SAGE: An Approach and Implementation Empowering Quick and Reliable Quantitative Analysis of Segmentation Quality

Danna Gurari[1], Suele Ki Kim[1], Eugene Yang[1], Brett Isenberg[2], Tuan A. Pham[2], Alberto Purwada[2], Patricia Solski[2], Matthew Walker[3], Joyce Y. Wong[2], and Margrit Betke[1]

[1] Department of Computer Science, Boston University

[2] Department of Biomedical Engineering, Boston University

[3] Department of Biology, Boston University

dgurari@bu.edu

Abstract

Finding the outline of an object in an image is a fundamental step in many vision-based applications. It is important to demonstrate that the segmentation found accurately represents the contour of the object in the image. The discrepancy measure model for segmentation analysis focuses on selecting an appropriate discrepancy measure to compute a score that indicates how similar a query segmentation is to a gold standard segmentation. Observing that the score depends on the gold standard segmentation, we propose a framework that expands this approach by introducing the consideration of how to establish the gold standard segmentation. The framework shows how to obtain project-specific performance indicators in a principled way that links annotation tools, fusion methods, and evaluation algorithms into a unified model we call SAGE. We also describe a freely available implementation of SAGE that enables quick segmentation validation against either a single annotation or a fused annotation. Finally, three studies are presented to highlight the impact of annotation tools, annotators, and fusion methods on establishing trusted gold standard segmentations for cell and artery images.

1. Introduction

Computer vision applications often rely on a method to accurately segment the contours of objects of interest in the scene. It is critical in both academia and industry to demonstrate that the segmentation algorithm consistently provides the desired outcome.

Performance analysis of segmentation algorithms varies depending on the application objectives. Zhang [14] proposed to group evaluation methods into three categories: "analytical methods", "empirical methods based on goodness measures", and "empirical methods based on discrepancy measures". He concluded that methods based on discrepancy measures, which indicate how similar a query segmentation is to a gold standard segmentation (e.g., shape similarity), are most powerful for segmentation evaluation. We refer to this evaluation method as the "discrepancy measure model". In this work, we focus on performance analysis of segmentations using discrepancy measures.

There has been little discussion about when to use which segmentation analysis method when calculating discrepancy scores. Numerous papers review *evaluation methods* for finding a discrepancy between two segmentations [5, 12, 14]. An active area of research lies in establishing an *annotation collection* process to obtain gold standard segmentations including studies about annotation tools and annotator expertise level [1, 4, 6, 7, 8, 9]. More recently, *annotation fusion* methods are being developed to produce a reliable gold standard segmentation from a collection of annotations for the cases when intra-annotator and inter-annotator variation may be high [2, 3, 11, 13].

Finding the appropriate methodology for analyzing segmentations is important for the following reasons: 1) gaining an appreciation of a segmentation algorithm design and 2) providing a reliable foundation that can support down-stream analyses based on these segmentations. For example, developers may prematurely dismiss good algorithms when their measures indicate poor results, whether due to unreliable gold standard segmentations or the wrong discrepancy measure. Additionally, scientists may reject downstream analyses, even when measures indicate strong segmentations, if the gold standard segmentations are not trusted.

It may be insufficient to approach segmentation analysis by only identifying the appropriate discrepancy measure to establish a score [14]. This is because the chosen gold standard segmentation also impacts the score [2]. Furthermore, access to various segmentation analysis tools and methods is critical for establishing accepted segmentations.

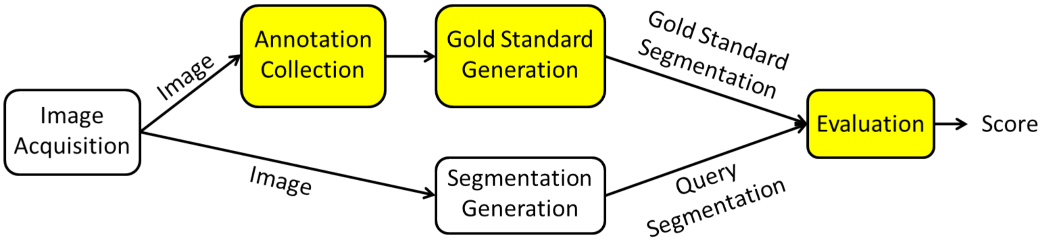

Figure 1. Overview of SAGE (**S**egmentation **A**nnotation Collection, **G**old Standard Generation, and **E**valuation), shown in yellow boxes, within the context of analyzing a query segmentation.

