PROCEEDINGS OF SPIE

SPIEDigitalLibrary.org/conference-proceedings-of-spie

Gamification concept for acquisition of medical image segmentation via crowdsourcing

Jauer, Malte-Levin, Spicher, Nicolai, Deserno, Thomas

Malte-Levin Jauer, Nicolai Spicher, Thomas M. Deserno, "Gamification concept for acquisition of medical image segmentation via crowdsourcing," Proc. SPIE 11601, Medical Imaging 2021: Imaging Informatics for Healthcare, Research, and Applications, 116010D (15 February 2021); doi: 10.1117/12.2582259

SPIE.

Event: SPIE Medical Imaging, 2021, Online Only

Gamification concept for acquisition of medical image segmentation via crowdsourcing

Malte-Levin Jauer[a], Nicolai Spicher[a], and Thomas M. Deserno[a]

[a]Peter L. Reichertz Institute for Medical Informatics of TU Braunschweig and Hannover Medical School, Mühlenpfordtstr. 23, Braunschweig, Germany

ABSTRACT

In many fields of medical imaging, image segmentation is required as a basis for further analysis and diagnosis. Convolutional neural networks are a promising approach providing high accuracy. However, large-scale annotated datasets are necessary to train these networks. As expert annotations are costly, crowdsourcing has shown to be an adequate alternative. In previous work, we examined how the workforce of a crowd should be distributed for obtaining annotations with an optimal trade-off between quantity and quality. In this work, we present a gamification approach by transforming the tedious task of image segmentation into a game. This approach aims at motivating users by having fun but nevertheless generating annotations of adequate quality. Therefore, this work presents a gamified crowdsourcing concept for medical image segmentation. We give an overview of incentives applied in state-of-the-art literature and propose two different gamification approaches on how the image segmentation task can be realized as a game. Finally, we propose a integrated game concept that combines both approaches with the following incentives: (a) points / scoring to reward users instantly for accurate segmentations, (b) leaderboard / rankings to let users accumulate scores for long-term motivation, (c) badges / achievements to give users a visual representation of their "strength" in segmentation, and (d) levels to visualize the learning curve of users in performing the segmentation. We give details on a first prototype implementation and describe how the game concept complies with the guidelines from our prior work.

Keywords: Gamification, Serious games, Crowdsourcing, Image segmentation, Convolutional neural networks, Training data

1. INTRODUCTION

Image segmentation is a frequently encountered task in many image processing pipelines and yields coherent areas in an image. In medical imaging, segmentation is required in multiple applications, such as the extraction of anatomical structures. Another application is the detection of regions of interest (ROI) on the surface of the human body acquired with digital cameras. This work is concerned with the segmentation of the human eye, more specifically, the delineation of the sclera (ROI). This reduces the segmentation task to a binary decision for each pixel in the image. Either the pixel belongs to the sclera, or not.

Fig. 1 shows an example image of a human eye (left side) and results of a prematurely terminated segmentation process by a non-expert user (right side): The blue area represents the region that the user correctly annotated as belonging to the ROI (true positive) while the green area represents the region that belongs to the ROI but was not annotated (false negative). Vice versa, the orange area represents the region that does not belong the ROI and the user correctly did not annotate it (true negative). The red area represents the region not belonging to the ROI but the user wrongly assigned it as belonging to the ROI (false positive).

While manual segmentation of medical images can be realized using various tools, this approach is often financially too expensive due to the costs of human experts.

Further author information:
M.J.: E-mail: malte-levin.jauer@plri.de
N.S.: E-mail: nicolai.spicher@plri.de (Corresponding author)
T.M.D.: E-Mail: thomas.deserno@plri.de

Medical Imaging 2021: Imaging Informatics for Healthcare, Research, and Applications
edited by Thomas M. Deserno, Brian J. Park, Proc. of SPIE Vol. 11601, 116010D
 · doi: 10.1117/12.2582259

