2nd Retinal Fundus Glaucoma Challenge: Structured description of the challenge design

CHALLENGE ORGANIZATION

Title

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

2nd Retinal Fundus Glaucoma Challenge

Challenge acronym

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

REFUGE-2

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.

Glaucoma is currently the leading reason of irreversible blindness in the world. This challenge brings together the medical image analysis community to develop novel automated classification and segmentation methods and to evaluate and compare them for glaucoma detection and optic disc/cup segmentation on a large dataset of retinal fundus images. 1st REFUGE was organized successfully in MICCAI 2018, where 28 teams registered out of which 11 submitted the test set results and participated in. In this year, our REFUGE-2 challenge will utilize a larger dataset (2000 images with glaucoma labels and disc/cup annotations) with more camera modalities (four fundus cameras), which are collected from multiple medical centers. The additional 800 images are captured using two new camera modalities. The REFUGE-2 challenge will consist of three tasks: 1) Classification of clinical Glaucoma; 2) Segmentation of Optic Disc and Cup; 3) Localization of Fovea, , which gives attention to the clinical application and technical research.

Challenge keywords

List the primary keywords that characterize the challenge.

fundus image, glaucoma detection, optic disc, optic cup.

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.

7th MICCAI Workshop on Ophthalmic Medical Image Analysis (OMIA7)

Duration

How long does the challenge take?

Half day.

Expected number of participants

Please explain the basis of your estimate (e.g. numbers from previous challenges) and/or provide a list of potential participants and indicate if they have already confirmed their willingness to contribute.

Expected number is about 20 teams.

In MICCAI 2018, we organized 1st REFUGE, where 28 teams registered out of which 11 submitted the test set results and participated in. This year we use a larger fundus image dataset with more modalities, which are mode widespread so we expect even larger interest from the ophthalmic image analysis community.

Publication and future plans

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

We will prepare a challenge review paper to IEEE transactions on medical imaging or Medical Image Analysis. The review paper of REFUGE-1 has been published in Medical Image Analysis 2020 (DOI: 10.1016/j.media.2019.101570).

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.

Expected participant team number is 20.
One projector and three microphones are needed.

TASK: Classification of clinical Glaucoma

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.

Glaucoma is currently the leading reason of irreversible blindness in the world. It is commonly caused by elevated intraocular pressure (IOP), which causes mechanical straining and torsion of optic nerve and loss of retinal nerve fibers. Glaucoma classification consists in categorizing an fundus image into glaucomatous or non-glaucomatous, based on its visual characteristics.

Keywords

List the primary keywords that characterize the task.

Glaucoma classification

ORGANIZATION

Organizers

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

Huazhu Fu, Inception Institute of Artificial Intelligence, UAE.
Yanwu Xu, Baidu Inc., China.
José Ignacio Orlando, CONICET/PLADEMA-UNICEN, Argentina.
Hrvoje Bogunovi, Medical University of Vienna, Austria.
Fei Li, Zhongshan Ophthalmic Center, Sun Yat-sen University, China.
Xiulan Zhang, Zhongshan Ophthalmic Center, Sun Yat-sen University, China.

b) Provide information on the primary contact person.

Yanwu Xu (ywxu@ieee.org)

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).

https://refuge.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.

No policy defined.

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.

A total of ~4000 USD awards will be provided by Baidu for the onsite challenge.

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.

We will score and rank submissions according to evaluation metrics. All the scores and ranks will be displayed publicly on the website leaderboard. Top-performing on-site participating teams will be invited to attened off-site competition and workshop. The final winners are based the perfomance of both on-site and off-site competitions.

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

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

The top-performing participating teams and individuals (based on the performance of the method) will be invited to contribute to a joint journal paper(s) with maximum 2 authors per team describing and summarizing the methods used and results found in this challenge. The paper will be submitted to a high-impact journal in the field. The participating teams could publish their own methods and results separately.

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.

The classification results should be provided in a single CSV file, named "classification_results.csv", with the first column corresponding to the filename of the test fundus image (including the extension ".jpg") and the second column containing the estimated classification probability/risk of the image belonging to a patient diagnosed with glaucoma (value from 0.0 to 1.0).

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.