Yet shared toolboxes integrating these have not been developed, leading to non-novel, time-consuming efforts to build such systems. Lastly, given that finding a meaningful performance score depends on establishing a trusted gold standard segmentation, it is unclear how, in practice, to establish a trusted gold standard segmentation.

The key contributions of this paper are:

- A principled approach for analyzing segmentation performance that connects annotation collection approaches, fusion methods, and evaluation algorithms into a unified framework we call SAGE.
- A freely available system implementing SAGE that is compatible on many platforms and operating systems and links existing annotation tools with popular fusion algorithms and evaluation algorithms enabling quick segmentation validation against either a single annotation or a fused annotation [10].
- Three studies using the toolbox that highlight the impact of annotation tools, annotator expertise, and fusion methods on establishing trusted, i.e., high-consensus, gold standard segmentations and so meaningful evaluation scores.

In Section 2, we describe SAGE and a toolbox that implements SAGE. In Section 3, we describe three studies that highlight ways to establish a trusted gold standard segmentation for cell and artery images. In Section 4 we present the results and in section 5 we analyze the results and discuss future work. Conclusions are given in Section 6.

2. Methods

We propose in this section a principled approach to analyze the quality of segmentations. We formulate it as a model called **S**egmentation **A**nnotation Collection, **G**old Standard Generation, and **E**valuation (SAGE). We then describe a freely available system implementing this framework.

2.1. SAGE Framework

SAGE indicates a pipeline of steps to consider when establishing a process to analyze segmentation performance. A flowchart summarizing this model is shown in Figure 1. SAGE connects methods for collecting segmentation annotations with algorithms for generating a gold standard and measures for evaluating how similar a segmentation is to the gold standard. It expands upon the discrepancy measure model which considers only selecting the appropriate evaluation measure to establish a score.

Since one would use SAGE in the context of analyzing the quality of a segmentation, one first must obtain an image and generate a query segmentation of an object in that image to analyze (lower path in Figure 1). This segmentation may be created either automatically or manually. One then would apply the SAGE model to analyze the quality of that segmentation (upper path in Figure 1). To use SAGE, one must first collect annotations, which may be obtained by one or more annotators. Next, one must establish a gold standard segmentation, which can be an original annotation or a fused annotation created by combining multiple annotations. Lastly, one must calculate a score using a discrepancy measure to assess how similar the query segmentation is to the gold standard.

2.2. Implementation

We describe here a freely available implementation of SAGE that links popular segmentation analysis tools in a single system. It is developed in Java in order to easily run on various computer hardware with various operating systems [10]. The system has been validated on Windows 7, Windows XP, and Mac OS X operating systems. The configurable choices for the system are described in detail below.

Annotation Collection: The system supports reading segmentations from the following annotation tools: LabelMe [9], ImageJ [8], and Amira [1]. More generally, the system supports reading segmentations in binary image format, as xml files indicating object boundary points connected by straight lines, and as xml files indicating all object points.

Gold Standard Generation: When more than one annotation per image is provided, the user can select an original annotation or a fused annotation to represent the gold standard. The system supports two fusion methods: Thresholded Probability Maps [7] and Simultaneous Truth and Performance Level Estimation (STAPLE) [13].

