PROCEEDINGS OF SPIE

SPIEDigitalLibrary.org/conference-proceedings-of-spie

STAPLE performance assessed on crowdsourced sclera segmentations

Jauer, Malte-Levin, Goel, Saksham, Sharma, Yash, Deserno, Thomas, Gijs, Marlies, et al.

Malte-Levin Jauer, Saksham Goel, Yash Sharma, Thomas M. Deserno, Marlies Gijs, Tos T. J. M. Berendshot, Christian J. F. Bertens, Rudy M. M. A. Nuijts, "STAPLE performance assessed on crowdsourced sclera segmentations," Proc. SPIE 11318, Medical Imaging 2020: Imaging Informatics for Healthcare, Research, and Applications, 113180K (2 March 2020); doi: 10.1117/12.2551297

SPIE.

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

STAPLE performance assessed on crowdsourced sclera segmentations

Malte-Levin Jauer[a], Saksham Goel[b], Yash Sharma[b], Thomas M. Deserno[a], Marlies Gijs[c], Tos T.J.M. Berendschot[c], Christian J.F. Bertens[c], and Rudy M.M.A. Nuijts[c]

[a]Peter L. Reichertz Institute for Medical Informatics of TU Braunschweig and Hannover Medical School, Braunschweig, Germany
[b]Indian Institute of Technology Bombay, Mumbai, India
[c]University Eye Clinic Maastricht, Maastricht University Medical Center+, Maastricht, The Netherlands

ABSTRACT

The Simultaneous Truth and Performance Level Estimation (STAPLE) algorithm is frequently used in medical image segmentation without available ground truth (GT). In this paper, we investigate the number of inexperienced users required to establish a reliable STAPLE-based GT and the number of vertices the user's shall place for a point-based segmentation. We employ "WeLineation", a novel web-based system for crowdsourcing segmentations. Within the study, 2,060 masks have been delivered by 44 users on 75 different photographic images of the human eye, where users had to segment the sclera. For all masks, GT was estimated using STAPLE. Then, STAPLE is computed using fewer user contributions and results are compared to the GT. Requiring an error rate lower than 2%, same segmentation performance is obtained with 13 experienced and 22 rather inexperienced users. More than 10 vertices shall be placed on the delineation contour in order to reach an accuracy larger than 95%. In average, a vertex along the segmentation contour shall be placed every 81 pixels. The results indicate that knowledge about the users performance can reduce the number of segmentation masks per image, which are needed to estimate reliable GT. Therefore, gathering performance parameters of users during a crowdsourcing study and applying this information to the assignment process is recommended. In this way, benefits in the cost-effectiveness of a crowdsourcing segmentation study can be achieved.

Keywords: Segmentation, Crowdsourcing, Ground Truth, STAPLE, Image Processing, Sclera

1. INTRODUCTION

For several objectives in medical image analysis, a reliable segmentation of regions of interest is necessary. Manual segmentations from experts are often considered as gold standard. Alternatively, either classical pixel- or texture-based segmentations can be used. With current advances in deep learning, the classical algorithms are more and more superseded by convolutional neural networks. However, reliable ground truth for at least a part of the datasets under consideration are essential to assess the segmentation performance or to train algorithms. In real world applications however, ground truth for the data often is missing.

The Simultaneous Truth and Performance Level Estimation (STAPLE) algorithm has been well established in medical image processing as method and mean to provide ground truth estimation for delineation of objects from multiple ratings as well as for comparing the performance of segmentation algorithms without ground truth.[1] The STAPLE algorithm has furthermore been used to determine the best-performing observer as well as to estimate the bias and variance of each rater.[2] It was applied in medical image segmentation to multiple expert segmentations[3,4] and algorithmic segmentations.[5,6] STAPLE was also applied in related domains where GT is missing, as for example in automatic R-wave detection in electrocardiograpy.[7] However, the number of segmentations needed for reliable GT generation has not yet been explored, in particular with respect to domain experts vs. inexperienced users.

Further author information: send correspondence to Malte-Levin Jauer, email: malte-levin.jauer@plri.de, phone: +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, 113180K
 · doi: 10.1117/12.2551297

In previous work, we have established a web-based platform that allows numerous users (the crowd) to provide reference delineation of objects in medical images.[8] In contrast to previous STAPLE investigations, where references were mostly generated by just a few domain experts (one up to three), we examined the results using many ratings per image (30 and more) but generated from mostly inexperienced users.