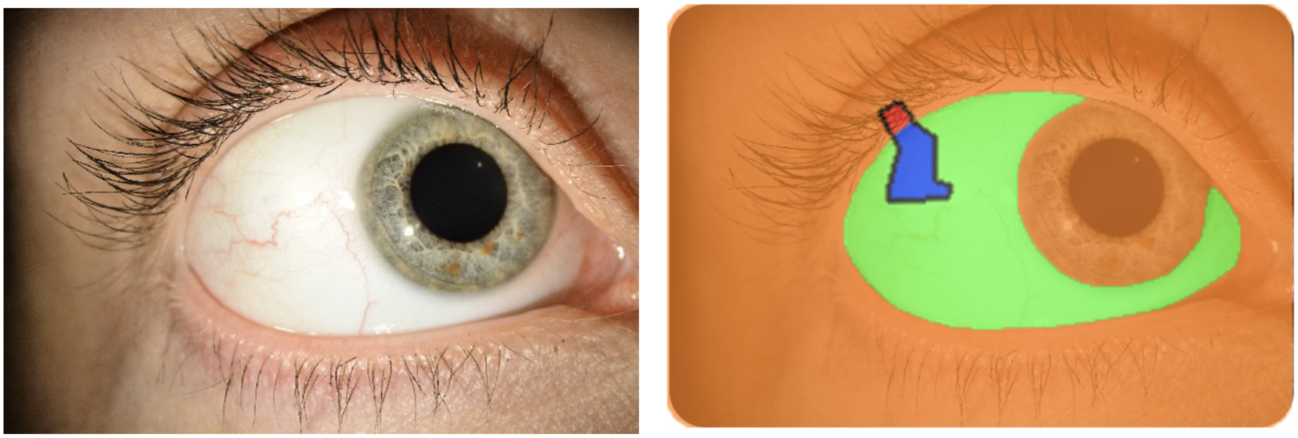

Figure 1. Image of a human eye, as exemplarily used for this work (left). Exemplary comparison of a user annotation with a ground-truth (right): true positives (blue), false positives (red), true negatives (orange), false negatives (green).

These experts are needed to create a ground-truth (GT) dataset consisting of annotations of high quality, i.e., a high number of true positive/negative pixels and a low number of false positive/negative pixels. This is especially critical for machine learning applications requiring large amounts of data for training. To eventually overcome the dependence on manual annotations, automatic algorithms have been developed. Currently, classical pixel- or texture-based approaches are often superseded by deep learning approaches,[1] using convolutional neural networks (CNNs) like U-Net[2] or SegNet.[3] Although these networks have the capability to handle complex segmentation tasks, they do not solve the problem of high costs for acquiring GT datasets.[4]

It has been shown that the concept of "crowdsourcing" can widen this bottleneck. The expensive annotations by a small group of experts are substituted by using a large group of non-experts[5–7] and multiple annotation per image from several non-experts are averaged to generate a consensus annotation of higher quality.

In previous work,[8] we investigated how CNNs can be trained effectively using crowd-sourced sclera annotations. We used training datasets of different sizes and quality levels for CNN training and analyzed the impact on CNN prediction accuracy. We gained the following insights, which we call *crowdsourcing requirements*:

- **R0**: The quality of annotations (true positive/true negative pixels) should be maximized
- **R1**: For each image, at least three annotations should be obtained
- **R2**: The number of annotated images should be maximized
- **R3**: The number of images annotated per user should be maximized (long-term motivation)

Evidently, R0 is a key requirement for any CNN training and therefore given for the sake of completeness only. With regard to R3, we observed from literature and previous investigations, that the qualitatively best annotations can be obtained from experienced (long-term) users. Therefore, although implied by R2, we added this aspect as an independent requirement.

In this work, we propose a concept how these requirements R0-R3 can be met using the means of gamification.

1.1 Background

The term crowdsourcing was first used in 2006 to describe a large group of people completing a task collaboratively which exceeds the capabilities of a single person.[5] It combines the availability of many individuals to efficiently fulfill task-based work as a distributed workforce. Crowdsourcing has not been consistently defined throughout literature but one can distinguish between three general types of crowdsourcing:

- The first type is called *citizen science*,[9] which is based on people assisting voluntarily with the intention to progress scientific research. Most volunteers have a personal interest in contributing to a specific research topic. For example, Zooniverse ([*]) is a well-known platform where people can publish citizen science projects like the classification of galaxies or animal species. In previous work,[10,11] we developed a platform called WeLineation[†] following the citizen science concept for segmentation.