Thresholded Probability Maps is an algorithm that takes N input segmentations and M segmentations and then labels a pixel as foreground when $\frac{M}{N} \geq p$ and background otherwise. STAPLE is an expectation-maximization algorithm that simultaneously generates gold standard segmentations and infers the performance of each input segmentation. For the formulation, each pixel is assigned 1 or 0 to indicate foreground and background respectively, T_i represents the value for the i-th pixel of the gold standard segmentation, D_{ij} represents the value for the i-th pixel of the j-th input segmentation, p_j represents the fraction of foreground pixels in the gold standard segmentation labeled as foreground in the segmentation for the j-th input segmentation, q_j represents the fraction of background pixels in the gold standard segmentation classified as background in the segmentation for the j-th input segmentation, and $j : D_{ij} = k$ denotes the set of indexes for which segmentation j has value k at pixel i. When the performance parameters p_j and q_j are given, pixels are labeled as foreground when W_i is greater than 0.5 and as background otherwise:

$$W_i \equiv f(T_i = 1|D_i, p, q) = \frac{a_i}{a_i + b_i} \quad (1)$$

$$a_i = f(T_i = 1) \prod_{j:D_{ij}=1} p_j \prod_{j:D_{ij}=0} (1 - p_j) \quad (2)$$

$$b_i = f(T_i = 0) \prod_{j:D_{ij}=0} q_j \prod_{j:D_{ij}=1} (1 - q_j) \quad (3)$$

The EM algorithm uses equation 4 to calculate the expected conditional log likelihood in the E-step and equations 5-6 to update the performance parameters for the M-step.

$$\begin{aligned} Q(\theta^t|\theta^{t-1}) = \sum_j [\sum_{i:D_{ij}=1} W_i^{(t-1)} \ln p_j + \\ \sum_{i:D_{ij}=1} (1 - W_i^{(t-1)}) \ln(1 - q_j) + \\ \sum_{i:D_{ij}=0} W_i^{(t-1)} \ln(1 - p_j) + \sum_{i:D_{ij}=0} (1 - W_i^{(t-1)}) \ln q_j] \quad (4) \end{aligned}$$

$$p_j^{(k)} = \frac{\sum_{j:D_{ij}=1} W_i^{(k-1)}}{\sum_i W_i^{(k-1)}} \quad (5)$$

$$q_j^{(k)} = \frac{\sum_{j:D_{ij}=0} (1 - W_i^{(k-1)})}{\sum_i (1 - W_i^{(k-1)})} \quad (6)$$

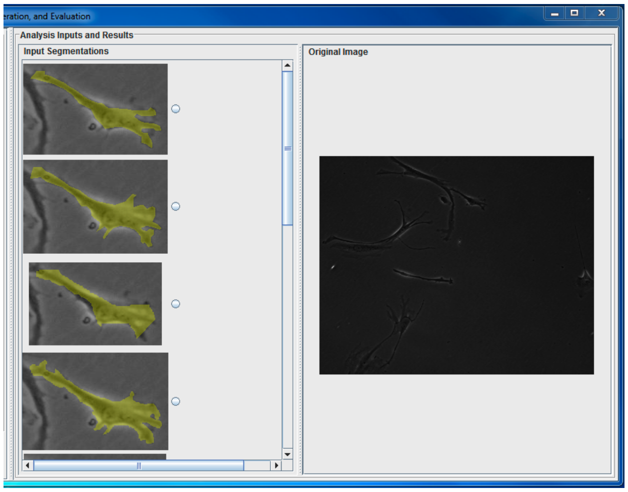

Figure 2. Interface of the toolbox for selecting a gold standard from annotations and fused annotation options.

When the system uses STAPLE, three starting conditions must be specified: initial performance parameters for input segmentations, probability a pixel in the image is foreground, and convergence threshold. The interface for selecting a gold standard from the original annotations and fusion segmentations is shown in Figure 2.

Evaluation Measures: The system supports the following six discrepancy measures commonly used to indicate how similar a query segmentation is to a gold standard segmentation - accuracy, precision, false positive rate, false negative rate, probability of error, and Hausdorff distance [5, 12, 14]. For the formulation of these measures, A represents the gold standard segmentation and B the query segmentation.