Challenge participants will be allow to make 2 submissions per day. The REFUGE-2 challenge will be hosted at the MICCAI 2020 conference in conjunction with OMIA workshop. There will also be an on-site part of the challenge when the second part of the test set will be released. The participants will have 1 hour on the day of the challenge to provide the results on the "on-site 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

10 July 2020: Registration opens.
15 July 2020: Training images are released.
20 July 2020: Annotations of the Training set are released.
15 Aug 2020: Off-site validation set is released.
18 Aug 2020: Submissions of results on off-site validation set are opened.
27 Aug 2020: Off-site validation set results submission deadline.
28 Aug 2020: Paper submission deadline.
01 Sep 2020: Off-site leaderboard is published and the best teams (with paper submitted) are invited for the on-site REFUGE-2.
20 Sep 2020: Camera-ready paper submission deadline.
08 Oct 2020: On-site REFUGE-2 (1 hour) in conjunction with OMIA workshop at 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).

All the data have been obtained IRB approval.

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.

Additional comments: When publishing the obtained results in scientific publications, you should make an appropriate citation in accordance with academic custom.

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 evaluation code on Github at https://github.com/ignaciorlando/refuge-evaluation

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

We also encourage all participating teams releasing their codes after challenge.

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.

A total of ~4000 USD awards will be provided by Baidu for the onsite ADAM challenge. No one explicitly could access to the test case labels before the challenge ended.

MISSION OF THE CHALLENGE

Field(s) of application

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

Examples:

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

Screening, Diagnosis.

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.

Patients visiting eye hospital.

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

Patients visiting eye hospital.

Imaging modality(ies)

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

Retinal fundus image.

Context information

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

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

Retinal fundus image.

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

None.

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.

Retinal fundus image.

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

Glaucoma detection in fundus image.

Assessment aim(s)

Identify the property(ies) of the algorithms to be optimized to perform well in the challenge. If multiple properties are assessed, prioritize them (if appropriate). The properties should then be reflected in the metrics applied (see below, parameter metric(s)), and the priorities should be reflected in the ranking when combining multiple metrics that assess different properties.

- Example 1: Find highly accurate liver segmentation algorithm for CT images.
- Example 2: Find lung tumor detection algorithm with high sensitivity and specificity for mammography images.

Corresponding metrics are listed below (parameter metric(s)).

Accuracy.

Additional points: Classification results will be compared to the clinical grading of glaucoma. Receiver operating curve will be created across all the test set images and an area under the curve (AUC) will be calculated. Each team receives a rank (1=best) based on the obtained AUC value. This ranking forms the classification leaderboard.

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).

Zeiss Visucam 500, Canon CR-2, topcon tirton, kowa.

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

-Dimension: 2D
-Timepoints: 1-time
-Position and orientation of the patient: sitting upright

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.

Zhongshan Ophthalmic Center, Sun Yat-sen University, China.

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).

Original fundus images captured by ophthalmologists and technicians;
Annotated with clinical diagnosis/records with multi-modality information.

Training and test case characteristics

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

Examples:

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

Training and test cases both represent a fundus image with glaucoma labels.

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

A total of 2000 color fundus photographs are available. The dataset is split into 3 subsets for training (1200), offline validation (400) and onsite test (400), stratified to have equal glaucoma presence percentage.

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

1200 training images are from previous REFUGE challenge as training data. 400 offline validation and 400 onsite test images in REFUGE2 are additional data collecting by the different fundus camera modalities, which have no overlapping with previous challenge.

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.

Class distribution in classification task is 2:8 (glaucoma vs. non-glaucoma).

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 reference standard for glaucoma presence were assigned based on the comprehensive evaluation of the subjects' clinical records, including follow-up fundus images, IOP measurements, optical coherence tomography images and visual fields (VF).

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.

Each image in the REFUGE data set includes a reference, trustworthy glaucomatous / non-glaucomatous label. These diagnostics were assigned based on the comprehensive evaluation of the subjects' clinical records, including follow-up fundus images, IOP measurements, optical coherence tomography images and visual fields (VF). The glaucomatous cases correspond to subjects with glaucomatous damage in the ONH area and reproducible glaucomatous VF defects. This last characteristic was defined as a reproducible reduction in sensitivity compared to the normative data set, in reliable tests, at: (1) two or more contiguous locations with p - value < 0.01 and (2) three or more contiguous locations with p - value < 0.05. ONH damage was defined as a vCDR > 0.7, thinning of the RNFL, or both, without a retinal or neurological cause for VF loss. Notice, then, that instead of using labels assigned based on a single CFP at a specific timepoint, the labels were retrieved from examinations of follow-up medical records using a pre-determined criterion, to ensure the reliability of the classification labels. 10%

c) Provide details on the subject(s)/algorithm(s) that annotated the cases (e.g. information on level of expertise such as number of years of professional experience, medically-trained or not). Provide the information separately for the training, validation and test cases if necessary.

