How to Collect Segmentations for Biomedical Images? A Benchmark Evaluating the Performance of Experts, Crowdsourced Non-Experts, and Algorithms

Danna Gurari, Diane Theriault, Mehrnoosh Sameki, Brett Isenberg, Tuan A. Pham, Alberto Purwada, Patricia Solski, Matthew Walker, Chentian Zhang, Joyce Y. Wong, Margrit Betke

Abstract

Analyses of biomedical images often rely on demarcating the boundaries of biological structures (segmentation). While numerous approaches are adopted to address the segmentation problem including collecting annotations from domain-experts and automated algorithms, the lack of comparative benchmarking makes it challenging to determine the current state-of-art, recognize limitations of existing approaches, and identify relevant future research directions. To provide practical guidance, we evaluated and compared the performance of trained experts, crowdsourced non-experts, and algorithms for annotating 305 objects coming from six datasets that include phase contrast, fluorescence, and magnetic resonance images. Compared to the gold standard established by expert consensus, we found the best annotators were experts, followed by non-experts, and then algorithms. This analysis revealed that online paid crowdsourced workers without domain-specific backgrounds are reliable annotators to use as part of the laboratory protocol for segmenting biomedical images. We also found that fusing the segmentations created by crowdsourced internet workers and algorithms yielded improved segmentation results over segmentations created by single crowdsourced or algorithm annotations respectively. We invite extensions of our work by sharing our data sets and associated segmentation annotations (`http://www.cs.bu.edu/~betke/BiomedicalImageSegmentation`*).*

1. Introduction

Imaging has become a common and important tool for advancing our understanding of biomedical processes, enabling observation both within and outside of living organisms (i.e., *in vivo* and *in vitro*) [1, 2]. In principle, collected images will contribute to the discovery of how the human body functions in both healthy and diseased states which will in turn greatly assist in the treatment and prevention of diseases and the engineering of biomaterials. However, the bottleneck limiting progress is often extracting information from the large number of captured images and typically depends on finding boundaries of biological structures (segmentations), whether analyzing morphology, classification, or motion. As a result, common questions asked by individuals analyzing biomedical images is "what segmentation collection approach is *sufficient* to consistently and efficiently find the desired boundaries of biological structures in their images?" and "given that derived biological interpretations are influenced by the accuracy of the boundaries of biological structures, what segmentation collection approach will yield the *most accurate* boundaries?"

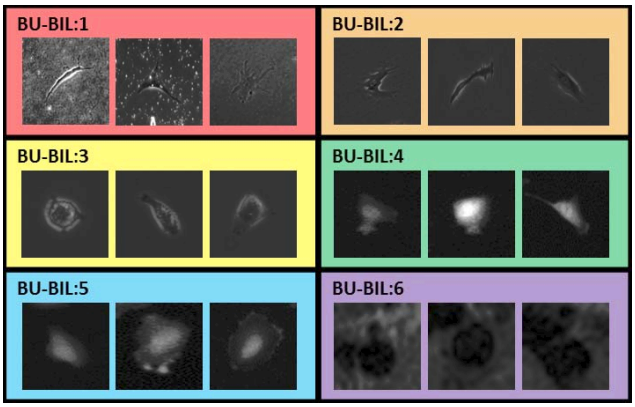

Figure 1. Representative images from the six datasets in the image library. Segmentation methods that accurately delineate boundaries of biological structures must handle appearance variations with respect to intensity, size, and shape; faint edges separating structures from the background; and backgrounds with clutter.

Often, *domain experts* draw the boundaries of biological structures using annotation software such as ImageJ [3] or Amira [4] [5]. The key motivating assumption for this approach is that human annotators trained on how to interpret images collected using particular biomedical image acquisition systems can distinguish between true object boundaries and image noise/artifacts and so draw highly accurate boundaries. However, this approach is time-consuming, expensive and does not scale.

To overcome the obstacles associated with relying on manual annotation by experts, developers have been in-

978-1-4799-6683-7/15 $31.00 © 2015 IEEE
DOI 10.1109/WACV.2015.160

Table 1. Salient properties characterizing each dataset in the image library and the number of objects per dataset.

ID	Modality	Object Type	Mag.	Avg. Pixel Count	Avg. Circularity	Avg. Object Intensity	Avg. Bkgrnd Intensity	# Objs
1	Phase Contrast	Rat smooth muscle cells	40x	35,613	0.15	64	61	35
2	Phase Contrast	Rabbit smooth muscle cells	10x	10,963	0.29	52	50	69
3	Phase Contrast	Fibroblasts	10x	3,937	0.53	58	50	47
4	Fluorescence	Lu melanoma cells	10x	836	0.53	48	17	61
5	Fluorescence	WM993 melanoma cells	10x	1,119	0.71	54	19	58
6	MRI	Rabbit aorta	10x	216	0.94	25	42	35