Therefore, this paper investigates:

1. How many inexperienced users are required to establish a reliable ground truth?
2. How many points are required required to establish a reliable ground truth?

2. METHODS

Our investigation is based on a dataset that was created using the "WeLineation" crowdsourcing platform for segmentation.[8] Within this study, mostly inexperienced users were performing the segmentations, only instructed by a short written description and one guidance figure (Fig 1). The dataset is composed of 75 photographs of human eyes (Fig. 2) with 24 up to 37 reference segmentations (Fig. 3) that have been created by in total 44 users with mostly no specific medical expertise. For a detailed description of the system and user statistics, we refer to the paper of Goel et al.[8]

To assess the number of users required to create reliable segmentation, we designed the following experiment. For each of the 75 images, STAPLE is performed using all the respectively available references, and the result of STAPLE is considered as "ground truth". Afterwards, STAPLE is run using fewer references and the results are compared to the ground truth, averaged over all images of the dataset.

Let G_i denote the ground truth of image i, $i = \{1, \ldots, 75\}$, and $S_{i,j}$ the resulting STAPLE mask for image i taking contributions from users 1 to j with $j = \{1, \ldots, 44\}$. The relative error $E_{i,j}$ is computed as the number of false positive and false negative pixels divided by the number of true pixels:

$$E_{i,j} = \frac{|S_{i,j} \cup G_i| - |S_{i,j} \cap G_i|}{|G_i|} \tag{1}$$

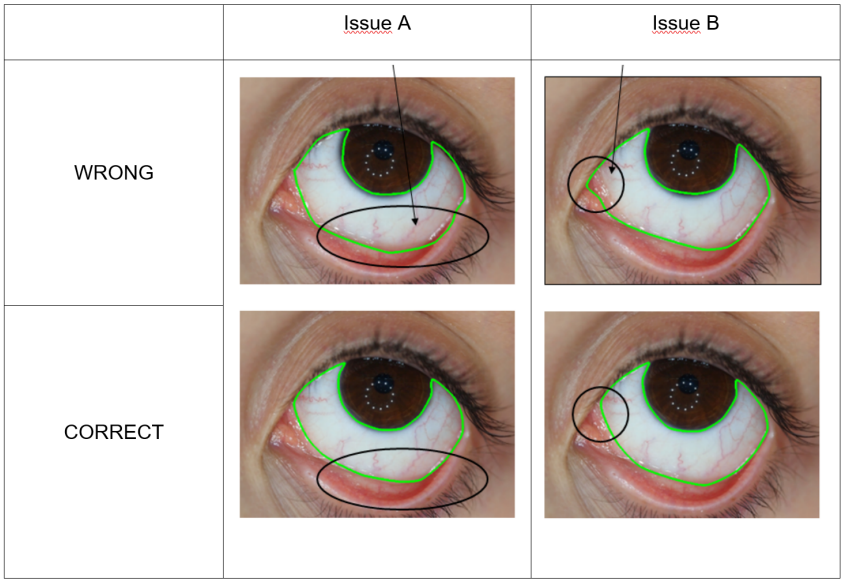

Figure 1. Exemplary images that are given to the study participants in the tutorial. The figure highlights common errors (Issue A and B) frequently done in sclera segmentation.

As the ground truth itself is calculated as a STAPLE result of all user contributions for a particular image, $S_{i,j_{\max}} = G_i$ holds, where $j_{\max}$ denotes the maximum number of segmentations for this image. We assume that adding more references will decreases the relative error, disregarding the particular image i:

$$E_{i,j} > E_{i,k} \qquad \forall \quad k > j \tag{2}$$

To determine the necessary amount of participants, we want the mean error computed over all the images:

$$E_j = \frac{1}{i} \sum_i E_{i,j} \tag{3}$$

to be lower as a given threshold, i.e., 5%, 2%, or 1%. We use the STAPLE result of all reference images to get the estimated performance of each user. Then, we successively compare STAPLE results with fewer segmentation masks ($j < j_{\max}$) by adding users either starting with the best or with the worst performing one, respectively. Hence, we compute a lower and upper bound of the error depending on the overall user performance or "experience level".