The reference standard for glaucoma presence were assigned based on the comprehensive evaluation of the subjects' clinical records, including follow-up fundus images, IOP measurements, optical coherence tomography images and visual fields (VF).

d) Describe the method(s) used to merge multiple annotations for one case (if any). Provide the information separately for the training, validation and test cases if necessary.

No need.

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.

- Quality control: only high-quality images are selected for manual annotation;
- Remove personal information and device information;
- The glaucoma labels are obtained from clinical diagnosis/records with multimodality information;
- Final labels were double checked by two other senior ophthalmologists for every image.

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.

None.

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

None.

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)

Classification results will be compared to the clinical grading of glaucoma. Receiver operating curve will be created across all the test set images and an area under the curve (AUC) will be calculated.

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

AUC is a common metric in clinical diagnosis evaluation.

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.

Each team receives a rank (1=best) based on the obtained AUC value. This ranking forms the classification leaderboard.

In our challenge, the on-site and off-site competitions will have different weights for the final ranking, as:
Final score = 0.3 * off-site ranking + 0.7 * on-site ranking,
where the off-site ranking is used to encourage the participants to obtain the staged result, and a lower weight (e.g., 0.3) guarantees the final ranking prefer on the test set.

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

Missing result will report an error information.

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

AUC is a standard evaluation measure for screening/diagnosis purposes.

Statistical analyses

a) Provide details for the statistical methods used in the scope of the challenge analysis. This may include

- description of the missing data handling,
- details about the assessment of variability of rankings,
- description of any method used to assess whether the data met the assumptions, required for the particular statistical approach, or
- indication of any software product that was used for all data analysis methods.

Missing data not allowed, and will receive an error notice. And significance of differences in the rankings will be tested by Wilcoxon Mann Whitney test.

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

The statistical significance of the differences in performance of the top-ranked teams was assessed by means of Wilcoxon signed-rank tests (= 0.05). The statistical significance of the differences between groups was assessed using a Wilcoxon rank-sum test due to the unpaired nature of the two sets (360 vs. 40 samples, respectively) for avoiding outlier data.

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.

None.

TASK: Segmentation of Optic Disc and Cup

SUMMARY

Abstract

Provide a summary of the challenge purpose. This should include a general introduction in the topic from both a biomedical as well as from a technical point of view and clearly state the envisioned technical and/or biomedical impact of the challenge.

The optic cup is the central cup-like area in the optic disc. The cup to disc ratio (CDR) is the comparison of the diameter of the cup to disc, which partially represents disease status. Determination of CDR varies among doctors and can be influenced by subjectivity. Thus, the accurate and automatic segmentation of optic disc and cup from fundus images is a fundamental task.

Keywords

List the primary keywords that characterize the task.

optic cup segmentation, optic disc segmentation, cup to disc ratio

ORGANIZATION

Organizers

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

Huazhu Fu, Inception Institute of Artificial Intelligence, UAE.
Yanwu Xu, Baidu Inc., China.
José Ignacio Orlando, CONICET/PLADEMA-UNICEN, Argentina.
Hrvoje Bogunovi, Medical University of Vienna, Austria.
Fei Li, Zhongshan Ophthalmic Center, Sun Yat-sen University, China.
Xiulan Zhang, Zhongshan Ophthalmic Center, Sun Yat-sen University, China.

b) Provide information on the primary contact person.

Yanwu Xu (ywxu@ieee.org)

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).

https://refuge.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.

No policy defined.

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.

A total of ~4000 USD awards will be provided by Baidu for the onsite challenge.

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.

We will score and rank submissions according to evaluation metrics. All the scores and ranks will be displayed publicly on the website leaderboard. Top-performing on-site participating teams will be invited to attened off-site competition and workshop. The final winners are based the perfomance of both on-site and off-site competitions.