- Another type of crowdsourcing is so-called *micro task based crowdsourcing* (MTBC) which follows a different approach by paying the crowd for the completion of tasks. This enables to recruite persons that do not have an intrinsic motivation. Commercial MTBC platforms such as Amazon MTurk ([‡]) and Microworkers ([§]) are already available. Several works reported satisfactory results using MTBC for obtaining medical image annotations.[12]

- The last type of crowdsourcing is based on *games with a purpose* (GWAP) which are also known as serious games. These games present the task to the player in an entertaining way.[13] Thereby a task is solved not by social motivation (citizen science) or financial incentives (MTBC) but only by the player's desire for entertainment.

1.2 Problem

We start our investigation from the perspective of a data scientist having a large image dataset that needs to be segmented. Furthermore, we assume that the dataset is too large for hiring experts for manual segmentation as ressources are limited. Additionally, we assume that "classical" algorithms for segmentation are not available or their accuracy is insufficient. In precedent work,[8] we proposed employing a deep learning framework which has proven suitable if the training dataset is adequate. The problem at hand consists in generating this GT training dataset with limited financial ressources.

In previous work, we already investigated the approach of *citizen science* for medical image segmentation.[10] We motivated students without prior knowledge or participants with some personal relation to the project. We experienced that although annotation quality was satisfactory for most users, only very few users were motivated in the long-term to provide annotations for the whole task.[10]

Kaufmann et al.[14] examined a popular MTBC platform and investigated motivating factors to work on paid crowdsourcing. Although it is reported that many workers mention the importance of payment, Kaufmann et al.[14] conclude that intrinsic factors often outrival the extrinsic ones when it comes to spending time on the platform. This highlights the importance of entertainment while doing a task for long periods. Motivation can also be increased when participants get the feeling of learning something. Although the results reported by Kaufmann et al.[14] also dependent on demographic values, they motivate the use of a gamified approach.

One advantage of the gamification method is that participants play the game for fun and not for money. In case the data (annotations) acquired by the users has an adequate quality for serving as GT, this method yields a good cost-benefit ratio. Additionally, literature reports several positive effects of gamification with respect to our defined crowdsourcing requirements. First of all, studies directly comparing to a non-gamified approach mention an increase of long-term engagement (R3).[15] Furthermore, increased quality of output[16,17] and less cheating, i.e. not performing a task adequately, is reported[16] (R0).

Therefore, we select the approach of gamification and pose the following research question: *How to design an image segmentation game that complies with defined crowdsourcing requirements using state-of-the-art gamification methods?*

[*] http://zooniverse.org
[†] http://welineation.plri.de
[‡] http://www.mturk.com
[§] http://www.microworkers.com

2. RELATED WORK

2.1 Game design guidelines

The success of gamification has its roots in the intrinsic motivation of users playing the game. This section outlines, which incentives are proposed in literature to reach this goal. Morschheuser et al.[18] reviewed gamification approaches of literature. Out of the ten identified commonly used incentives for gamification, they identified four most commonly used:[18]

Scores: Scores shall be used as a metric to measure the success of the players. Using this metric, the players are rewarded. The nature of the reward depends on the application-specific scoring mechanism. For applications, where a lot of simple tasks have to be performed, the score is usually increased in relation to the quantity of completed tasks. In cases, where the quality of a completed task can be measured, the score is usually linked to quantity and quality. The quality can be calculated if the expected result is already known. In most cases however, the result is unknown and the score is calculated as the level of agreement to the results of other players.[18] It has also been shown that the scoring can be combined with other game mechanisms, e.g., using time limits[19,20] to realize a more engaging experience.

