few years have shown the validity of this metric.

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.

The performance rank will be based on the rank of the f1-score.

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

The submission will be docker based with model and prediction script requirement from the participants. This will eliminate the possibility of missing results since the organizing team will evaluate performances on test cases.

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

Using f1-score for ranking seems most reasonable for such a classification problem.

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.

The main differentiator of algorithms will be a classification f1-score . However we will also perform statistical significance checks on model performance across tasks and cases.

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

The ranking will be classification based hence no special statistical analyses method explored.

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.

N/A

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.

Dergachyova, O., D. Bouget, A. Huaulmé, X. Morandi, and P. Jannin. 2016. “Automatic Data-Driven Real-Time Segmentation and Recognition of Surgical Workflow.” International Journal of Computer Assisted Radiology and Surgery. https://doi.org/10.1007/s11548-016-1371-x.

Garraud, C., B. Gibaud, C. Penet, G. Gazuguel, G. Dardenne, and P. Jannin. 2014. “An Ontology-Based Software

Suite for the Analysis of Surgical Process Model." In Proceedings of Surgetica'2014, 243–45. Chambery, France.

Huaulmé, A., F. Despinoy, S.A. Heredia Perez, K. Harada, M. Mitsuishi, and P. Jannin. 2019. "Automatic Annotation of Surgical Activities Using Virtual Reality Environments." International Journal of Computer Assisted Radiology and Surgery, June. https://doi.org/10.1007/s11548-019-02008-x.

Maier-Hein, L., M. Eisenmann, A. Reinke, S. Onogur, M. Stankovic, P. Scholz, T. Arbel, et al. 2018. "Why Rankings of Biomedical Image Analysis Competitions Should Be Interpreted with Care." Nature Communications 9 (1). https://doi.org/10.1038/s41467-018-07619-7.

Mitsuishi, M., A. Morita, N. Sugita, S. Sora, R. Mochizuki, K. Tanimoto, Y.M. Baek, H. Takahashi, and K. Harada. 2013. "Master-Slave Robotic Platform and Its Feasibility Study for Micro-Neurosurgery: Master-Slave Robotic Platform for Microneurosurgery." The International Journal of Medical Robotics and Computer Assisted Surgery 9 (2): 180–89.