To determine the number of points that shall be used, we correlate the number of vertices that have been placed by the user for a certain image with the accuracy of the corresponding mask. The STAPLE algorithm already yields estimates for the specificity and sensitivity of each rater for each image. Let $p_{i,j}$ and $q_{i,j}$ denote the sensitivity and specificity for the segmentation of image i by user j, respectively, the segmentation accuracy $A_{i,j}$ can be determined as follows

$$A_{i,j} = \frac{p_{i,j} \cdot |G_i| + q_{i,j} \cdot (Z - |G_i|)}{|G_i|} \tag{4}$$

where $|G_i|$ denotes the number of pixels in the mask G_i and Z is the number of pixels in the photograph. Let $N_{i,j}$ denote the number of vertices along the contour of user j for image i. Then, we can determine the accuracy of a certain mask as a function of the number of points $A(N)$.

3. RESULTS

As assumed, the relative pixel error decreases with the number of users (Fig. 4), disregarding the order of the users' performance: The blue curve is based on taking users with lowest performance first for the particular image (ascending order); orange indicates using the respective best performers first (descending order). The shown results are averages computed over the whole set of 75 images.

Figure 2. Exemplary images of the dataset created using the WeLineation crowd-source platform.

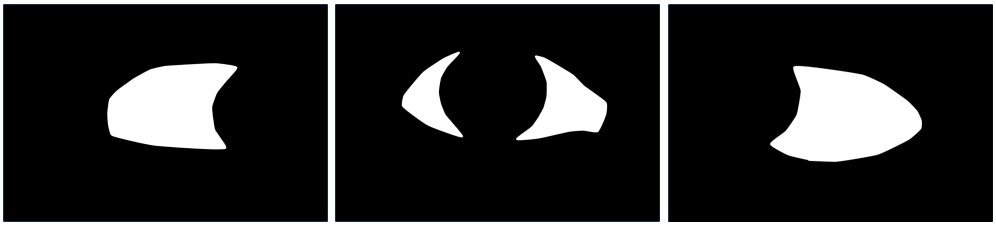

Figure 3. Exemplary crowd-sourced segmentation masks.

The funnel between both curves can be seen as lower and upper error bounds with respect to the user's performance. That means, a horizontal line indicates how many users are required for a relative pixel error E with respect to the user performance. Exemplary, if a 2% error threshold is desired, either 13 users with good performance (intersection with orange curve) or 22 low-performing users (intersection with blue curve) are necessary.

Accordingly, a vertical comparison indicates how much the error varies depending on the user performance. Assuming five segmentations, the error to the estimated ground truth is $0.03 < E_5 < 0.10$.

The mean number of vertices placed by the users within one image is 23.35, with a standard deviation of 7.99. The correlation of the number of vertices that have been placed by the users with the accuracy $A(N)$ of each of the 2,060 masks was calculated and visualized (Fig. 5). Disregarding a few outliers, the accuracy A is larger than 0.95 if the number of vertices ranges between 10 and 40. The mean accuracy for all masks is 0.986 with a standard deviation of 0.011.

Since the number of vertices depends on the contour length, we normalized the means of all masks per image to the length of G_i. Therefore, this scatter plot is composed of 75 dots only (Fig. 6). As visible in the plot, the length between two vertices along the contour in most cases is between 65 and 95 pixel with a mean value of 80.732, whereas we used images of a resolution of $1,200 \times 802$ pixels.

4. DISCUSSION

In this paper, we analyzed the number of non-expert users required and the number of vertices to be placed by a user in order to obtain reliable ground truth with STAPLE. Depending on the performance, reliable ground truth is obtained with 13 to 22 users. This shows that segmentations with low error can be generated with notably smaller number of manual segmentations when using the best performers first.

With respect to the number of vertices, good accuracy ($A > 95\%$) is reached by segmentation masks composed of a wide range of vertices. As a better measure than the number of vertices, the length between two vertices was identified. For a valid segmentation, vertices shall be placed every 81 pixels in average.