tegrating *segmentation algorithms* into publicly available image analysis systems and researchers have been designing new algorithms to tackle open segmentation challenges [3, 6, 7]. Older methods including thresholding (e.g., Otsu Thresholding [8]), feature-based (e..g, Hough Transform [9]), and region growing (e.g., Seeded Watershed [10]) algorithms are still actively used, in part because of their ease of use and widespread availability in bioimage analysis systems. Level-set based algorithms are more recent developments; their success typically depend on selecting an appropriate initial contour which gets evolved into the final boundary [6, 16]. Although the continued development and wide-spread sharing of new segmentation tools are valuable for assisting with the effort required to analyze the large number of images, the number of automation methods are becoming too numerous to explore for both non-experts and experts. A challenge for individuals trying to choose from the abundance of options is how to infer from isolated studies reported for lab-specific datasets which tool will work well for their biomedical image sets since there are no comparison studies that include algorithms from the past 15 years and analyze algorithms on more than a couple of datasets [14, 15].

An alternative option is to leverage recently available, easy-to-use *crowdsourcing* systems that make it plausible for manual annotations to be a scalable solution to the segmentation problem [17]. This begs the question of whether large groups of non-trained humans can be leveraged to consistently draw accurate boundaries for biomedical image sets.

The purposes of this work are to facilitate making an informed choice quickly about which segmentation collection approach will work well for biomedical image sets and to highlight limitations of existing methods. The key contributions of this work are:

- Evaluating and comparing the performance of biomedical image segmentation by trained experts, non-experts and automated segmentation algorithms
- Demonstrating a reliable process for using online, paid crowdsourced workers as part of the laboratory protocol for segmenting biomedical images
- Publicly sharing a library of images collected and used for biomedical research with associated expert annotations

2. Biomedical Image Library (BU-BIL)

We compiled a generalized image library using images recorded for biology and biomedical research studies at Boston University for which high-quality image segmentations were required (**Table 1**). Our image library includes six datasets that represent three imaging modalities and six object types (**Table 1**). We instructed the providers of the datasets to choose images that capture the various environmental conditions and imaging noise that arose in their studies. We asked these experts to then select objects from those images that reflect the natural diversity of shape and appearances that these objects can exhibit. We finally cropped the image subregions containing the identified objects to create our image library (discussed below). The outcome was a library with 305 objects from 235 images. We verified by visual inspection that the image library includes a variety of object appearances, backgrounds, and properties distinguishing objects from the background (**Fig. 1**). We call this collection the Boston University Biomedical Image Library (BU-BIL) and share it publicly and describe these datasets below (`http://www.cs.bu.edu/~betke/BiomedicalImageSegmentation`).

Phase Contrast Images of Cells (datasets 1–3): Images were collected by observing the cells with a Zeiss Axiovert S100 microscope, a Ludl motorized stage, and a cooled Princeton Instruments CCD camera. In each experiment, a density of 10^3 cells/cm^2 was selected to reduce cell-cell interactions. For *dataset 1*, the neonatal rat smooth muscle cells (Coriell Cell Repositories, NJ) were cultured on PAAM hydrogel that contained embedded 0.75-μm fluorescent beads to facilitate imaging of gel deformation, and incubated at 37°C in 5% CO_2 for a minimum of 18 hours. Image dimensions were 1,024 by 811 pixels and pixels were recorded using eight bits. For *datasets 2–3*, the vascular muscle cells from New Zealand White and Watanabe Heritable Hyperlipidemic (WHHL) rabbit aortas (Brown Family Research) and fibroblasts of the Balb/c 3T3 mouse strain (American Type Culture Collection, VA)

were cultured at 37°C in 5% CO_2 in Dulbecco's modified Eagle's medium (Invitrogen, NY) supplemented with penicillin, streptomycin, L-glutamine, and 10% bovine calf serum (Hyclone, UT). Six hours before image acquisition, the cells were seeded onto a tissue culture plastic substrate. Image dimensions were 1,300 by 1,030 pixels for both datasets. Dataset 2 was recorded using one byte per pixel and dataset 3 was recorded using 2 bytes per pixel.

Fluorescence Images of Cells (datasets 4–5): Images were collected by observing the cells with a Zeiss Axiovert S100 microscope, a Ludl motorized stage, and a cooled Princeton Instruments CCD camera (1,300 x 1,030 pixels, 1-byte/pixel). The 1205 Lu and WM993 melanoma cells (Wistar Institute) were each cultured at 37°C in 5% CO_2 in Dulbecco's modified Eagle's medium supplemented with penicillin, streptomycin, L-glutamine, and 10% bovine calf serum (Invitrogen, NY). Cells were patterned onto a dish using a microfabricated polydimethylsiloxane (PDMS) stencil with a 600 micron hole. After 6 hours incubation at 37°C in 5% CO_2, the stencil was peeled away and media was added to the dish. The patterned cells were placed in a custom constructed microscope incubator to maintain stable culture conditions.

