WeLineation: crowdsourcing delineations for reliable ground truth estimation

Goel, Saksham, Sharma, Yash, Jauer, Malte-Levin, Deserno, Thomas

Saksham Goel, Yash Sharma, Malte-Levin Jauer, Thomas M. Deserno, "WeLineation: crowdsourcing delineations for reliable ground truth estimation," Proc. SPIE 11318, Medical Imaging 2020: Imaging Informatics for Healthcare, Research, and Applications, 113180C (2 March 2020); doi: 10.1117/12.2551279

SPIE.

Event: SPIE Medical Imaging, 2020, Houston, Texas, United States

WeLineation: Crowdsourcing delineations for reliable ground truth estimation

Saksham Goel[a], Yash Sharma[a], Malte-Levin Jauer[b], and Thomas M. Deserno[b]

[a]Indian Institute of Technology Bombay, Mumbai, India
[b]Peter L. Reichertz Institute for Medical Informatics of TU Braunschweig and Hannover Medical School, Braunschweig, Germany

ABSTRACT

Crowdsourcing is a concept to encourage humans all over the world to generate ground truth for classification data such as images. While frameworks for binary and multi-label classification exist, crowdsourcing of medical image segmentation is covered only by few work. In this paper, we present a web-based platform supporting scientists of various domains to obtain segmentations, which are close to ground-truth references. The system is composed of frontend, authentication, management, processing, and persistence layers which are implemented combining various javascript tools, the django web framework, an asynchronous celery task, and a PostgreSQL database, respectively. It is deployed on a kubernetes cluster. A set of image data accompanied by a task instruction can be uploaded. Users can be invited or subscribe to join in. After passing a guided tutorial of pre-segmented example images, segmentations can be obtained from non-expert users from all over the world. The Simultaneous Truth and Performance Level Estimation (STAPLE) algorithm generates estimated ground truth segmentation masks and evaluates the users performance continuously in the backend. As a proof of concept, a test-study with 75 photographs of human eyes was performed by 44 users. In just a few days, 2,060 segmentation masks with a total of 52,826 vertices along the mask contour have been collected.

Keywords: Crowdsourcing, Segmentation, Delineation, Gamification, Ground truth, STAPLE

1. INTRODUCTION

Segmentation remains a great challenge in medical image processing, and artificial intelligence needs training data with appropriate ground truth. To reach acceptable performance, reliable solutions for real-world application still require manual segmentation or task-specific annotations of large training datasets by domain experts.[1] Both options are limited by the significant amount of time to spend or costs to pay.

The increasing engagement of crowdsourcing[2–4] tries to cope with this bottleneck. While several attempts have been made to build crowdsourcing frameworks for binary and multi-label classification tasks, the area of crowdsourcing in medical image segmentation still is relatively unexplored. However, it has been acknowledged already to potentially provide accessible health care at lower costs. Crowdsourcing is described as viable solution for medical image annotation, although only few work reports on the precise setup of pilot experiments including parameters, such as the number of annotators per image or an optimal user incentive.[5]

Therefore, we developed a web-based crowdsourcing application that allows researchers to obtain close to ground-truth segmentations of their medical image data set and performed a proof-of-concept study to describe relevant system parameters.

2. METHODS

The general system configuration comprises of (i) a web frontend including the point-based segmentation tool, (ii) an authentication mechanism for access control, (iii) a management component to handle segmentation tasks and image-to-user assignments, (iv) a processing component that asynchronously computes the Simultaneous Truth

Further author information: (Send correspondence to Malte-Levin Jauer.)
E-mail: malte-levin.jauer@plri.de, Telephone: 49 531 391 9504

Medical Imaging 2020: Imaging Informatics for Healthcare, Research, and Applications, edited by Po-Hao Chen, Thomas M. Deserno, Proc. of SPIE Vol. 11318, 113180C
 · doi: 10.1117/12.2551279

Table 1: Open Source Libraries Used

WeLineation component	Software libraries
Frontend	Bootstrap MD, FontAwesome, hover.js
Annotation Tool	PaperJS, Simplify.js, jQuery
Dashboard, Calendar	charts.js, pace.js, progressbar.js, D3.js
Tutorial	intro.js
Data plotting	matplotlib
Authentication	Django Authentication
Processing	Python SciPy, NumPy, Pillow
Management, Gateway, Load balancing	gunicorn
Static files	nginx
Request handler	Django web framework
Distributed asynchronous task management	Celery
Message broker, Cache management	Redis
Container management	Docker
Persistence, Database	PostgreSQL
Deployment and Container Orchestration	Kubernetes, GitLab CI/CD

and Performance Level Estimation (STAPLE) algorithm,[6] (v) a persistence layer for the storage of segmentation masks, meta- and user data, and (vi) an export layer to allow usage of generated segmentations in post processing computations, e.g. deep learning frameworks, or statistical evaluations for further diagnosis (Fig. 1).