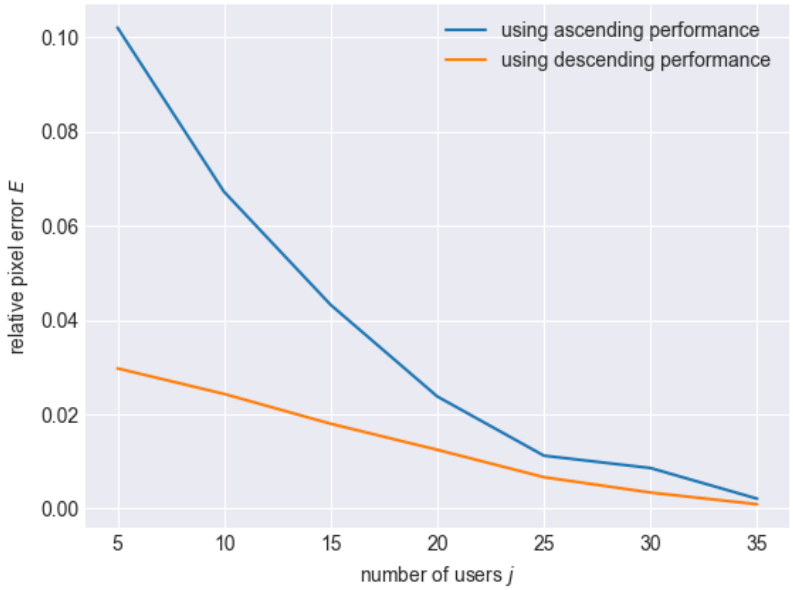

Figure 4. Relative error of STAPLE-based ground truth using different numbers of delineations, averaged over all images.

Although this is the first time that such values have been determined systematically, our study has some limitations. The shape of the eye in the photographs is convex, and results for a objects with other shapes may differ. Furthermore, differences depending on the characteristics of the crowd might exist but were not examined in this work.

However, the obtained relationships can serve as an orientation for crowdsourcing applications, and users can be advised to place a sufficient number of vertices by the system to ensure valid segmentations. Together with the number of necessary user contributions, this will affect the overall effort that has to be performed by the crowd, which in the end is directly related to time (and money) required.

Hence, determining the performance of users is essential already in the process of segmentation. The crowdsourcing system shall present images in a different order to users performing with different quality. Specifically, only well-performing users should be assigned images that have not been segmented by someone else. This way, the segmentation quality can be improved while lowering the necessary number of contributions. For example, to yield relative errors below 5%, 15 segmentations per image are required when the worst performing users contribute first, possibly because their performance is unknown to the system. If in contrast, the performance is known and the best performing users are preferred, only 5 segmentations per image are required to yield even lower relative pixel errors.

In conclusion, our study indicates that a few experts can be substituted by a larger number of novices that do not provide the certain domain knowledge without a loss in general performance, which may impact medical imaging informatics in general.

5. OUTLOOK

During this work, performance analysis was conducted on a per image or per mask basis. However, it is also of interest to observe the performance of distinct users across the whole dataset. In future, we will identify users with a generally better performance and investigate how STAPLE results change, when users with an overall performance are preferred instead users with the best image-specific performance, as it was done so far. Moreover it is of interest, whether characteristics of overall best-performers can be identified and how this knowledge can

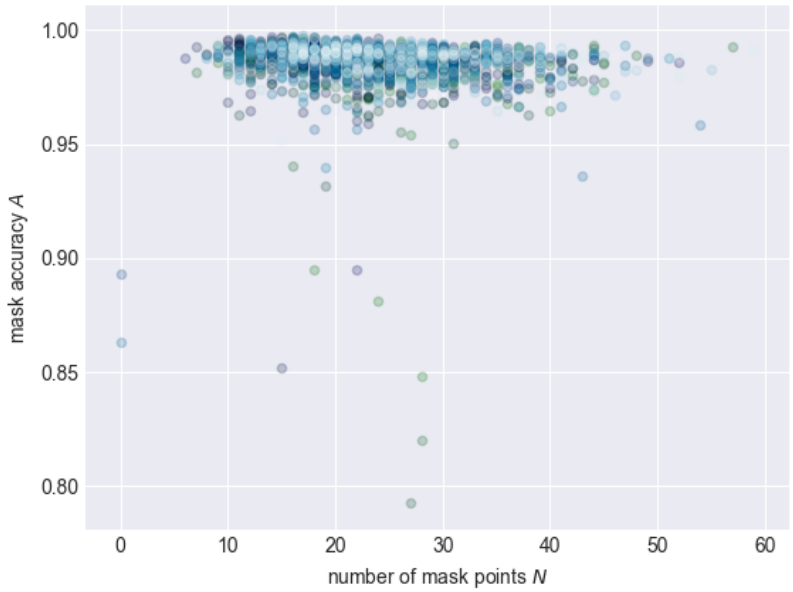

Figure 5. Accuracy w.r.t. the number of vertices that are used for delineation.