Leaderboards: The concept of a leaderboard describes a visual representation of player's scores in relation to the other players. Usually, the players are ranked using their current score or the sum of scores obtained from playing the game multiple times. Although leaderboards are used by a majority of the gamification approaches, the success depends on the application. Some studies reported increased player motivation due to the competitive nature induced by leaderboards. However, also contrary reports exist, where player motivation was reduced. Morschheuser et al. noted that in most of these cases, the usage of all-time leaderboards led to frustration of new players that get the feeling of never being able to reach the top.[18] However, in some applications it might be more important to keep a small number of players (the top players of the leaderboard) motivated in the long run, as these players can be trusted to obtain good results. In this regard, long-term leaderboards could be the right choice.[18]

Badges / Achievements: Badges visually represent that a player has successfully finished a specific task.[15] For example, badges are icons displayed next to the player's name or in-game displayed on the players avatar. Sometimes, a list of the player's achievements is displayed. In a nutshell, badges are a means of steering the motivation of a player towards a specific goal, e.g., solving a task within a certain time limit. It is also common to reward a badge for "wins in a row".

Level systems: Level systems are used to reward the effort and success of the user over a long time.[13] In many cases, some game mechanics are "unlocked" only after reaching a certain level. Levels may change the visual appearances of the player's avatar, making success visible to others which increases competitiveness and reward of success.[18]

3. MATERIALS AND METHODS

In the following, the individual components of the gaming concept are introduced: First, game mechanics have to be devised that allow one or multiple players to interact with a given image and result in an annotation of this image. We propose two different approaches: 1) A naive game for a single player and 2) a more elaborated implementation for a competitive multiplayer game. Second, we propose multiple gamification incentives based on the literature review. We propose *Dr. Columbus*, a game concept integrating mechanics and incentives.

3.1 Segmentation acquisition

3.1.1 Naive segmentation game

A naive approach to transform the task of image segmentation into a game would be a direct mapping of the task and only changing the user interface. Typically, image segmentation is performed by using computer programs which display the image in high resolution and provide tools, e.g. a lasso function, for delineating the different regions with the help of the computer mouse. We transform this task to a game by displaying the detail of an image in the background to the user. In front the cursor is shown and is controlled by the W/A/S/D keys of the keyboard which allows to move the icon and to see other parts of the image.

The task of the user is to run along the complete sclera. Once the cursor returns to the start location, the game is finished. The enclosed area between start and end location is stored as belonging to the sclera while all other area is stored as not belonging to the sclera.

Pros and Cons: This game allows to acquire annotations efficiently. However, it does not meet the crowd-sourcing requirements and gamification criteria: There is no long-term motivation, no competitive incentive to increase the number of annotations obtained per player, and no apparent reason for the player to perform the annotation with good quality.

3.1.2 Elaborated segmentation game

The prior game concepts has several shortcomings. The online game *Paper.io 2*[¶] which served as a reference for the following concept.

Similar to the naive approach, the image to be segmented is drawn in the background. The game supports multiple players which are all represented by an icon in the foreground with controls identical to the naive approach. Each player moves his icon and tries to occupy the largest possible area by drawing a line behind him and increasing his own user area once he has drawn a full circle. Furthermore, a player can "steal" the area of an opponent by driving through the opponents territory. However, while leaving his/her own area, a player is "vulnerable" and opponents can pass through his line which ends the game for this player. The game ends when only one player is left. Additionally, acquiring territory in the region representing the sclera (Fig. 1 (green area)) the player gains points, whereas a region not associated with the sclera (orange/red area) will cost the player points.

Pros and cons: This game overcomes some disadvantages of the naive concept. It motivates the users to participate in the game frequently and for a longer term. However, the resulting segmentation annotations are influenced by the interaction with the opponents. This results in a potential lower quality of the segmentations and presumably the image is not fully delineated by a single player due to the competitive multiplayer nature.

3.2 Scoring

The scoring mechanism is an essential component of the gamification process. To score a segmentation, we assume to have a GT available, e.g., acquired from an expert.

With respect to this GT, we can determine true and false positives as well as true and false negatives from the user annotation. In our application, the sclera of the patient's eye represents the "positive" area of the map (green and blue in Figure 1). Collecting user area in this space rewards the player with points. This collected area is classified as true positive (TP) pixels in the segmentation (blue in Figure 1).