The system uses **accuracy** to calculate the fraction of the true cell region captured by the segmented region as $\frac{|A \cap B|}{|A|}$; **precision** to calculate the average overlap between the two regions as $\frac{|A \cap B|}{|A \cup B|}$; **false positive rate** to calculate the fraction of background pixels in the true segmentation labeled as foreground in the segmentation; **false negative rate** to calculate the fraction of foreground pixels in the true segmentation that are classified as background in the segmentation; **probability of error** to calculate the probability of mislabeling an object pixel as background or a background pixel as object as $PE = P(O) * P(B|O) + P(B) * P(O|B)$ where $P(B|O)$ is the false negative rate, $P(O|B)$ is the false positive rate, and $P(O)$ and $P(B)$ are the prior probabilities of object and background pixels respectively in the images; and **directed Hausdorff distance** to find the point in A furthest from any point in B and calculate the Euclidean distance from that point to its nearest point in B as $h(A, B) = \max_{a \in A}\{\min_{b \in B}\{d(a, b)\}\}$ where $d(a, b)$ is the Euclidean distance between points a and b.

Table 1. Description of image library for annotation

ID	# of Images	Imaging Modality	Object	Resolution	Avg. Object Pixel Count	Format
1	35	Phase Contrast	Neonatal rat smooth muscle cells	1024×811	35,649	tif
2	48	Phase Contrast	Fibroblast cells of the Balb/c 3T3 mouse strain	1030×1300	3,914	tif
3	36	Phase Contrast	Vascular smooth muscle cells from rabbit aortas	1030×1300	9,880	jpg
4	35	MRI	Left renal artery and the iliac bifurcation of a New Zealand White Rabbit	512×512	180	bmp

Table 2. Description of annotator experience

ID	Education Level	Worked with cell images	Worked with MRI images	Used ImageJ	Used Amira
A	Undergrad	3 mths	None	Yes	No
B	Post-doc	14 yrs	3 mths	Yes	No
C	PhD student	10 yrs	1 yr	Yes	No
D	Post-doc	2 mths	None	Yes	No
E	PhD student	3 wks	1 yr	Yes	No

3. Experiments

We ran three case studies using the toolbox to highlight various ways to establish trusted gold standard segmentations in practice. These studies examine which annotation tools to use, who should annotate, and whether fusion methods should be used. The measure used to evaluate whether a gold standard segmentation should be trusted is consensus amongst domain experts. We first characterize the image libraries and annotators and then describe the experimental design for each study.

3.1. Image Library for Annotation and Annotators

The intent of creating the image library was to provide a generalized collection of images representing various image acquisition modalities, object types, and image acquisition parameters. The image library contains a total of 154 images coming from four datasets. The first three datasets were collected by observing the cells with a Zeiss Axiovert S100 microscope and capturing images using a Princeton Instruments 1300YHS camera. For the first dataset, the cells were cultured at 37°C in 5% CO_2 on a PAAM hydrogel with embedded fluorescent beads with a size of 0.75 microns. For the second dataset, the cells were cultured at 37°C in 5% CO_2 on a PAAM hydrogel. For the third dataset, the cells were cultured at 37°C in 5% CO_2 on tissue culture plastic. The fourth dataset contains MRI images of a left renal artery obtained axially using a 3T MRI scanner (Philips Achieva). A single object from each dataset, present throughout the sequence of images, was identified to annotate. The specifications of the datasets are summarized in Table 1.

Five domain experts participated as annotators in the experiments. They had different education levels, experiences with the image types, and experiences with annotation tools, as summarized in Table 2.

3.2. Studies

Study 1: Impact of Annotation Tool. The five annotators were asked to annotate the first 154 images with two annotation tools, ImageJ [8] and Amira [1], using their own judgement. ImageJ, like LabelMe [9], uses a collection of user specified points connected by straight lines to produce a 2D segmentation. Amira collects user brush strokes to produce a 2D binary mask indicating all pixels in an object.

Annotator A annotated using a touchpad to interface with a laptop running a Mac operating system and would annotate in 2-3 hour intervals before taking a break. Annotator B annotated using a mouse to interface with both a desktop and laptop running typically on a Linux operating system and would annotate in 1-2 hour intervals before taking a break. Annotator C annotated using a touchpad to interface with a laptop running a Windows 7 operating system and would annotate in 1 hour intervals before taking a break. Annotator D annotated primarily using a mouse to interface with a laptop running a Windows 7 operating system and would annotate in 2 hour intervals before taking a break. Annotator E annotated using a mouse to interface with a desktop running a Windows 7 operating system and would annotate in 3-6 hour intervals before taking a break.