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

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

The top-performing participating teams and individuals (based on the performance of the method) will be invited to contribute to a joint journal paper(s) with maximum 2 authors per team describing and summarizing the methods used and results found in this challenge. The paper will be submitted to a high-impact journal in the field. The participating teams could publish their own methods and results separately.

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.

The segmentation results should be provided in a “segmentation” folder, as one image per test image, with the segmented pixels labeled in the same way as in the reference standard (bmp (8-bit) files with 0: optic cup, 128: optic disc, 255: elsewhere). Please, make sure that your submitted segmentation files are named according to the original image names and with the same extension.

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.

Challenge participants will be allow to make 2 submissions per day. The REFUGE-2 challenge will be hosted at the MICCAI 2020 conference in conjunction with OMIA workshop. There will also be an on-site part of the challenge when the second part of the test set will be released. The participants will have 1 hour on the day of the challenge to provide the results on the "on-site 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

10 July 2020: Registration opens.
15 July 2020: Training images are released.
20 July 2020: Annotations of the Training set are released.
15 Aug 2020: Off-site validation set is released.
18 Aug 2020: Submissions of results on off-site validation set are opened.
27 Aug 2020: Off-site validation set results submission deadline.
28 Aug 2020: Paper submission deadline.
01 Sep 2020: Off-site leaderboard is published and the best teams (with paper submitted) are invited for the on-site REFUGE-2.
20 Sep 2020: Camera-ready paper submission deadline.
08 Oct 2020: On-site REFUGE-2 (1 hour) in conjunction with OMIA workshop at 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).

All the data have been obtained IRB approval.

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.

Additional comments: When publishing the obtained results in scientific publications, you should make an appropriate citation in accordance with academic custom.

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 evaluation code on Github at https://github.com/ignaciorlando/refuge-evaluation

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

We also encourage all participating teams releasing their codes after challenge.

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.

A total of ~4000 USD awards will be provided by Baidu for the onsite ADAM challenge. No one explicitly could access to the test case labels before the challenge ended.

MISSION OF THE CHALLENGE

Field(s) of application

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

Examples:

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

Decision support.

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.

Patients visiting eye hospital.

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

Patients visiting eye hospital.

Imaging modality(ies)

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

Retinal fundus image.

Context information

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

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

Retinal fundus image.

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

None.

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.

Retinal fundus image.

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

Optic disc/cup segmentation in fundus image.

Assessment aim(s)

Identify the property(ies) of the algorithms to be optimized to perform well in the challenge. If multiple properties are assessed, prioritize them (if appropriate). The properties should then be reflected in the metrics applied (see below, parameter metric(s)), and the priorities should be reflected in the ranking when combining multiple metrics that assess different properties.

- Example 1: Find highly accurate liver segmentation algorithm for CT images.
- Example 2: Find lung tumor detection algorithm with high sensitivity and specificity for mammography images.

Corresponding metrics are listed below (parameter metric(s)).

Accuracy.

Additional points: Submitted segmentation results will be compared to the reference standard. the disc and cup Dice indices (DSC), and the cup-to-disc ratio (CDR) will be calculated as segmentation evaluation measures. Each team receives a rank (1=best) for each evaluation measure based on the mean value of the measure over the set of test images. The segmentation score is then determined by adding the three individual ranks (2xDSC and 1xCDR). The team with the lowest score will be ranked #1 on the segmentation leaderboard.

For segmentation of optic disc and cup, the results were compared with the gold standard segmentation using the Dice index (DSC) for OD/OC separately, and the mean absolute error (MAE) of the vertical cup-to-disc ratio (vCDR) estimations. Finally, the segmentation ranking is based on
Classification score = 0.35*DSC_cup + 0.25*DSC_disc + 0.4*MAE
Since the MAE of the vCDR is calculated based on the seg- mentation of OC and OD, we set a larger weight for vCDR than to each individual segmentation term. Moreover, it is standard to first segment the OD region and then extract the OC from the cropped OD area. Hence, we assigned a larger weight to the OD segmentation results than to the OC.

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).

Zeiss Visucam 500, Canon CR-2, topcon tirton, kowa.

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

-Dimension: 2D
-Timepoints: 1-time
-Position and orientation of the patient: sitting upright

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.

Zhongshan Ophthalmic Center, Sun Yat-sen University, China.

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).