Magnetic Resonance Images of Aortas (dataset 6): Magnetic resonance images (MRIs) were collected axially of the aorta of two New Zealand White Rabbits. A 3T Philips Achieva MRI scanner was used to collect each series of images of physical locations along the aorta at cross-cuts 4mm apart showing the volume of the aorta that extends from the left renal bifurcation to the iliac bifurcation (512 x 512 pixels, 1-byte/pixel). The iliac and left renal bifurcation are both roughly perpendicular to the aorta. The aorta runs approximately perpendicular to the axial scan direction. Each pixel represents approximately 0.23 x 0.23 mm. The dataset includes a complete MRI scan with 22 images and a partial MRI scan with 13 images

Image Cropping: We cropped all images so that there is exactly one dominant object in the foreground. To do this, an expert-drawn segmentation is used to detect the object location, and increase the bounding box size by a percentage of the original bounding box dimensions, which maintains the original region proportions. For datasets 1-5, we used 50% and for dataset 6 we used 125%. The datasets represent biological structures that range in size from approximately hundreds to tens of thousands of pixels (**Table 1**).

3. Methods

We collected multiple annotations for each of the 305 objects in our image library using trained domain experts; online, paid crowdsourced workers; and algorithms. Expert annotations are freely shared (`http://www.cs.bu.edu/~betke/BiomedicalImageSegmentation`).

3.1. Expert-Drawn Annotations

A total of ten trained domain experts participated as annotators. Some of the annotators were also the creators of the image data. They had a vested interest in the quality of the segmentations they produced because they needed accurate object boundaries for their biomedical research studies.

The annotators created segmentations using three computer annotation tools: ImageJ [3], Amira [4], and an iPad touchpad drawing program [18]. ImageJ takes as input user specified points and connects them sequentially with straight lines to produce a 2D segmentation. Both Amira and the touchpad drawing program take as input user brush strokes to produce a 2D binary mask indicating all pixels in an object. All domain experts had experience with biomedical images and ImageJ. We instructed the annotators to identify the object regions using their own judgment.

3.2. Crowdsourced-Drawn Annotations

We collected seven crowdsourced segmentation annotations for each of the 305 objects.

The annotators created segmentations using the on-line image annotation tool LabelMe [19]. LabelMe supports tracing the boundary of objects by taking as input user specified points and connecting them sequentially with straight lines. The annotator finishes annotating an object by clicking on the starting point or right clicking with the computer mouse. If a mistake is made, the annotator can delete and redraw the object boundary.

We recruited annotators from the Amazon Mechanical Turk internet marketplace. We posted each drawing task for each image to Mechanical Turk as a HIT paired with a price to be paid upon completion of the task. An internet worker could review the HIT before accepting the job. Workers were first shown step-by-step annotation instructions followed by pictures exemplifying good and bad annotations (**Fig. 2a**). After accepting the HIT, a worker was then presented the drawing interface to create the object boundary (**Fig. 2b**). A worker could submit a HIT after meeting either of the two criteria for finishing the annotation. We paid workers $0.02 for each submitted HIT and accepted all submitted HITs. We only accepted workers that had previously completed at least 100 human intelligence tasks (HITs) and received at least a 92% approval rating.

3.3. Computer-Drawn Annotations

We evaluated six publicly available algorithms that represent four key classes of algorithms commonly reported in the literature for biomedical images [20]: thresholding (i.e., *Otsu thresholding* [8]), feature-based (i.e., *Hough transform for circles* [9]), region-growing (i.e., *seeded watershed* [10]), and deformable models (i.e., *Chan Vese level set method* [11], *Lankton region-based level set method* [13], and *Shi approximation level set method* [12]).