All annotators first annotated using ImageJ on all images in various orders. Then, within one week, all annotators annotated using Amira on all images in various orders.

The SAGE implementation was then run over all ImageJ annotations with each person having their annotations treated as a gold standard. For each of the five gold standard sets, the system was used to calculate the following six evaluation measures indicating how each of the other non-gold standard annotations compared against the gold standard: accuracy, precision, false positive rate, false negative rate, probability of error, and Hausdorff distance. This process was repeated for the Amira annotations.

Study 2: Impact of Annotators. Study one data is used to compare annotators qualitatively and quantitatively.

Study 3: Impact of Gold Standard Generation. Four experts participated in this study. First, a library of annotations was created to include ten annotation options for each of 98 images in the image library. Five of the annotation options were the ImageJ annotations produced by the five annotators. The other five annotation options were generated using fusion methods implemented in SAGE on the five input annotations. The five fusion methods are consecutively as follows: Thresholded Probability Map with $p = 0.2$ (union of annotations); Thresholded Probability Map with $p = 1$ (intersection of annotations); Thresholded Probability Map with $p = 0.6$ (majority vote); STAPLE initialized with global foreground set to 0.1, convergence threshold set to 0, and all performance parameters initialized to 0.7; STAPLE initialized with global foreground set to 0.1, convergence threshold set to 0, and performance parameters initialized to the average of performance parameter values assigned by the four experts participating in the study.

Then, the four experts used the SAGE implementation to select, from the ten annotations shown simultaneously, the segmentation best representing the gold standard. All experts were presented the original images in the same order and reviewed the 98 images in one sitting. For each image, the order of the corresponding annotations in the user interface was randomized to prevent the experts from learning which annotation represented what source.

4. Results

Study 1: Impact of Annotation Tool. Qualitative results of a set of annotations for an image from each dataset are shown in Figure 3 where relative size of objects are preserved. The quantitative results were pre-processed to include only data where the five annotators annotated the same object resulting in 153 valid ImageJ images and 152 valid Amira images. For each annotation tool, the average score for each evaluation measure over all annotator comparisons across all images was calculated. Quantitative results are shown in Table 3.

Study 2: Impact of Annotators. For the post-processed data, the average evaluation score over all images for every permutation of two annotators for each evaluation measure was calculated using SAGE. Quantitative results for Amira and ImageJ annotations are shown in Table 4.

Table 3. Average evaluation measure score for annotations obtained using different annotation tools are shown where *I-* indicates ImageJ annotations and *M-* indicates Amira annotations and D_i indicates the i-th dataset.

Tool	Acc	Prec	FPR	FNR	POE	HD
I-All	0.85	0.72	0.0018	0.15	0.0035	16
M-All	0.87	0.76	0.0018	0.13	0.0034	14
I-D_1	0.86	0.74	0.006	0.14	0.011	29
M-D_1	0.87	0.77	0.0058	0.13	0.011	30
I-D_2	0.86	0.75	0.0004	0.14	0.0008	12
M-D_2	0.89	0.80	0.0003	0.11	0.0007	10
I-D_3	0.86	0.75	0.0010	0.14	0.002	18
M-D_3	0.87	0.77	0.0009	0.13	0.002	16
I-D_4	0.82	0.65	0.0002	0.18	0.0004	4
M-D_4	0.85	0.73	0.0001	0.15	0.0002	3

Study 3: Impact of Gold Standard Generation: From the 98 images, where experts voted for the best from 10 segmentations, we found agreement between none of the annotators for 27 images, two annotators for 49 images, three annotators for 18 images, and four annotators for 4 images. Where there was consensus, there were five cases of voting ties. From the 76 cases of voting consensus for a particular annotation, 26 were for B, 13 were for A, 13 were for the Probability Threshold Map fusion method with $p = 0.6$, 8 were for E, 7 were for D, 4 were for STAPLE with uniform performance parameters initialized, 3 were for STAPLE with performance parameters established by the experts, and 2 were for the Probability Threshold Map fusion method with $p = 1$. Annotator C and Probability Threshold Map fusion method with $p = 0.2$ did not receive any consensus votes. Fused methods accounted for 9.86% of the consensuses.