Naturally, we want to reward TP pixels and penalize false positives (FP) pixels. Therefore, we substract wrongly annotated pixels from the right ones. To get a measure independent from the size of the GT, we relate the result to all pixels classified as "sclera" in the GT (TP + FN). Exemplary, we define our base score as 500, e.g., the number of points that are obtained when the ROI is annotated correctly. As this measure dependents on the quality of annotation, we refer to it as *qualityscore*

$$Q = \frac{TP - FP}{TP + FN} \cdot 500. \tag{1}$$

[¶] https://paperio.site/

Table 1. The game concept adheres to the requirements for crowdsourcing medical image segmentations.

Gamification measure	**Dr. Columbus implementation**	**R0**	**R1**	**R2**	**R3**
Scoring	Player is rewarded for area "discovered". Falsely discovered area decreases the score.	X	X	X	
Leaderboard	Columbus of the day/week/month		X	X	X
Badges / Achievements	Special player icons only available with badges	X			X
Level system	Level displayed next to players name				X

Players need to annotate more correct than incorrect pixels to get a positive score Q. We also reward the effort of each game played with 5 points using an *effortscore* E. This score is additionally increased when playing r multiple games in a row:

$$E = r \cdot 5. \tag{2}$$

The final *userscore* is calculated as the sum of both and is a non-negative value

$$U = max\{0\,,\, Q + E\}. \tag{3}$$

3.3 Leaderboards

Leaderboards motivate players by competing with each other. To avoid frustration of new players, we use three leaderboards for a daily, weekly, and monthly time span, respectively.

Users are ranked with respect to U obtained in the respective period. This way, new users can rise quickly to be the leader of the day. The weekly and monthly leaderboard aim at mid-term motivation of the users.

3.4 Level system

We further introduce levels that stay constant throughout the day, weeks, and months. Levels can be increased after a certain U is acquired. The level of each player is displayed in-game next to his/her username as well as on the leaderboard so that players can compare with each other.

3.5 Badges

As fourth incentive of gamification, we apply the concept of "badges". Earned badges are displayed as a pictogram of a medal in the corresponding color over the player icon in-game. In our case, we do not want the badges to be redundant to the score or level system, which mainly target at motivating the users to stay in the game and annotate more images.

Instead, we use this mechanic to honor for high quality annotations. To this end, the ratio of TP to FP pixels is saved for each game of the player. If this ratio is above 80% for the last 5 annotations of the user, he/she will be awarded with a bronze batch. Silver and gold badges are awarded for a ratio above 90% and 95%, respectively. The feasibility of the given percentages has to be validated experimentally. Note that, in contrast to levels, badges can be lost again, if the annotation quality decreases.

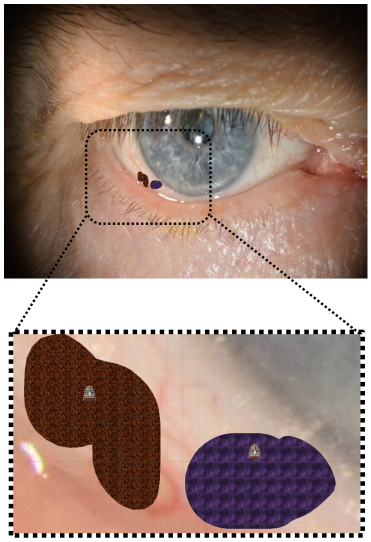

Figure 2. Screenshot of Dr. Columbus.