For the implementation of these components, we combine open source libraries, such as javascript tools, the Django web framework, asynchronous celery, PostgreSQL, and deploy the application on a kubernetes cluster (Tab. 1). The system is accessible at: https://welineation.plri.de.

2.1 Frontend

The frontend includes all web pages available to the user. Most importantly, these are the dashboard and the vertices-based annotation tool.

2.1.1 Dashboard

After logging in, the user is presented the dashboard, where the available segmentation tasks are shown (Fig. 2). Furthermore, users are visualized their progress in the specific task, their overall segmentation performance and

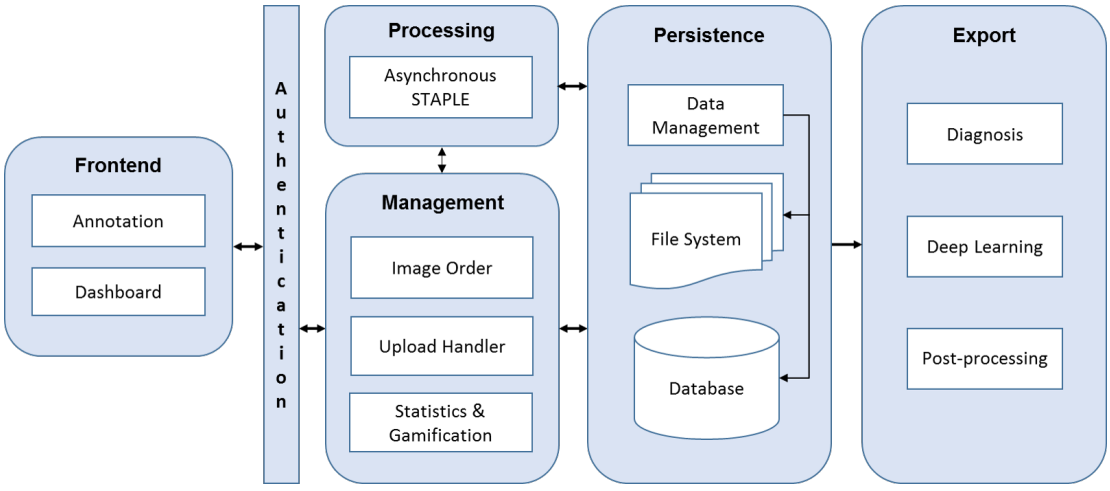


Figure 1: Block diagram of system components

a menu bar to navigate within the application. Unless the tutorial of a task has been passed successfully, it is started when the user selects the corresponding task. Then, the segmentation page is displayed. The user is also presented with an activity calendar indicating the number of contributions per day.

2.1.2 Annotation

The annotation tool provides a view of the images to be segmented manually (Fig. 3a). In order to support various input devices such as notebook, touch-pad, tablet, mouse, pen, etc., there are two input modes to draw a mask on top of the image:

1. clicking individual vertices along the contour. The vertices are then interconnected to form a closed contour.
2. clicking the mouse button and holding it while following along the segment's contour. In this case, vertices are generated continuously. Then, the Ramer–Douglas–Peucker algorithm[7] is applied to reduce redundant points and simplify the point-based representation of the deliniation curve (Fig. 3b).

In all cases and stages, users are able to add, remove, and relocate points. During all interactions, the contour is smoothed to a cubic curve using Bézier interpolation (Fig. 3c). Depending on the task's configuration, one or more contours can be drawn per user and image. Continuing to the next image or leaving the task automatically saves the annotation. Both the segmentation points as well as a binary segmentation mask are stored in the WeLineation system.

2.2 Management

The WeLineation platform is designed to support various tasks and users simultaneously. It therefore requires a certain management layer. This layer is composed of three main functions: Upload handler, image order, and statistics & gamification.

2.2.1 Upload Handler

The upload system supports task masters uploading and managing datasets, task instructions, and performance test images. After uploading the data, a distance metric must be determined. The system provides options including Jaccard, Dice, and modified Hausdorff distances.[8] The distance metric is used for performance tests that can be established optionally for each task and the STAPLE algorithm. If selected, all potential users need to pass the test within an adjustable threshold before they are allowed to enter this task. Tasks can be made available publicly or only to particular users.

