

2015

Combining crowd worker, algorithm, and expert efforts to find boundaries of objects in images

<https://hdl.handle.net/2144/14059>

Boston University

BOSTON UNIVERSITY
GRADUATE SCHOOL OF ARTS AND SCIENCES

Dissertation

**COMBINING CROWD WORKER, ALGORITHM, AND EXPERT
EFFORTS TO FIND BOUNDARIES OF OBJECTS IN IMAGES**

by

DANNA GURARI

M.S., Washington University in St. Louis, 2005
B.S., Washington University in St. Louis, 2005

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

2015

© Copyright by
DANNA GURARI
2015

Approved by

First Reader

Margrit Betke, PhD
Professor of Computer Science

Second Reader

Stan Sclaroff, PhD
Professor of Computer Science

Third Reader

Henry Kautz, PhD
Professor of Computer Science

Acknowledgments

First and foremost, I thank my advisor Margrit Betke. This dissertation, from start to finish, would not have been possible without her. I am humbled and grateful that she dedicated five years of her life to foster my professional growth. From the bottom of my heart, I thank her for her unconditional belief in me and willingness to stand by my side through the many rejections. And I thank her for her commitment to make me feel that me and my research are important.

I thank Stan Sclaroff who has often offered himself as a second advisor. I very much cherish his constant honesty and clarity in thought. And I feel privileged that he dedicated so much time to help me learn how to clearly formulate and voice my ideas so they would be interesting and accessible to a broader audience.

I am deeply grateful to Henry Kautz, Tammy Riklin Raviv, and Mark Crovella who kindly offered their time to serve on my committee. I thank each of them for their enthusiasm about my research and encouragement for me to think bigger.

A special thank you to Seule Ki Kim and Mehrnoosh Sameki for trusting me to mentor each of them. I feel my greatest accomplishment over the past five years comes from the role each of them allowed me to play in their lives. Our shared experiences have been truly meaningful and joyous.

Thanks to Kristen Grauman, Robert Pless, Terry Boulton, and Allen Palmer for their mentorship. I am grateful each of them took an interest in me and my development professionally and personally when there are so many individuals who could benefit from their wisdom.

I feel lucky to have worked so closely with Zheng Wu, Diane Theriault, Suyog Jain, Kun He, and Jianming Zhang. I thank each of them for the stimulating conversations, great advice, and countless shared hours of working side by side.

Without my collaborators and friends in the biomedical research community, I am not sure I would have found my research passion. Thank you to Joyce Wong, Chentian

Zhang, Matt Walker, Brett Isenberg, Tuan Pham, Alberto Purwada, Patricia Solski, Chris Hartman, Ulla Hansen, Jennifer Willoughby, Bhavik Nathwani, Andriy Fedorov, Jason Yang, Robert Bao, Wendy Salmon, Hunter Elliott, and David Hildebrand for all the efforts to help me discover and contribute to research problems that target society's health care problems.

I thank all the members of Boston University's Image and Video Computing group. I feel lucky to have learned from colleagues with such a richly diverse set of research interests and perspectives. And a special thanks to the students who attended so many of my practice talks to help me improve my presentations: Mike Breslav, Shugao Ma, Qinxun Bai, Fatih Cakir, Sarah Bargal, Andrew Kurauchi, Wenxin Feng, and Ajjen Joshi.

Thanks to my oldest and newest friends as well as extended family for helping me stay sane and smiling. I am delighted to still be sharing and making memories with Max Gakh, Emily Hathaway, Rachel Loftspring, Laura Krause, Ari Krichiver, Pete Leuenberger, and members of Capoeira Luanda, even if only occasionally. I treasure, will miss dearly, and anticipate many more years of memory making to come with my Boston/Cambridge friends who have spiced up my life over the past five years: Dalia Topelson Ritvo, Kyle Woerner, Daniela Woerner, Avlana Eisenberg, and Jason Yang. A thanks to the many members of Sinha Capoeira for the many wonderful shared experiences in Boston. And a thanks to Matt for loving me whole-heartedly - including during my weakest moments of imperfection and vulnerability as well as my greatest moments of strength and happiness. I thank him for being my reason to dare greatly as I step into my next stage of life.