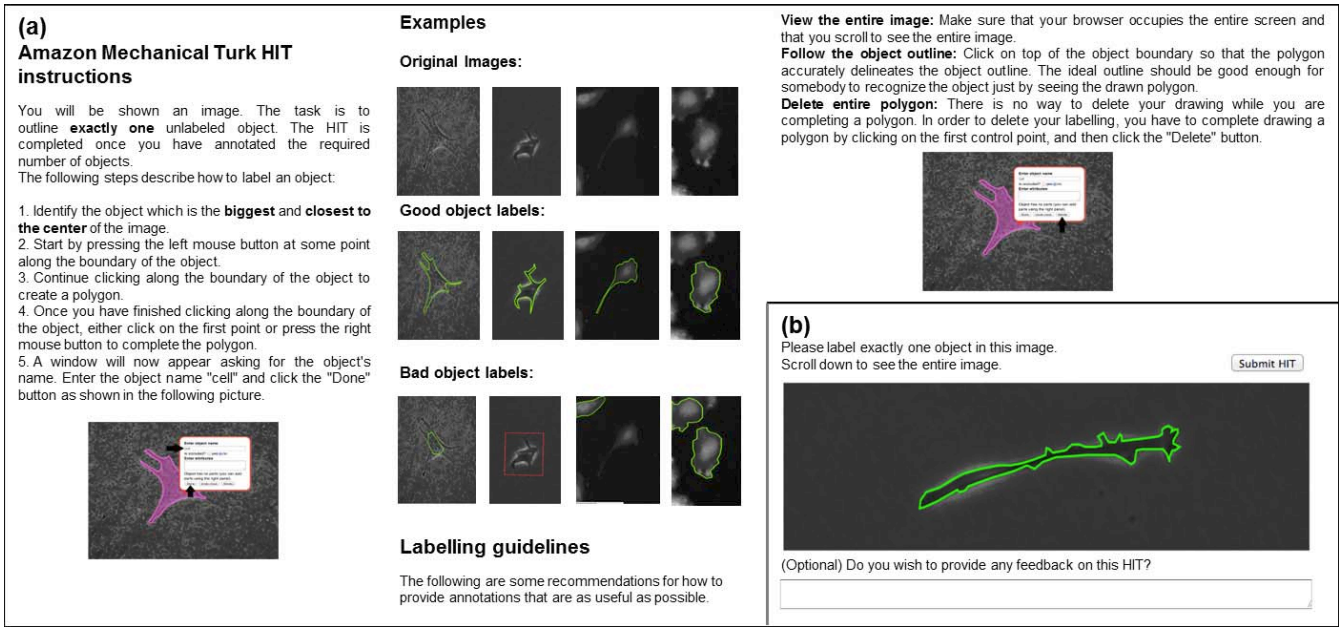

Figure 2. Crowdsourcing user interface. An example of (**a**) the annotation instructions given for datasets 1-5 and (**b**) a segmentation annotation created using the interface that internet workers use to complete the drawing task, LabelMe.

Otsu thresholding (Otsu) is based on the assumption that biological structures ("foreground") have different intensity values than the background [8]. It finds the value that minimizes the average variance between all foreground and background pixels respectively and then assigns all pixels with intensities below that value as background and the rest of the pixels as foreground.

Hough transform with circles (HoTr) finds the set of circles that have at least a pre-specified number of pixels on their boundary in the edge map of the image [9]. We combine these circles to create the final segmentation.

Seeded watershed (SeWa) is based on the assumption that the biological structure and background can be separated based on intensity homogeneity and spatial proximity [10]. The algorithm starts from initial markers and then iteratively adds unassigned neighboring pixels to one of the markers until every pixel is assigned to the region of exactly one marker. The algorithm runs on the gradient map of the image. We automatically set two initial markers: we used the convex hull of the Hough Transform for Circles segmentation for the background marker and the eroded Hough Transform for Circles segmentation for the foreground marker.

The three *level set* based methods deform an initial contour to a final contour, separating image foreground from background so that some method-specific image partition condition is enforced. *Chan Vese level set method (ChVe)* evolves the initial contour to try to separate the image into two homogeneous intensity regions [11]. The *Shi approximation level set method (Shi)* computationally speeds up the evolution process by replacing slow real-valued calculations with faster integer-based calculations [12]. *Lankton region-based level set method (Lank)* evolves the initial contour by using the local neighborhood statistics for each pixel in order to adjust how to separate the sub-region into two homogeneous intensity regions [13]. For all three methods, we automatically created initial contours using the boundary of a circle drawn at the center of the image region with a diameter slightly smaller than the smallest image dimension. For all three methods, we set a maximum number of 2000 iterations before algorithm termination.

We built a system that facilitates applying all the segmentation algorithms on all images in the library with one command. The system processes all images sequentially. For each image, the workflow is to apply a segmentation algorithm, post-process by filling any holes and keeping the largest object, and finally save the resulting binary segmentation as an image. We wrapped publicly available code for each of the six segmentation algorithms into six modules that adapt the the original code interface into a shared, compatible interface in the system (**Table 2**).

3.4. Fused Annotations

We evaluate segmentations created by an ensemble algorithm to examine how combining multiple segmentations compares with stand-alone segmentations. We used *Probability Maps (P-map)* which takes as input N segmentations and outputs a single segmentation where a pixel is labeled as foreground when at least M of the segmentations label it as foreground and background otherwise. We chose this method because it is simple to understand and does not require tuning a set of complex algorithm parameters. We

Table 2. List of segmentation sources evaluated in the study and associated publicly available code and systems used.

Segmentation Source (Acronym)	Publicly Available System/Code
Expert Annotators (Expe)	Amira [4]; ImageJ [3]; iPad touchpad drawing program [18]
Non-Expert Annotators (NoEx)	LabelMe [19]
Otsu Thresholding [8] (Otsu)	MATLAB [7]; ImageJ plug-in [3]
Hough Transform for Circles [9] (HoTr)	MATLAB [7]; ImageJ plug-in [3]
Seeded Watershed [10] (SeWa)	MATLAB [7]; ImageJ plug-in [3]
Chan Vese level set method [11] (ChVe)	MATLAB [6]
Shi approximation level set method [12] (Shi)	MATLAB [6]
Lankton region-based level set method [13] (Lank)	MATLAB [6]

then post-process the segmentation result by filling holes and keeping only the largest object.