5. Discussion and Future Work

We first discuss the benefit of using the SAGE model. Then, we analyze the impact of the annotation tools, annotators, and fusion methods on establishing trusted gold standard segmentations in practice.

SAGE Model: Design Analysis. The results of our studies support the flow of modules used in our SAGE model. The annotation collection process should precede gold standard generation since varying the collection methods leads to differences in the gold standard as observed qualitatively in Figure 3 and quantitatively in Table 4. The gold standard generation step should precede the evaluation measure step because varying the gold standard generation process (e.g., using various fusion methods with var-

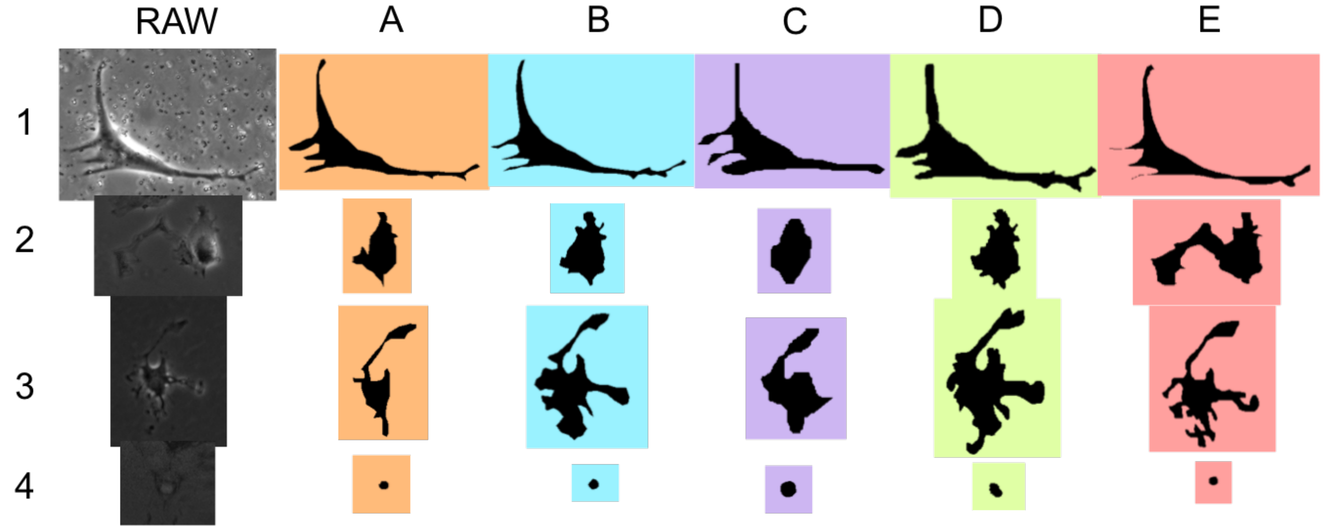

Figure 3. Qualitative results showing a set of annotations collected using ImageJ from the five annotators (A-E) for an image from each dataset (1-4).

Table 4. Average evaluation score over all images for every pair of annotations for each evaluation measure are shown where *I-* indicates ImageJ annotations and *M-* indicates Amira annotations. False positive rate and probability of error scores are all $value \times 10^{-2}$
.