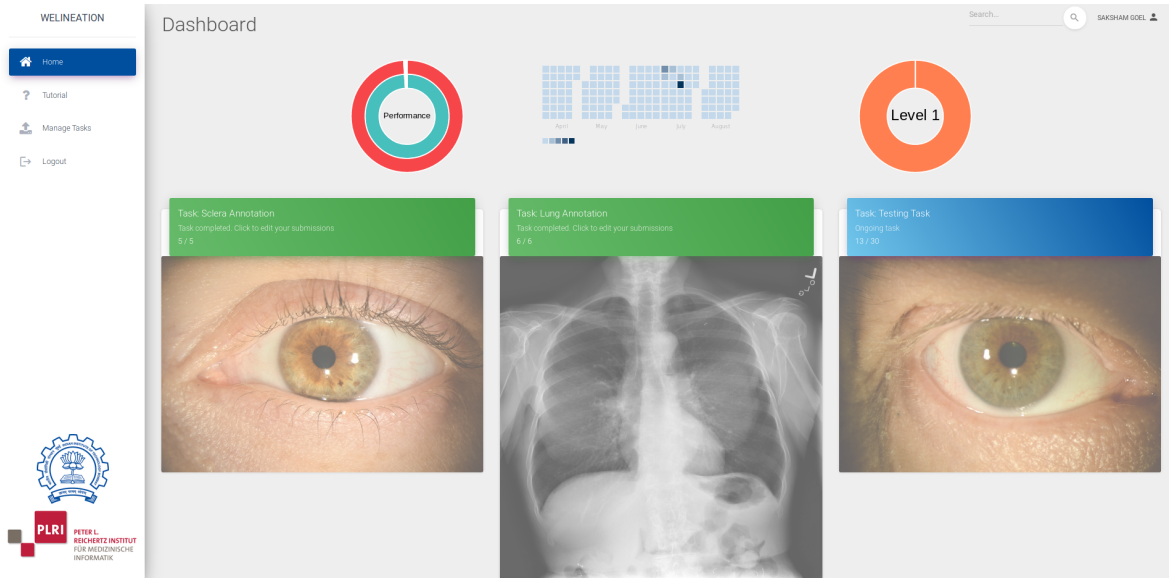


Figure 2: User interface with dashboard

2.2.2 Image Order

When the user selects a specific task and –if required– has passed the performance test, the images to be segmented are assigned in a specific order. Currently, the system supports the assignment of images in

- ascending order by file name,
- descending order by file name, or
- random order.

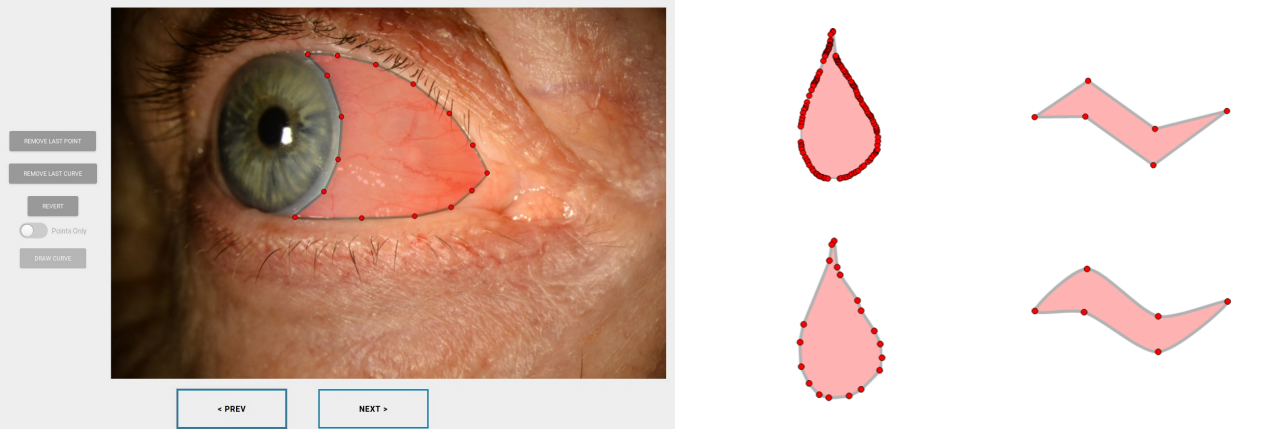
Furthermore, the system can assign only images first that have not been segmented by any user yet. This way, the uploaded dataset gets segmented homogeneously in the process. In later stages, it will be possible to select a user-specific ordering with respect to the user's overall or task-specific performances.