4. Experiments

To evaluate the segmentation sources, we analyzed a total of 6,148 segmentations created by 10 experts, 58 crowdsourced workers, and six algorithms. The studies were designed to examine 1) which source among experts, non-experts, and algorithms yields the most accurate segmentations?, 2) how well does each of the segmentation sources generalize to different biological structure characteristics and image modalities?, and 3) what are the limitations of each segmentation source?

4.1. Performance Evaluation Methodology

To evaluate segmentation quality, we computed scores that indicate how closely annotations match gold standard segmentations, i.e., representations of "true" biological structure regions, using the region overlap ratio, a standard evaluation metric. This metric computes the number of pixels common to both the annotation and gold-standard regions that are in the combination of regions (i.e., $\frac{|A \cap B|}{|A \cup B|}$, where A represents the set of pixels in the gold standard segmentation and B represents the set of pixels in the annotation). Scores range from 0 to 1 with higher scores reflecting greater similarity and so better performance.

To establish high-quality gold standard segmentations, we used the consensus between expert-drawn segmentations. For each image, we applied the fused annotation method (**Section 3.4**), using as input all available expert annotations and setting M to the minimum value that returns a majority vote.

4.2. Analysis of Segmentation Sources

We computed the overlap ratio for every segmentation produced by all experts, non-experts, and algorithms. These scores are the foundation for our subsequent analyses.

We first independently analyzed for each of the three segmentation sources all scores over the entire image library, the subset of phase contrast images (datasets 1-3), the subset of fluorescence images (datasets 4-5), and the subset of magnetic resonance images (dataset 6).

We next analyzed the variability within each of the three segmentation sources for each dataset. For experts, we evaluated based on each annotation set, which is defined as a particular annotator using a single annotation tool. For non-experts, we evaluated based on each batch from the seven batches of crowdsourced annotations we collected per image. For algorithms, we evaluated based on each set of algorithm drawn segmentation results generated.

Finally, we analyzed whether combining segmentations could lead to improved results for the non-expert and algorithmic sources. We applied the fused annotation method (**Section 3.4**) independently to the set of non-expert and algorithm annotations, and chose $M = 4$ because its the minimum value that returns a majority vote. We then computed the overlap ratio for all resulting segmentations.

4.3. Image Library Characterization

We characterized the diversity of biological structures and environmental conditions in the image library to support analyses that suggest which algorithms cater to particular image conditions versus generalize well. Gold standard segmentations were used to compute the area, circularity, i.e., degree of deviation from a circle, and average intensity of the biological structure as well as average background intensity for each image region.

5. Results

5.1. Analysis of Segmentation Sources

We found overall that the experts consistently drew more accurate segmentations than non-experts who consistently drew more accurate segmentations than algorithms, when evaluating by comparing the median score of all analyzed segmentations against the gold standard segmentations (**Fig. 3**; All). The median score over the entire image library is 0.85 for experts, 0.82 for non-experts, and 0.36 for algorithms. With respect to how annotation quality relates to imaging modality, we found that all three segmentation sources consistently drew segmentations best matching gold

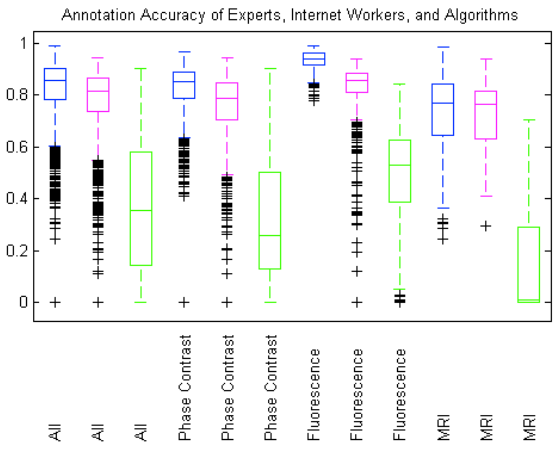

Figure 3. Region overlap ratio scores for segmentations created by experts (red), non-experts (green), and algorithms (blue), averaged over all data, and data of each of the three image modalities. For each annotation source, the central mark of the box denotes the median score and the box edges the 25th and 75th percentiles scores. The whiskers extend to the most extreme data points not considered outliers, and the outliers are plotted individually (black). Surprisingly, the quality of annotations of internet workers follows closely that of experts, and algorithms perform on average much worse. Automated segmentation techniques struggle particularly with interpreting the outlines of cells in phase contrast images and aortas in MRIs. The best annotations were collected for fluorescence images, followed by phase contrast images, and then MRIs for all three annotation sources.