	AB	AC	AD	AE	BA	BC	BD	BE	CA	CB	CD	CE	DA	DB	DC	DE	EA	EB	EC	ED
I-Acc	0.95	0.92	0.97	0.94	0.81	0.87	0.94	0.89	0.68	0.76	0.81	0.75	0.72	0.82	0.82	0.80	0.81	0.90	0.88	0.93
M-Acc	0.89	0.91	0.93	0.94	0.88	0.92	0.94	0.95	0.80	0.81	0.86	0.88	0.82	0.85	0.88	0.91	0.78	0.80	0.84	0.85
I-Prec	0.78	0.63	0.70	0.76	0.78	0.67	0.77	0.81	0.63	0.67	0.69	0.67	0.70	0.77	0.69	0.74	0.76	0.81	0.67	0.74
M-Prec	0.80	0.73	0.77	0.74	0.80	0.75	0.80	0.76	0.73	0.75	0.76	0.74	0.77	0.80	0.76	0.78	0.74	0.76	0.74	0.78
I-FPR	0.17	0.35	0.32	0.23	0.29	0.24	0.17	0.07	0.24	0.15	0.17	0.10	0.20	0.07	0.11	0.04	0.10	0.12	0.23	0.08
M-FPR	0.10	0.24	0.18	0.30	0.19	0.29	0.20	0.35	0.13	0.08	0.14	0.24	0.13	0.06	0.21	0.26	0.10	0.05	0.15	0.11
I-FNR	0.05	0.08	0.03	0.06	0.19	0.13	0.06	0.11	0.32	0.24	0.19	0.25	0.28	0.18	0.18	0.20	0.12	0.10	0.07	0.19
M-FNR	0.10	0.09	0.07	0.06	0.11	0.09	0.06	0.05	0.21	0.19	0.14	0.12	0.18	0.16	0.12	0.09	0.22	0.20	0.16	0.15
I-POE	0.23	0.43	0.35	0.30	0.23	0.42	0.30	0.27	0.43	0.42	0.42	0.42	0.35	0.30	0.42	0.33	0.30	0.27	0.42	0.33
M-POE	0.29	0.36	0.30	0.38	0.29	0.37	0.26	0.39	0.36	0.37	0.34	0.37	0.30	0.26	0.34	0.34	0.38	0.39	0.37	0.34
I-HD	13	19	14	13	15	20	11	11	17	16	14	15	18	14	21	14	16	13	22	13
M-HD	16	15	13	16	12	12	10	13	15	15	11	14	17	17	14	15	17	18	16	14

ious tuned parameters) while keeping the annotation collection process constant (same collection of input annotations) and evaluation measure constant causes the output score to vary [2]. Finally, the annotation collection process is independent from the gold standard generation step because varying the annotation collection process while keeping the evaluation measure constant and gold standard selection process constant (using a single input annotation as is), causes the output score to vary as shown in Table 4.

The results therefore suggest that SAGE more accurately describes the factors impacting the performance score than the the discrepancy measure model. Thus it shows that SAGE is a useful model to use when analyzing segmentation quality.

Study 1: Impact of Annotation Tool. Images in Figure 3 exemplify the variety of annotation challenges in the four datasets, where objects in dataset 4 are small, the background in dataset 1 contains clutter, and objects in datasets 2 and 3 have involved contour details.

Quantitatively, the annotator agreement when using Amira is on average greater than or equal to the annotator agreement when using ImageJ for all 6 measures over all four datasets. Note that higher values are better for the accuracy and precision measures, while lower values are better for the other four measures. In contrast to the findings in Meyer et al's work [7], which found that there was no significant difference between annotation methods, this suggests that inter-annotator variation can be reduced based on the annotation method used.

Future work will explore the cause of this improvement.

The annotators suggested that the improvement may be because Amira supports easily erasing and adding pixels to the segmentation whereas correction is a more involved process with ImageJ. Also, Amira identifies an annotation with a transparent overlay on the image while ImageJ only displays the segmented line or the filled region making comparison against the original image difficult.

Study 2: Impact of Annotators. Images in Figure 3 exemplify the differences between how annotators annotate images. Quantitatively, the set of measures reveal that education level and experience may not be the greatest factors for achieving annotator consensus. Annotators A and B agree more (columns AB and BA) than B and C (columns BC and CB), the most experienced annotators, with respect to Hausdorff distance, probability of error, and precision while the other measures indicate comparable similarity between annotators. Annotators A and B share similar agreement (columns AB and BA) to that between B and D (columns BD and DB), the most educated annotators, with respect to accuracy, precision, false negative rate, probability of error, and Hausdorff distance. One suggested cause of the high agreement between A and B was their shared training for what defines the gold standard, as they were the only pair from the five annotators that conducted research together. Future work will explore the impact of shared instructions for how to annotate on annotator consensus.