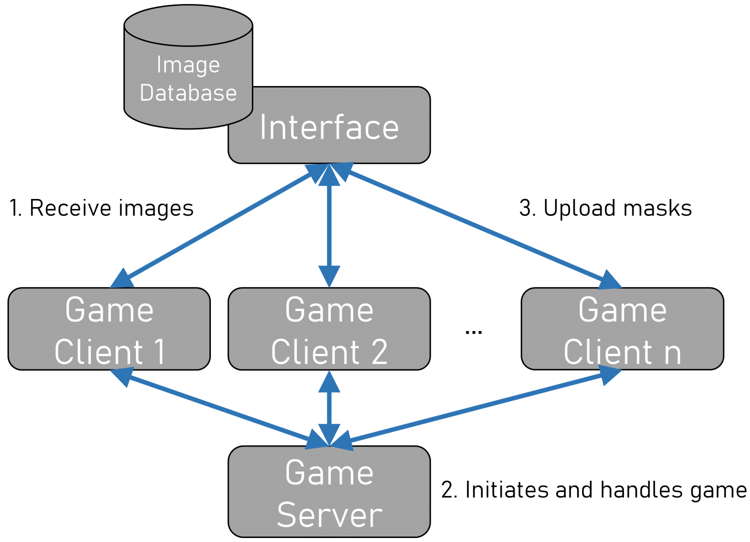

Figure 3. System architecture of Dr. Columbus.

3.6 Integrated game concept: Dr. Columbus

Bringing together the naive and elaborated game mode and applying the presented gamification incentives, we propose the game "Dr. Columbus". The name is motivatived by players "discovering" new areas of an image during annotation. In Dr. Columbus, the elaborated game mode is normally played. Players can learn how to segment the images based on a GT annotation.

Similar to the Q score, but not visible to the players, we calculate a so called "trustworthiness" score, describing how accurate and how reliable a certain player is in the elaborated mode. For players with high trustworthiness, a random session of the naive game is initated between the elaborate game rounds. We expect a player of high trustworthiness to generate a high-quality full segmentation on an image. For the naive mode, 100 points are assigned unconditionally. In this way, we combine the advantages from both game modes by motivating the users by means of gamification, but still generating new annotations as required by crowdsourcing.

4. RESULTS

We described a GWAP concept that uses state-of-the art gamification incentives to meet defined crowdsourcing requirements based on our previous work (see Table 1). Using a scoring incentive, players are rewarded for fulfilling accurate image segmentation while playing a game.

Using GT to check the quality of annotations, fulfills requirement R0. Due to the calculation of the different score and the competitive element of opponents "stealing" points, this incentive motivates players to perform multiple annotations, which contributes to requirements R1 and R2. Using a leaderboard, we motivate players to stay in the game in the mid- and long-term. As this accumulates annotations over a long time, this approach meets requirements R1, R2, and R3. The badge incentive is designed to motivate players to keep their quality of annotation high throughout the game (R0), whereas the level system fosters players long-term motivation by persistent level visualizations.

We developed a prototype with a screenshot being shown in Figure 2 (bottom image). In this scene, a game round of the elaborated segmentation game with two players just commenced. By moving on the screen, players discover new areas of the whole image (top image). The system architecture of the game is depicted in Figure 3. The images to be annotated are stored within a persistent database. At the beginning of a new game, each player/client retrieves the same image from the interface to the image database. Subsequently, the game is initiated and handled by a game server. After the game ended, the clients display the acquired points to the players and upload the generated masks in the background via the interface to the image database.

5. DISCUSSION AND OUTLOOK

Missing or too small GT training data is a bottleneck of applying CNNs in image segmentation. We aim to overcome this bottleneck by using gamification for the acquisition of annotations as training data. Our use-case is the segmentation of the sclera in pictures of the human eye. In previous work[8] we derived requirements for crowdsourcing within this context: The quality of annotations should be maximized; for each image, at least three annotations should be obtained; the number of annotated images should be maximized; and the number of images annotated per user should be maximized.

In this work, we proposed a GWAP concept for gamification of image segmentation. The concept includes two different game modes and supports four gamification incentives. We aim for a study in the near future to evaluate the influence of the incentives on annotation quantity and quality. Additionally, future work will assess how this GWAP concept can be integrated efficiently with other crowdsourcing approaches, i.e. citizen science and MTBC. Furthermore, it would be interesting to analyze wether one of the three approaches is superior for medical image segmentation.

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 Dec;42:60–88.

[2] Ronneberger O, Fischer P, Brox T. U-Net: Convolutional networks for biomedical image segmentation. arXiv:150504597 [cs]. 2015 May;.

[3] Badrinarayanan V, Kendall A, Cipolla R. SegNet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans Pattern Anal Mach Intell. 2017 Dec;39(12):2481–95.