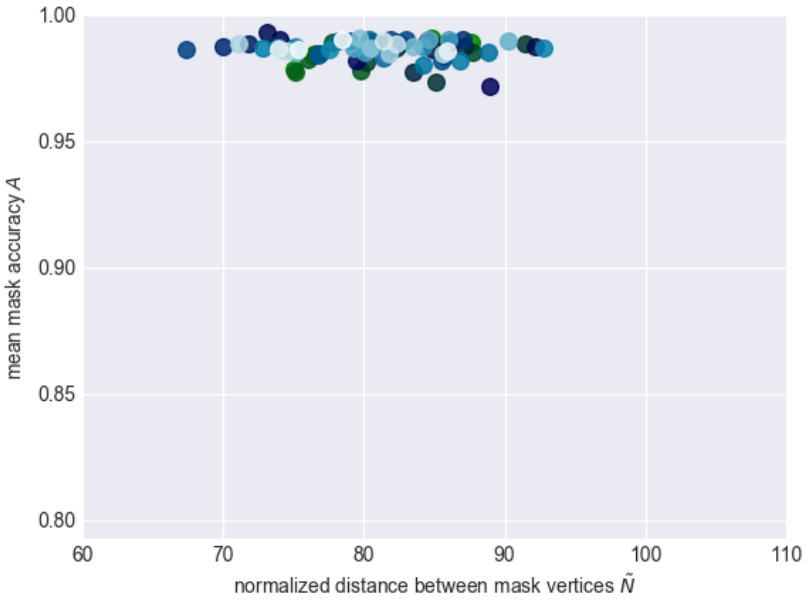

Figure 6. Mean mask accuracy w.r.t. the pixel distance between vertices.

be used e.g. to motivate other participants with gamification measures. Finally, the obtained STAPLE results will be evaluated with respect to their application as training samples for deep neuronal networks.

REFERENCES

[1] 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.

[2] Warfield SK, Zou KH, Wells WM. Validation of image segmentation by estimating rater bias and variance. Philos Trans A Math Phys Eng Sci 2008;366(1874):2361–2375.

[3] Gordon S, Lotenberg S, Long R, Antani S, Jeronimo J, Greenspan H. Evaluation of uterine cervix segmentations using ground truth from multiple experts. Comput Med Imag Grap 2009;33(3):205–216.

[4] Rohlfing T, Russakoff DB, Maurer CR. Extraction and application of expert priors to combine multiple segmentations of human brain tissue. In: Ellis RE, Peters TM, editors. Lect Notes Comput Sci. vol. 2879. Springer; 2003. p. 578–585.

[5] Dewalle-Vignion AS, Betrouni N, Baillet C, Vermandel M. Is STAPLE algorithm confident to assess segmentation methods in PET imaging? Phys Med Biol 2015;60(24):9473.

[6] Krenn M, Dorfer M, del Toro OAJ, Müller H, Menze B, Weber MA, et al. Creating a large-scale silver corpus from multiple algorithmic segmentations. In: Menze B, Langs G, Montillo A, Kelm M, Mller H, Zhang S, et al, editors. Lect Notes Comput Sci. vol. 9601. Springer; 2015. p. 103–115.

[7] Kashif M, Jonas SM, Deserno TM. Deterioration of R-Wave detection in pathology and noise: a comprehensive analysis using simultaneous truth and performance level estimation. IEEE Trans Biomed Eng 2017;64: 2163–2175.

[8] Goel S, Sharma Y, Jauer ML, Deserno TM. WeLineation: crowdsourcing delineations for reliable ground truth estimation. In: Deserno TM, Chen PH, editors. Proc SPIE. vol. 11318; 2020. (in press).