Original fundus images captured by ophthalmologists and technicians;
Annotated with clinical diagnosis/records with multi-modality information.

Training and test case characteristics

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

Examples:

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

Training and test cases both represent a fundus image with the pixel-wise labels.

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

A total of 2000 color fundus photographs are available. The dataset is split into 3 subsets for training (1200), offline validation (400) and onsite test (400), stratified to have equal glaucoma presence percentage.

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

1200 training images are from previous REFUGE challenge as training data. 400 offline validation and 400 onsite test images in REFUGE2 are additional data collecting by the different fundus camera modalities, which have no overlapping with previous challenge.

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.

Class distribution in classification task is 2:8 (glaucoma vs. non-glaucoma).

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.

Manual pixel-wise annotations of the optic disc and cup were obtained by SEVEN (was 3 as proposed) independent GLAUCOMA SPECIALISTS from Zhongshan Ophthalmic Center, Sun Yat-sen University, China. The reference standard for the segmentation task was created from the seven annotations, which were merged into single annotation by another SENIOR GLAUCOMA SPECIALIST.

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 procedure consisted in manually drawing a tilted ellipse covering the OD and the OC, separately, by means of a free annotation tool with capabilities for image review, zoom and ellipse fitting. A single segmentation per image was afterwards obtained by taking the majority voting of the anotations of the seven experts. A senior specialist with more than 10 years of experience in glaucoma performed a quality check afterwards, analyzing the resulting masks to account for potential mistakes. When errors in the annotations were observed, this additional reader analyzed each of the seven segmentations, removed those that were considered failed in his/her opinion and repeated the majority voting process with the remaining ones.

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.

Manual annotations of the OD and the OC were provided by seven independent glaucoma specialists from the Zhongshan Ophthalmic Center (Sun Yat-sen University, China), with an average experience of 8 years in the field (ranging from 5 to 10 years). All the ophthalmologists independently reviewed and delineated the OD/OC in all the images, without having access to any patient information or knowledge of disease prevalence in the data.

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.

Majority vote.

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.

- Quality control: only high-quality images are selected for manual annotation;
- Remove personal information and device information;
- 7 independent annotations from ophthalmologists;
- Unified ground truth were determined by two other ophthalmologists for every image, by 1) remove low-quality annotations; 2) average chosen annotations.

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.

-Inter-/ intra-observer variability.
However, in total 9 ophthalmologists are involved in the annotation process, to make the reference as accurate as possible.

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

None.

ASSESSMENT METHODS

Metric(s)

a) Define the metric(s) to assess a property of an algorithm. These metrics should reflect the desired algorithm properties described in assessment aim(s) (see above). State which metric(s) were used to compute the ranking(s) (if any).

- Example 1: Dice Similarity Coefficient (DSC)
- Example 2: Area under curve (AUC)

For segmentation of optic disc and cup, the results were compared with the gold standard segmentation using the Dice index (DSC) for OD/OC separately, and the mean absolute error (MAE) of the vertical cup-to-disc ratio (vCDR) estimations.

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

Dice score is usually in medical segmenation evaluation.

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.

Finally, the segmentation ranking is based on
Classification score = 0.35*DSC_cup + 0.25*DSC_disc + 0.4*MAE
Since the MAE of the vCDR is calculated based on the seg- mentation of OC and OD, we set a larger weight for vCDR than to each individual segmentation term. Moreover, it is standard to first segment the OD region and then extract the OC from the cropped OD area. Hence, we assigned a larger weight to the OD segmentation results than to the OC.

In our challenge, the on-site and off-site competitions will have different weights for the final ranking, as:
Final score = 0.3 * off-site ranking + 0.7 * on-site ranking,
where the off-site ranking is used to encourage the participants to obtain the staged result, and a lower weight (e.g., 0.3) guarantees the final ranking prefer on the test set.

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

Missing result will report an error information.

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

DSC is a standard measure of segmentation performance.
Cup-to-Disc ratio has a direct clinical relevance as it is a measure used in ophthalmology and optometry to assess the progression of glaucoma.

Statistical analyses

a) Provide details for the statistical methods used in the scope of the challenge analysis. This may include

- description of the missing data handling,
- details about the assessment of variability of rankings,
- description of any method used to assess whether the data met the assumptions, required for the particular statistical approach, or
- indication of any software product that was used for all data analysis methods.