2.2.3 Statistics & Gamification

The WeLineation system currently records metadata from the users' delineations, such as timestamps of the submissions, number of used mask vertices and the users' performances as estimated by STAPLE. As a first step to give an incentive to the user, the average performance is shown on the dashboard. Furthermore, an experience level is increased for each delineation performed by the user and visualized in the task's progress bar and dashboard's pie chart. The level of the users increases according to their performance and contribution to the task. Further gamification steps, such as high-speed racing, accuracy and speed competition, team gaming or named high score, are planned for later stages of the system.

2.3 Authentication

The authentication wall controls access to the platform. It is possible to restrict access to certain datasets only to specific users. The authentication component also differs roles such as user, task master, and system administrator. It allows level-based task access per user.



(a) Annotation tool (b) Point simplification (c) Bézier interpolation

Figure 3: The frontend segmentation tool (a) is using the Ramer-Douglas-Peucker simplification algorithm[7] (b) for the removal of not needed segmentation points. Between the segmentation points, a cubic Bézier interpolation is used (c).

2.4 Processing

The processing part of the WeLineation system continuously estimates the ground truth from all delineation masks gathered from the crowd. The processing is done asynchronously in the backend of the system. Furthermore, the performance of each user is assessed. Both is done simultaneously using the STAPLE algorithm, which is used in many applications.[9,10] Basically, STAPLE is an iterative algorithm that computes a mean segmentation and then evaluates the quality of each user. Based on the user's performance, the weights of users contribution to the mean are adjusted in the next iteration of STAPLE. The ground truth is estimated using an expectation-maximization algorithm. Then, the STAPLE parameters are stored in the WeLineation database and shown in the user's dashboard.

2.5 Persistence

The data management component of the persistence layer is responsible for all stateful data. While images are stored directly on the file system, the database holds the following information:

1. *Account*: This table is managing the login credentials and roles of every user. Also, the average performance is stored persistently.
2. *Task*: This table holds the different delineation tasks uploaded to the system, their settings, and their instructions.
3. *Image*: This table task-specifically links to all images on the file system.
4. *Mask*: This table stores the essential information regarding all delineations performed by the crowd, including position and number of vertices, performance determined by STAPLE, and other meta data per mask.
5. *Segment*: This table is storing references to the binary masks generated by STAPLE, which are stored as image files directly on the file system.

2.6 Export

For further analysis of the results and using the segmentations for diagnostical purposes, the whole dataset of a task can be exported. An export is composed of all delineations as binary mask as well as point sets, the STAPLE results, and the users performances. Additionally, the data contained in the tables mentioned in Section 2.5 are exported.

3. PROOF-OF-CONCEPT STUDY

To provide a proof-of-concept, we performed a study on 75 photographs of human eyes, with the task of sclera segmentation. Depending on the gaze direction of the eye, one or two segments are required per image. the tutorial was composed of four images.

4. RESULTS

Within just a week, 44 users contributed 2060 masks consisting of 52826 vertices in total (Tab. 2). A particular analysis shows that the minimum and maximum number of masks per image is 27 and 37, respectively, and 21 users completed the whole set of images (Fig. 4).

To validate the output from the STAPLE algorithm, we compared the results with manual segmentations for 30 images by experienced ophthalmologists. The average error rate is 1.1% and goes as low as 0.5% on some images. Qualitatively, images are very close to the expert level ground truth (Fig. 5).

We also analyzed the time it takes the users to annotate an image with a segmentation. In this regard, users performed very differently, with a standard deviation of 110.4. The average time to segment an image is 83.8 seconds (Fig. 6). The performance also varies, but does usually not correlate with the time spent to segment an image.

Table 2: Results of the Proof-of-Concept Study

Tasks defined	3
Tasks performed	1
Images in performed task	75
Total users	44
Active users	40
Completing users	21
Masks created	2,060
Vertices placed	52,826
Average Sensitivity	0.910 ± 0.088
Average Specificity	0.995 ± 0.008

5. DISCUSSION

Establishing ground truth for segmentation of medical images is a well known problem in medical imaging informatics. We have developed a web-based system for crowdsourcing segmentations of inexperienced users. A proof-of-concept study was performed that supports the following claims:

1. Less than 30 non-expert users with minimal knowledge in anatomy can create expert-like references.
2. With a careful design of the user interface, the annotation time is about one minute per image for the presented task.
3. Crowdsourcing establishes reliable ground truth on large datasets, which is needed to train machine learning algorithms, e.g. deep neural networks.
4. Therefore, crowdsourcing of image segmentation holds immense potential for medical image processing and computer-aided diagnosis.