To the anonymous reviewers at workshops, conferences, and journals as well as the anonymous Mechanical Turk workers, I am thankful they shared their brain power and time to support and grow my research vision.

I would not be who I am without the Kane family. A thank you to Ilan Kane and Galia Kane for helping raise me and loving me unconditionally as family. And a thank you to Ronlee Kane and Doreen Kane for being the best "cousins" one could ask for.

My next thanks go to Giselle Limentani, Dave DeMagistris, and Marguerite DeMag-

istris. I thank Marguerite who, at the age of 93, offered me her perspective and reminded me to live my life with courage. No words can express my gratitude to Giselle and Dave. They adopted me as one of their own children, offering me unconditional love, tenderness, kindness, compassion, and generosity. I am deeply grateful they offer me a peaceful, loving home with them that is constantly filled with laughter and joy.

Without my siblings, I would certainly be a different soul. A thanks to Inbal Gurari for the much big sister support and advice she has shared through her life to help me navigate through the experiences of my life - and, of course, for her sharing the joy of my nephew and niece, Noam Gurari and Lihi Gurari. Thanks to Erez Gurari for perpetually spoiling me to a big brother who generously shares everything from his time to his home. A special thanks to Itai Gurari and Dara Warner for taking an interest in my research and playing a key role in the problems and methods I explored. A special thanks to my twin sister Netta Gurari for her countless hours of offering encouragement and guidance to help me navigate through the PhD process smoothly.

A thank you to my mother, Shaula Gurari, who always believes I am the best... in fact, the best PhD in Boston! I thank her for always doing everything she can to support me to achieve bigger dreams. Her unconditional love and support has carried me through many rough life patches and has been critical to get me to where I am today.

My final thank you goes to my father Eitan Gurari. As a computer science professor, he was my main mentor for many years. He inspired in me a love and fascination for the computer science field. I thank him for planting a seed in my heart that has turned into a research passion which continues to grow and blossom in my life. I lovingly carry him forever in my memory and heart.

COMBINING CROWD WORKER, ALGORITHM, AND EXPERT EFFORTS TO FIND BOUNDARIES OF OBJECTS IN IMAGES

(Order No.)

DANNA GURARI

Boston University, Graduate School of Arts and Sciences, 2015

Major Professor: Margrit Betke, Professor of Computer Science

ABSTRACT

While traditional approaches to image analysis have typically relied upon either manual annotation by experts or purely-algorithmic approaches, the rise of crowdsourcing now provides a new source of human labor to create training data or perform computations at run-time. Given this richer design space, how should we utilize algorithms, crowds, and experts to better annotate images? To answer this question for the important task of finding the boundaries of objects or regions in images, I focus on image segmentation, an important precursor to solving a variety of fundamental image analysis problems, including recognition, classification, tracking, registration, retrieval, and 3D visualization. The first part of the work includes a detailed analysis of the relative strengths and weaknesses of three different approaches to demarcate object boundaries in images: by experts, by crowdsourced laymen, and by automated computer vision algorithms. The second part of the work describes three hybrid system designs that integrate computer vision algorithms and crowdsourced laymen to demarcate boundaries in images. Experiments revealed that hybrid system designs yielded more accurate results than relying on algorithms or crowd workers alone and could yield segmentations that are indistinguishable from those created by biomedical experts. To encourage community-wide effort to continue working on developing methods and systems for image-based studies which can have real and measurable impact that benefit society at large, datasets and code are publicly-shared (<http://www.cs.bu.edu/~betke/BiomedicalImageSegmentation/>).