Missing data not allowed, and will receive an error notice. And significance of differences in the rankings will be tested by Wilcoxon Mann Whitney test.

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

The statistical significance of the differences in performance of the top-ranked teams was assessed by means of Wilcoxon signed-rank tests (= 0.05). The statistical significance of the differences between groups was assessed using a Wilcoxon rank-sum test due to the unpaired nature of the two sets (360 vs. 40 samples, respectively) for avoiding outlier data.

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.

None.

TASK: Localization of Fovea (Optional Task)

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.

Fovea is located at the centre of retina. It is temporal to the optic disk between the main superior and inferior vascular arcades. Accurate localization of the foveal region is important to any diagnosis method that is based on the statistical categorization of vision threatening lesions in the retina. We propose this task for optional, which will not be counted into the final ranking. But we will provide a separate rank page for online challenge. Thus, the final challenge ranking is still based on Tasks 1&2.

Keywords

List the primary keywords that characterize the task.

macular center

ORGANIZATION

Organizers

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

Huazhu Fu, Inception Institute of Artificial Intelligence, UAE.
Yanwu Xu, Baidu Inc., China.
José Ignacio Orlando, CONICET/PLADEMA-UNICEN, Argentina.
Hrvoje Bogunovi, Medical University of Vienna, Austria.
Fei Li, Zhongshan Ophthalmic Center, Sun Yat-sen University, China.
Xiulan Zhang, Zhongshan Ophthalmic Center, Sun Yat-sen University, China.

b) Provide information on the primary contact person.

Yanwu Xu (ywxu@ieee.org)

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).

https://refuge.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.

No policy defined.

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.

This task is an optional task.

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.

We will score and rank submissions according to evaluation metrics. All the scores and ranks will be displayed publicly on the website leaderboard. This task is optional task and NOT counted into the final leaderboards.

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).

This task is optional task and NOT counted into the final leaderboards.

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.

The localization results should be provided in a single CSV file, named "fovea_location_results.csv", with the first column corresponding to the filename of the test fundus image (including the extension ".jpg"), the second column containing the X-coordinate and the third column containing the Y-coordinate. Please, make sure that your

submitted segmentation files are named according to the original image names and with the same extension.

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.

Task 3 (fovea localization) is OPTIONAL for MICCAI ONSITE challenge and NOT counted into the final leaderboards.

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

10 July 2020: Registration opens.
15 July 2020: Training images are released.
20 July 2020: Annotations of the Training set are released.
15 Aug 2020: Off-site validation set is released.
18 Aug 2020: Submissions of results on off-site validation set are opened.
27 Aug 2020: Off-site validation set results submission deadline.
28 Aug 2020: Paper submission deadline.
01 Sep 2020: Off-site leaderboard is published and the best teams (with paper submitted) are invited for the on-site REFUGE-2.
20 Sep 2020: Camera-ready paper submission deadline.
08 Oct 2020: On-site REFUGE-2 (1 hour) in conjunction with OMIA workshop at 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).

All the data have been obtained IRB approval.

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.

Additional comments: When publishing the obtained results in scientific publications, you should make an appropriate citation in accordance with academic custom.

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 evaluation code on Github at https://github.com/ignaciorlando/refuge-evaluation

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

We also encourage all participating teams releasing their codes after challenge.

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 task is optional task and NOT counted into the final leaderboards.

MISSION OF THE CHALLENGE

Field(s) of application

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

Examples:

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

Decision support.

Task category(ies)

State the task category(ies).

Examples:

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

Localization.

Cohorts

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

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

Patients visiting eye hospital.

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

Patients visiting eye hospital.

Imaging modality(ies)

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

Retinal fundus image.

Context information

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

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

Retinal fundus image.

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

None.

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.

Retinal fundus image.

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

Localization of Fovea in fundus image.

Assessment aim(s)

Identify the property(ies) of the algorithms to be optimized to perform well in the challenge. If multiple properties are assessed, prioritize them (if appropriate). The properties should then be reflected in the metrics applied (see below, parameter metric(s)), and the priorities should be reflected in the ranking when combining multiple metrics that assess different properties.

- Example 1: Find highly accurate liver segmentation algorithm for CT images.
- Example 2: Find lung tumor detection algorithm with high sensitivity and specificity for mammography images.

Corresponding metrics are listed below (parameter metric(s)).