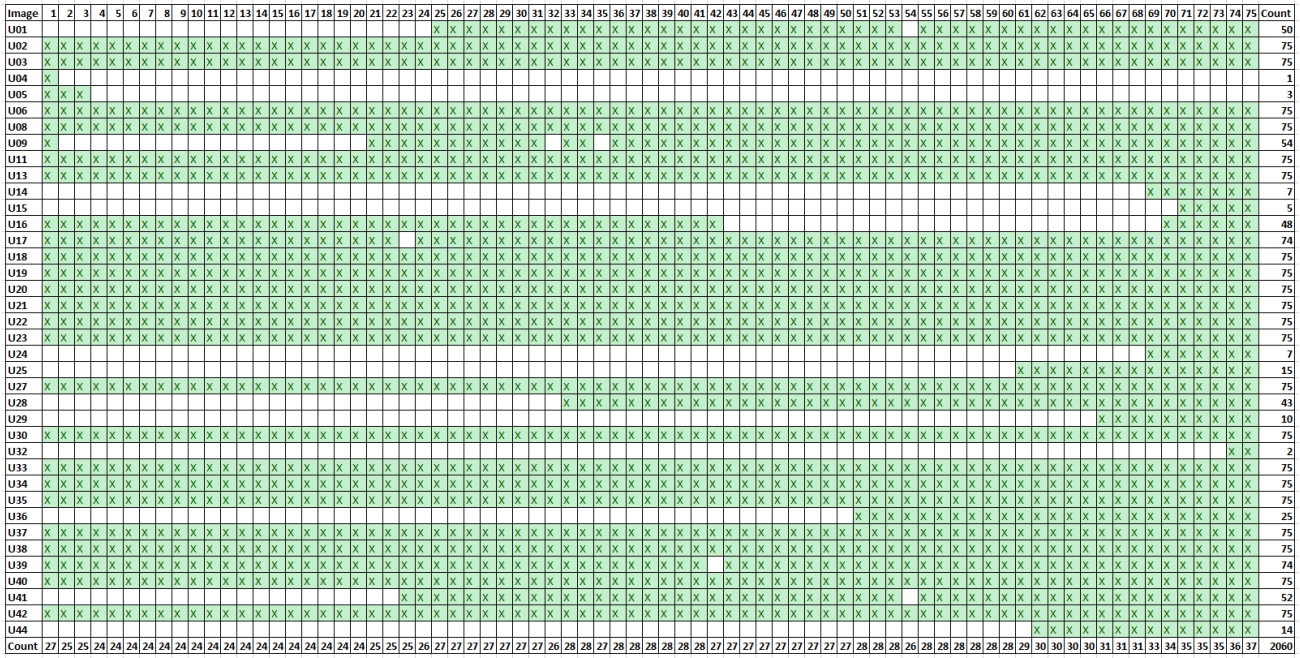
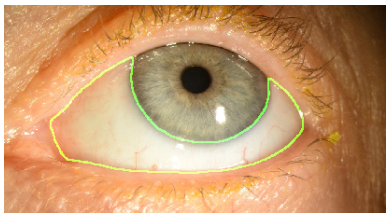
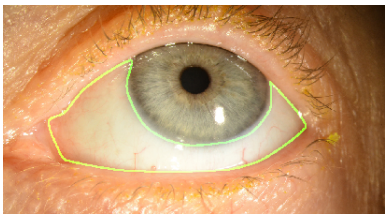


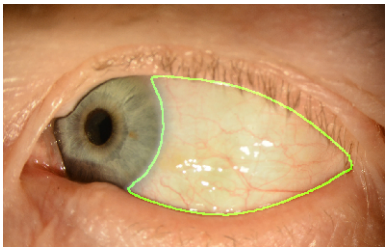
Figure 4: Matrix of user contribution for the set of task images