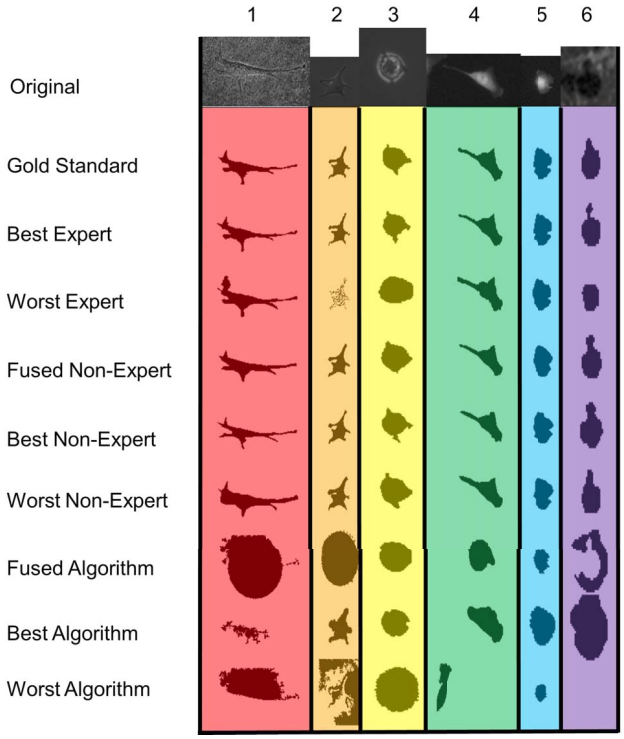

Figure 4. Representative segmentation results. Raw images (row 1), followed by fused, highest-scoring, and lowest-scoring segmentations for experts (rows 2–4), non-experts (rows 5–7), and algorithms are shown for a biological structure from each dataset in the image library (cols 1–6).

standard segmentations for the studied fluorescence images, followed by phase contrast images, and finally magnetic resonance images (**Fig. 3**; Fluorescence, Phase Contrast, MRI). These observations that errors in drawn boundaries are often increasingly severe for experts, non-experts, and algorithms and for fluorescence, phase contrast, and magnetic resonance images are exemplified in **Figure 4**. We found that outliers often stemmed from annotating the incorrect object for humans and identifying no object for algorithms (e.g., **Fig. 4**; col 6, "Worst Algorithm").

We observed that the consistency of quality between annotations was the greatest for experts, followed by non-experts, and finally the least between algorithms (**Fig. 3**). Within each of the three annotation sources, we observe for each dataset there was variability in quality between different sets of collected annotations with respect to the median score and the amount of variability of agreement with the gold standard (**Fig. 5a-c**). Among the six tested algorithms, we found that the gold standard segmentations are most accurately captured by *HoTr* for dataset 1 with a median score of 0.31, *HoTr* for dataset 2 with a median score of 0.59; *SeWe* for dataset 3 with a median score of 0.66; and *Otsu* for dataset 4 with a median score of 0.63; *HoTr* for dataset 5 with a median score of 0.63; and *SeWe* for dataset 6 with a median score of 0.59.

We found that combining segmentations with the fused annotation method led to improved results for both non-experts and algorithms. For non-experts, the median score for the fused annotations was higher than all individual annotation sets for every dataset (**Fig. 5b**). For algorithms, the median score for the fused annotations was higher than all individual annotation sets for datasets 4 and 5 which are the fluorescence datasets (**Fig. 5c**).

We found that 58 workers created all crowdsourced annotations. The drawing tasks for datasets 1 through 6 were completed by 18, 24, 22, 27, 24, and 23 unique workers, respectively, taking on average 60 s, 50 s, 38 s, 36 s 43 s, and 47 s per object, respectively.

5.2. Image Library Characterization

We found that structures in the fluorescence and magnetic resonance images mostly appear rounder, i.e., circularity values closer to 1, than structures observed in the phase contrast images, i.e., circularity values closer to 0 (**Table 1**). This is exemplified in **Fig. 4** with structures in datasets 1 and 2 appearing less round than structures in the other datasets. The difference between the average pixel intensity for the biological structure and background reported in **Table 1**