Contents

1	Introduction	1
2	Evaluation Methodology	7
2.1	Methods	9
2.1.1	SAGE Framework	9
2.1.2	Implementation	10
2.2	Experiments	13
2.2.1	Image Library for Annotation and Annotators	13
2.2.2	Studies	14
2.3	Results	16
2.4	Discussion and Future Work	18
2.5	Conclusions	20
3	Comparative Analysis of Segmentations Created by Experts, Algorithms, and Crowd Workers	22
3.1	Biomedical Image Library (BU-BIL)	24
3.2	Methods	27
3.2.1	Expert-Drawn Annotations	27
3.2.2	Crowdsourced-Drawn Annotations	27
3.2.3	Computer-Drawn Annotations	28
3.2.4	Fused Annotations	30
3.3	Experiments	31
3.3.1	Performance Evaluation Methodology	31

3.3.2	Analysis of Segmentation Sources	31
3.3.3	Image Library Characterization	32
3.4	Results	32
3.4.1	Analysis of Segmentation Sources	32
3.4.2	Image Library Characterization	36
3.5	Discussion	36
3.6	Conclusions	37
4	Crowdsourcing: Domain Expertise Helps & Hurts	38
4.1	Related Work	40
4.2	Datasets and Annotation Methods	41
4.2.1	Image Sets - Defining Levels of Familiarity	42
4.2.2	Open-Ended Drawing Task	43
4.2.3	Closed-Ended Voting Task	44
4.3	Evaluation Methods	45
4.3.1	Characterizing Performance of Crowd Workers	46
4.3.2	Measuring Significance of Observed Results	47
4.3.3	Correlating Worker Effort to Quality of Work	48
4.4	Crowdsourcing Studies	49
4.4.1	Study 1: Drawing on Everyday and Biomedical Images	49
4.4.2	Study 2: Drawing on Images Flipped Upside Down	51
4.4.3	Study 3: Influence of Data Familiarity for Voting Task	54
4.5	Discussion	56
4.6	Conclusions	59
5	Hybrid System - Drawing with Quality Control	60
5.1	Formative Studies	61
5.1.1	Study F1: Expert-Drawn Segmentations on Black and White Images	62
5.1.2	Studies F2 and F3: One Expert Votes for Best Segmentation	63

5.1.3	Study F4: Multiple Experts Vote for Best Segmentation for Everyday Images	64
5.1.4	Study F5: Multiple Experts Vote for Best Segmentation for Biomedical Images	66
5.1.5	Lessons for Crowdsourcing Taken from the Five Formative Studies	67
5.2	Methods	69
5.2.1	SAVE Framework	70
5.2.2	Annotation Collection Implementation	70
5.2.3	Vote Collection Implementation	71
5.2.4	Evaluation Implementation	73
5.2.5	Four SAVE Systems	74
5.2.6	Quantitative Segmentation Performance Analysis	74
5.3	Experiments	74
5.3.1	Image Libraries	75
5.3.2	Crowdsourcing Platform and Participants	75
5.3.3	Performance Analysis for Four SAVE Implementations	76
5.3.4	Characterizing Crowd Behavior	77
5.3.5	Characterizing Successes of Image Segmentation Algorithms	78
5.4	Results	78
5.4.1	Performance Analysis for Four SAVE Implementations	78
5.4.2	Characterizing Crowd Behavior	80
5.4.3	Characterizing Successes of Image Segmentation Algorithms	83
5.5	Discussion	84
5.6	Conclusions	86
6	Hybrid System - Human Initializes Algorithm	88
6.1	Methods	90
6.1.1	Segmentation System	90

6.1.2	Crowdsourced Initial Contour Module	92
6.2	Experiments	92
6.3	Results	93
6.4	Discussion	95
6.5	Conclusions	96
7	Hybrid System - Predicting Computing Source	97
7.1	Predicting Segmentation Quality	100
7.1.1	Prediction Model	100
7.1.2	Training Data Generation	101
7.1.3	Prediction Features	102
7.2	Segmentations by Humans or Computers?	104
7.3	Experiments and Results	106
7.3.1	Predicting Quality of Candidate Segmentation	106
7.3.2	Interaction Tools - Human vs Computer Input	108
7.3.3	Interaction Tools - Human vs Computer Output	112
7.4	Conclusions	116
8	Closing Remarks	117
	Bibliography	118
	Curriculum Vitae	126