[4] Hesamian MH, Jia W, He X, Kennedy P. Deep learning techniques for medical image segmentation: Achievements and challenges. J Digit Imaging. 2019 Aug;32(4):582–96.

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

[6] Amgad M, Elfandy H, Hussein H, Atteya LA, Elsebaie MAT, Abo Elnasr LS, et al. Structured crowdsourcing enables convolutional segmentation of histology images. Bioinformatics. 2019 Sep;35(18):3461–7.

[7] Grote A, Schaadt NS, Forestier G, Wemmert C, Feuerhake F. Crowdsourcing of histological image labeling and object delineation by medical students. IEEE Trans Med Imaging. 2019 May;38(5):1284–94.

[8] Jauer ML, Wesemeyer T, Deserno TM. Ground-truth quality vs. quantity for training of deep neural networks for medical image segmentation. In: Proc. SPIE Medical Imaging 2021: Imaging Informatics for Healthcare, Research, and Applications. San Diego, California, United States; 2021. .

[9] Meakin JR, Ames RM, Jeynes JCG, Welsman J, Gundry M, Knapp K, et al. The feasibility of using citizens to segment anatomy from medical images: Accuracy and motivation. PLoS ONE. 2019 Oct;14(10):e0222523.

[10] Goel S, Sharma Y, Jauer ML, Deserno TM. WeLineation: Crowdsourcing delineations for reliable ground truth estimation. In: Proc. SPIE Medical Imaging 2020: Imaging Informatics for Healthcare, Research, and Applications. Houston, Texas, United States; 2020. .

[11] Jauer ML, Goel S, Sharma Y, Deserno TM, Gijs M, Berendshot TTJM, et al. STAPLE performance assessed on crowdsourced sclera segmentations. In: SPIE Medical Imaging 2020: Imaging Informatics for Healthcare, Research, and Applications. vol. 11318. International Society for Optics and Photonics; 2020. p. 113180K.

[12] Heim E. Large-Scale Medical Image Annotation with Quality-Controlled Crowdsourcing [Dissertation]. University of Heidelberg; 2018.

[13] Albertarelli S, Dassenno F, Galli L, Pasceri G. The rise of serious games and gamified application in software development. In: 2015 2nd ACM International Conference on Mobile Software Engineering and Systems; 2015. p. 172–173.

[14] Kaufmann N, Schulze T, Veit D. More than fun and money. Worker Motivation in Crowdsourcing – A Study on Mechanical Turk. AMCIS 2011 Proc - Submiss. 2011 Aug;.

[15] Lee TY, Dugan C, Geyer W, Ratchford T, Rasmussen J, Shami NS, et al. Experiments on Motivational Feedback for Crowdsourced Workers. In: ICWSM; 2013. .

[16] Eickhoff C, Harris CG, de Vries AP, Srinivasan P. Quality through flow and immersion: Gamifying crowdsourced relevance assessments. In: Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '12. New York, NY, USA: Association for Computing Machinery; 2012. p. 871–880.

[17] Goncalves J, Hosio S, Ferreira D, Kostakos V. Game of words: Tagging places through crowdsourcing on public displays. In: Proceedings of the 2014 Conference on Designing Interactive Systems. DIS '14. New York, NY, USA: Association for Computing Machinery; 2014. p. 705–714.

[18] Morschheuser B, Hamari J, Koivisto J. Gamification in crowdsourcing: A review. In: 2016 49th Hawaii International Conference on System Sciences (HICSS); 2016. p. 4375–84.

[19] Harris CG. The beauty contest revisited: Measuring consensus rankings of relevance using a game. In: Proceedings of the First International Workshop on Gamification for Information Retrieval. GamifIR '14. New York, NY, USA: Association for Computing Machinery; 2014. p. 17–21.

[20] He J, Bron M, Azzopardi L, de Vries A. Studying user browsing behavior through gamified search tasks. In: Proceedings of the First International Workshop on Gamification for Information Retrieval. GamifIR '14. New York, NY, USA: Association for Computing Machinery; 2014. p. 49–52.