Study 3: Impact of Gold Standard Generation. Furthering the previous analyses of fusion methods [2, 13], we investigate whether the fusion methods are perceived to provide improved segmentations over the original annotations. Results indicate a low preference for fusion methods over original annotations for our datasets. Future work will investigate whether fusion methods are preferred for different applications.

6. Conclusions

Knowledge of the various segmentation analysis methodologies and access to segmentation analysis tools are critical for establishing trusted segmentations. We presented a framework to obtain project specific segmentation performance indicators in a principled way that links annotation collection processes with gold standard generation methods and evaluation algorithms. Furthermore, by turning this framework into a toolbox supporting popular tools and algorithms, we enable researchers to focus on the most important research issues of developing improved algorithms and establishing reliable gold standard segmentations. Three user studies run with the toolbox demonstrate the impact of annotation tools, annotator expertise, and fusion methods on establishing reliable gold standard segmentation. Analyses indicate a preference for the annotation tool Amira over ImageJ and for original annotations over fused annotations.

Acknowledgments

The authors thank Netta Gurari and Diane Theriault for their useful discussions and gratefully acknowledge funding from the National Science Foundation (IIS-0910908).

References

[1] Amira, software platform for visualizing, manipulating, and understanding life science and bio-medical data. Retrieved August 17, 2012, from http://amira.com. 1, 2, 4

[2] A. M. Biancardi, A. C. Jirapatnakul, and A. P. Reeves. A comparison of ground truth estimation methods. *International Journal of Computer Assisted Radiology and Surgery*, 5(3):295–305, 2010. 1, 6, 7

[3] S. R. Cholleti, S. A. Goldman, A. Blum, D. G. Politte, and S. Don. Veritas: Combining expert opinions without labeled data. *In Proceedings of the 2008 20th IEEE International Conference on Tools with Artificial intelligence*, 2008. 1

[4] S. Dasiopoulou, E. Giannakidou, G. Litos, P. Malasioti, and I. Kompatsiaris. A survey of semantic image and video annotation tools. *Knowledge-Driven Multimedia Information Extraction and Ontology Evolution*, 6050:196–239, 2011. 1

[5] D. P. Huttenlocher, G. A. Klanderman, and W. J. Rucklidge. Comparing images using the Hausdorff distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(9):850–863, 1993. 1, 3

[6] J. Little, A. Abrams, and R. Pless. Tools for richer crowd source image annotations. *In Workshop on Applications of Computer Vision (WACV)*, pages 369–374, 2012. 1

[7] C. R. Meyer, T. D. Johnson, and G. M. et al. Evaluation of lung MDCT nodule annotation across radiologists and methods. *Acad Radiology*, 13(10):1254–1265, 2006. 1, 3, 6

[8] W. Rasband. ImageJ, 1997–2012. U.S. National Institutes of Health, Bethesda, Maryland, USA, http://imagej.nih.gov/ij. 1, 2, 4

[9] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman. Labelme: a database and web-based tool for image annotation. *International Journal of Computer Vision*, 77(1–3):157–173, 2008. 1, 2, 4

[10] SAGE: An implementation empowering quick and reliable quantitative analysis for segmentation comparison. Retrieved January 10, 2013, from http://www.cs.bu.edu/~betke/SAGE. 2

[11] A. Sorokin and D. Forsyth. Utility data annotation with amazon mechanical turk. *Proceedings of First IEEE Workshop on Internet Vision at CVPR*, pages 1–8, 2008. 1

[12] J. K. Udupa, V. R. LeBlanc, Y. Zhuge, C. Imielinska, H. Schmidt, L. M. Currie, B. E. Hirsch, and J. Woodburn. A framework for evaluating image segmentation algorithms. *Comp Med Imaging and Graphics*, 30(2):75–87, 2006. 1, 3

[13] S. K. Warfield, K. H. Zou, and W. M. Wells. Simultaneous truth and performance level estimation (STAPLE): An algorithm for the validation of image segmentation. *IEEE Transactions on Med Imaging*, 23(7):903–921, 2004. 1, 3, 7

[14] Y. J. Zhang. A survey on evaluation methods for image segmentation. *Patt Recog*, 29(8):1335–1346, 1996. 1, 3