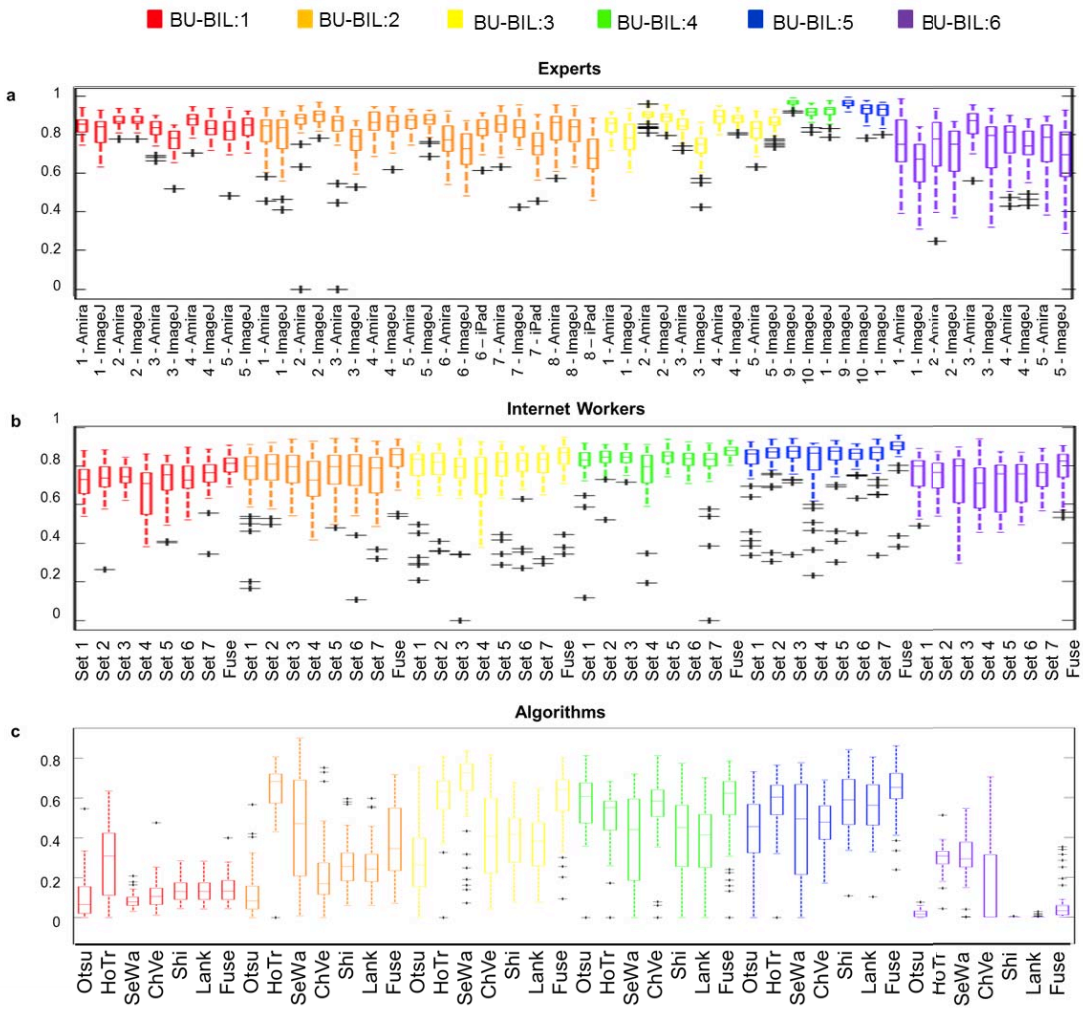

Figure 5. Variability within region overlap ratio scores obtained for each annotation set for each dataset (phase contrast in red, orange, and yellow; fluorescence in green and blue; MRI in purple. See Fig. 3 for the explanation of a box plot visualization). The top plot (**a**) summarizes scores based on different combinations of an expert, annotation tool used by that expert, and dataset. The plot reveals that the performance of experts differs noticeably, especially for phase contrast data, and that annotations of phase contrast images with Amira were more accurate than with ImageJ. The middle plot (**b**) shows scores averaged over the results of each of the seven batches of crowdsourced segmentation annotations collected per each object and the fused annotation created by combining all seven annotations per object. The fused annotation approach yielded the highest median score for all datasets (last box for each color). The bottom plot (**c**) shows that the performance of the algorithms varies widely across datasets. The fused annotation approach was a clear winner for the fluorescence data.

reflects what can be observed in **Fig. 4**, where structures in the fluorescence and magnetic resonance images have a stronger contrast to the background than structures in phase contrast images.

6. Discussion

Our results indicate that all experts and non-experts consistently drew imperfect, yet high-quality segmentations while no single algorithm consistently performed well for allstudied images. We also found that experts, non-experts and algorithms share which image modality/object type was most difficult for them to annotate. Annotations of cells on fluorescence data was most accurate and annotations of aortas on MRI data least accurate. We aimed to conduct our studies on datasets that together represent a diversity of appearances for biological structure types, environmental conditions, and imaging modalities. We suggest BU-BIL and the analyzed segmentation methods as a starting point towards learning which sources generalize well versus cater to particular image conditions.

It is valuable for the research community to realize that the contributions of untrained internet workers can be very close in quality to those of domain experts trained to interpret biomedical images. Such crowdsourced work can be solicited through online annotation systems with easy-to-use graphical user interfaces to inexpensively and quickly

obtain boundaries for biomedical images with consistent accuracy. Our results lead us to suggest that the contributions of online crowdsourced workers without domain-specific backgrounds may be successfully included in a laboratory protocol for segmenting biomedical images.

We were surprised to observe that, among the set of freely-shared algorithms evaluated in this study, no single algorithm worked well in general and that older algorithms regularly outperformed newer algorithms. While we hypothesize that the level set based algorithms may be optimized by tuning parameters and contour initializations to yield better results for specific datasets, we caution against assuming that such tuned methods will effortlessly lead to improved results across the board. We suggest that the observed performance inconsistency of newer segmentation methods should instead motivate future work. This work needs to answer the question how to select an algorithm, among a given set, based on image context so that the best performing algorithm is applied when it will perform best.