Accuracy.

Additional points: The evaluation criterion is the Average Euclidean Distance between the estimations and ground truth, which is the lower the better.

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).

Zeiss Visucam 500, Canon CR-2, topcon tirton, kowa.

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

-Dimension: 2D
-Timepoints: 1-time
-Position and orientation of the patient: sitting upright

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.

Zhongshan Ophthalmic Center, Sun Yat-sen University, China.

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).

Original fundus images captured by ophthalmologists and technicians;
Annotated with clinical diagnosis/records with multi-modality information.

Training and test case characteristics

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

Examples:

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

Training and test cases both represent a fundus image with the pixel-wise labels.

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

A total of 2000 color fundus photographs are available. The dataset is split into 3 subsets for training (1200), offline validation (400) and onsite test (400), stratified to have equal glaucoma presence percentage.

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

1200 training images are from previous REFUGE challenge as training data. 400 offline validation and 400 onsite test images in REFUGE2 are additional data collecting by the different fundus camera modalities, which have no overlapping with previous challenge.

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.

Class distribution in classification task is 2:8 (glaucoma vs. non-glaucoma).

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.

Manual pixel-wise annotations of the fovea (macular center) were obtained by 7 independent GLAUCOMA SPECIALISTS. The reference standard for localization task was created by using the average of selected annotations from the 7 annotations, for each individual images by another independent GLAUCOMA SPECIALIST.

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 procedure consisted in manually drawing a tilted ellipse covering the macular region, by means of a free annotation tool with capabilities for image review, zoom and ellipse fitting. A single segmentation per image was afterwards obtained by taking the majority voting of the anotations of the seven experts. A senior specialist with more than 10 years of experience in glaucoma performed a quality check afterwards, analyzing the resulting masks to account for potential mistakes. The center coordinates of the mask are used as the final ground-truth.

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.

Manual annotations of the macular center were provided by seven independent glaucoma specialists from the Zhongshan Ophthalmic Center (Sun Yat-sen University, China), with an average experience of 8 years in the field (ranging from 5 to 10 years). All the ophthalmologists independently reviewed and delineated the macular center in all the images, without having access to any patient information or knowledge of disease prevalence in the data.

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.

Majority vote.

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.

- Quality control: only high-quality images are selected for manual annotation;
- Remove personal information and device information;
- 7 independent annotations from ophthalmologists;
- Unified ground truth were determined by two other ophthalmologists for every image, by 1) remove low-quality annotations; 2) average chosen annotations.

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.

-Inter-/ intra-observer variability.
However, in total 9 ophthalmologists are involved in the annotation process, to make the reference as accurate as possible.

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

None.

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)

The evaluation criterion is the Average Euclidean Distance (AED) between the estimations and ground truth, which is the lower the better.

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

AED is a common metric for localization evaluation.

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.

Please NOTE that Task 3 (fovea localization) is OPTIONAL for MICCAI ONSITE challenge and NOT counted into the final leaderboards. However, it is ESSENTIAL for the ONLINE challenge on the TEST dataset.

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

Missing result will report an error information.

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

Please NOTE that Task 3 (fovea localization) is OPTIONAL for MICCAI ONSITE challenge and NOT counted into the final leaderboards. Fovea is an important part of fundus image. We propose this task for optional, which will not be counted into the final ranking. But we will provide a separate rank page for online challenge. Thus, the final challenge ranking is still based on Tasks 1&2.

Statistical analyses

a) Provide details for the statistical methods used in the scope of the challenge analysis. This may include

- description of the missing data handling,
- details about the assessment of variability of rankings,
- description of any method used to assess whether the data met the assumptions, required for the particular statistical approach, or
- indication of any software product that was used for all data analysis methods.

Missing data not allowed, and will receive an error notice. And significance of differences in the rankings will be tested by Wilcoxon Mann Whitney test.

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

The statistical significance of the differences in performance of the top-ranked teams was assessed by means of Wilcoxon signed-rank tests (= 0.05). The statistical significance of the differences between groups was assessed using a Wilcoxon rank-sum test due to the unpaired nature of the two sets (360 vs. 40 samples, respectively) for avoiding outlier data.

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.

None.

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.

J. I. Orlando et al., "REFUGE Challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs," Med. Image Anal., vol. 59, p. 101570, Jan. 2020.