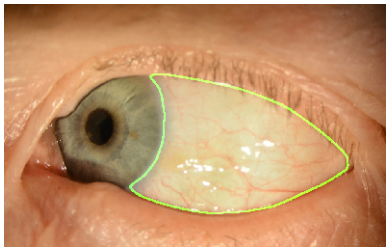
(a) STAPLE-based ground truth



(b) Expert's mask



(c) STAPLE-based ground truth



(d) Expert's mask

Figure 5: Comparing the algorithm to an expert

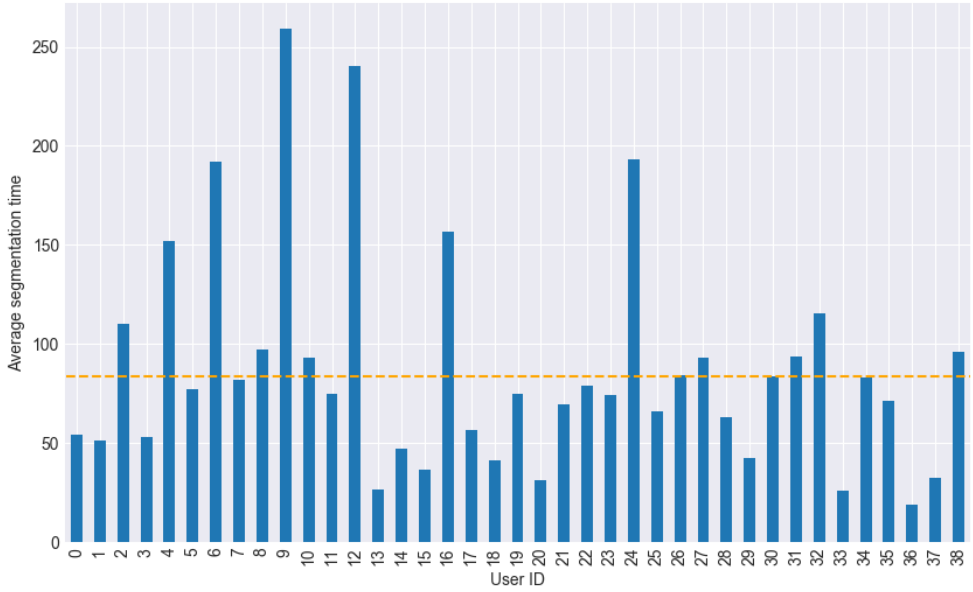


Figure 6: Average segmentation times per user. Each bar shows the average of a certain user over all images he/she segmented. The global average over all users is 83.8 seconds (orange dotted line).

In future, we plan to add different segmentation assignments incorporating the users' performances, different incentives (gamification, challenges, etc) and to evaluate how the crowd reacts to these measures. Also, tasks with pre-segmented images can be designed, where users are asked to only perform a fine tuning of the already given vertices. Furthermore, the processing unit may be extended with machine learning components to estimate the vertices initialization.

REFERENCES

[1] Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis. Med Image Anal 2017;42:60 – 88.

[2] Buhrmester M, Kwang T, Gosling SD. Amazon's Mechanical Turk: a new source of inexpensive, yet high-quality, data? Perspect Psychol Sci 2011;6(1):3–5.

[3] Russell BC, Torralba A, Murphy KP, Freeman WT. LabelMe: a database and web-based tool for image annotation. Int J Comput Vision 2008;77(1-3):157–173.

[4] Inel O, Khamkham K, Cristea T, Dumitrache A, Rutjes A, van der Ploeg J, et al. Crowdtruth: machine-human computation framework for harnessing disagreement in gathering annotated data. In: Mika P, Tudorache T, Bernstein A, Welty C, Knoblock C, Vrandei D, et al, editors. International Semantic Web Conference. Springer; 2014. p. 486–504.

[5] Wazny K. Applications of crowdsourcing in health: an overview. J Glob Health 2018;8(1):010502.

[6] Warfield SK, Zou KH, Wells WM. Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. IEEE Trans Med Imaging 2004;23(7):903.

[7] Hershberger JE, Snoeyink J. Speeding up the Douglas-Peucker line-simplification algorithm. University of British Columbia, Department of Computer Science; 1992.

[8] Dubuisson MP, Jain AK. A modified Hausdorff distance for object matching. In: Proceedings of 12th international conference on pattern recognition. vol. 1. IEEE; 1994. p. 566–568.

[9] Archip N, Jolesz FA, Warfield SK. A validation framework for brain tumor segmentation. Acad Radiol 2007; 14(10):1242–1251.

[10] Xu Z, Asman AJ, Singh E, Chambless L, Thompson R, Landman BA. Collaborative labeling of malignant glioma. In: Proc IEEE Int Symp Biomed Imaging; 2012. p. 1148–1151.