7. Conclusions

Analyses on biomedical images often rely on finding boundaries of biological structures and so are influenced by the accuracy of the used segmentations. To examine how to consistently and efficiently collect high quality segmentations, we evaluated 6,148 segmentations created by experts, non-experts, and algorithms on our proposed biomedical image library representing fluorescence, phase contrast, and magnetic resonance images showing cells and aortas. Our study demonstrates that crowdsourced workers are a viable source for replacing domain experts in consistently collecting high-quality segmentations for biomedical images. Our results also reveal that none of the studied algorithms performed well for all datasets in the image library and all algorithms yielded lower quality results than segmentations produced by crowdsourced workers. We facilitate extensions of this work by sharing our image library with all annotations (`http://www.cs.bu.edu/~betke/BiomedicalImageSegmentation`).

Acknowledgments

The authors gratefully acknowledge funding from the National Science Foundation (IIS-1421943, IIS-0910908).

References

[1] M. Helmstaedter, K. L. Briggman, and W. Denk. High-accuracy neurite reconstruction for high-throughput neuroanatomy. *Nature Neuroscience*, 14(8):1081–1088, 2011. 1

[2] A. S. Krupnick et al. Quantitative monitoring of mouse lung tumors by magnetic resonance imaging. *Nature Methods*, 7(1):128–142, 2012. 1

[3] W. Rasband. ImageJ, 1997–2012. U.S. National Institutes of Health, Bethesda, Maryland, USA. 1, 2, 3, 5

[4] Amira, software platform for visualizing, manipulating, and understanding life science and bio-medical data. Retrieved August 17, 2012, from `http://amira.com`. 1, 3, 5

[5] D. Gurari et al. SAGE: An Approach and Implementation Empowering Quick and Reliable Quantitative Analysis of Segmentation Quality. *IEEE Workshop on Applications In Computer Vision (WACV)*, 475–481, 2013. 1

[6] T. Dietenbeck et al. Creaseg: A free software for the evaluation of image segmentation algorithms based on level-set. *IEEE Image Proc (ICIP)*, pages 665–668, 2010. 2, 5

[7] MATLAB. The Mathworks, Inc., Natick, MA. 2, 5

[8] N. Otsu. A threshold selection method from gray-level histograms. *IEEE Trans. Syst. Man, Cybern. B, Cybern.*, 9(1):62–66, 1979. 2, 3, 4, 5

[9] D. Ballard. Generalizing the Hough transform to detect arbitrary shapes. *Pattern Recognition*, 13(2):111–122, 1981. 2, 3, 4, 5

[10] L. Vincent and P. Soille. Watersheds in Digital Spaces: An Efficient Algorithm Based on Immersion Simulations *IEEE Transactions on Pattern Analysis and Machine Learning*, 13(6):583–598, 1991. 2, 3, 4, 5

[11] T. Chan and L. Vese. Active contours without edges. *IEEE Transactions on Image Processing*, 10(2):266–277, 2001. 3, 4, 5

[12] Y. Shi and W. C. Karl. A real-time algorithm for the approximation of level-set based curve. evolution. *IEEE Transactions on Image Processing*, 17(5):645–656, 2008. 3, 4, 5

[13] S. Lankton and A. Tannenbaum. Localizing region-based active contours. *IEEE Transactions on Image Processing*, 17(11):2029–2039, 2008. 3, 4, 5

[14] L. P. Coelho, A. Shariff, and R. F. Murphy. Nuclear Segmentation in Microscope Cell Images: A Hand Segmented Dataset and Comparison of Algorithms *I S Biomed Imaging (ISBI)*, 518–521, 2009. 2

[15] B. Moller and S. Posch. Comparing Active Contours for the Segmentation of Biomedical Images. *International Symposium on Biomedical Imaging*, 736–739, 2012. 2

[16] D. Gurari et al. How to Use Level Set Methods to Accurately Find Boundaries of Cells in Biomedical Images? Evaluation of Six Methods Paired with Automated and Crowdsourced Initial Contours. *The Interactive Medical Image Computation Workshop (MICCAI IMIC)*, 9 pp, 2014. 2

[17] B. M. Good and A. I. Su. Crowdsourcing for Bioinformatics. *Bioinformatics*, 29:1925–1933, 2013. 2

[18] S. K. Kim et al. I'm Cell: A touch pad tool for annotating cell images. *Proceedings of the 1st Biomedical Signal Analysis Conference*, 2014. 3, 5

[19] B. C. Russell et al. LabelMe: a database and web-based tool for image annotation. *International Journal of Computer Vision*, 77(1-3):157–173, 2005. 3, 5

[20] E. Meijering. Cell segmentation: 50 years down the road. *IEEE Signal Processing Magazine*, 29(5):140–145, 2012. 3