Chapter 1

Introduction

The ubiquitous use of cameras and the advance in imaging technology for medical and scientific visualization have resulted in an unprecedented number of images to be analyzed. In response, an explosion of new image-based applications are emerging in both academia and industry which reap a multitude of benefits to society. Demarcating the boundaries of objects (segmentation) is commonly a critical step in image-based applications whether trying to observe object silhouettes [94], collect measurements (features) [47, 77], match images of the same scene (registration) [25, 76], follow objects over time (tracking) [54, 70], differentiate between different types of objects (classification) [87], find similar images in a database (image retrieval) [59, 81], or analyze shapes [74] or behaviors [69]. Consequently, much related work is devoted to obtaining high-quality segmentations whether from domain experts, algorithms, or crowd workers (**Figure 1.1**). Given this rich design space, how should we utilize algorithms, crowds, and experts to consistently collect high quality segmentations?

Domain expertise may be required to understand how object boundaries should be delineated and careful attention may be necessary to manually separate complex object shapes from other objects and/or the background. Manual annotation studies investigate ways to reduce the inter and intra-annotator variability that arises when collecting segmentation drawings from domain experts [65, 89].

Many *computer vision algorithms* were proposed over the past 40 years with different built-in assumptions about image properties that enable them to accurately demarcate object boundaries for a range of object and background appearances observed from

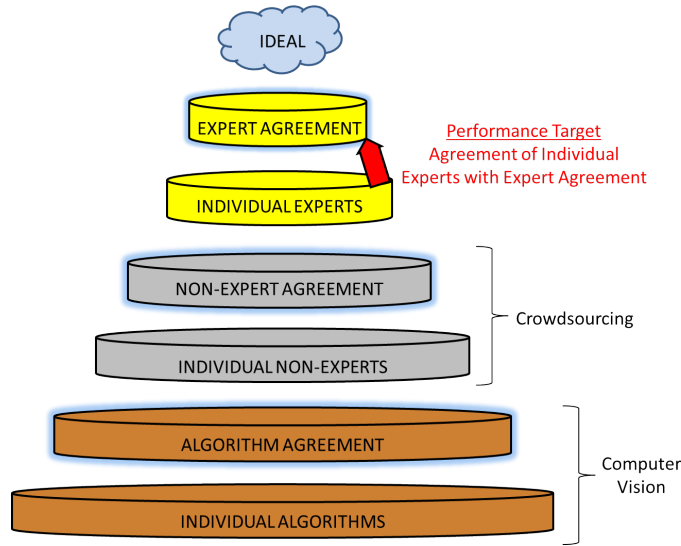


Figure 1.1: Which segmentation collection method will work best in demarcating objects of interest? Overview of resources available to create segmentations. Higher positions for each color in the hierarchy reflect a greater depth of training for the task, the “Agreement” disc for each color reflects a greater breadth of expertise. The width of each disc approximates the relative potential to scale the resource to perform the task. The objective when choosing from these resource options is to maximize quality (higher in the hierarchy) while minimizing cost (lower in the hierarchy).

differing acquisition systems and environmental conditions [6, 15, 16, 48, 35, 68] (**Figure 1.2a,b**). However, such built-in assumptions restrict the wide-spread applicability of such algorithms. Consequently, it can be faster for individuals to manually trace boundaries themselves than to risk repeatedly applying different algorithms until finding one to trust (assuming a suitable option exists).

More recently, as an alternative option, researchers across several fields have proposed to offload the labor-intensive, segmentation task to *crowdsourced workers* [8, 26, 35, 36, 37, 79]. The widespread importance of this approach is exemplified when looking even just at its popularity within the computer vision community. LabelMe [79], a freely-shared, web-based drawing tool regularly used for crowdsourcing, has a website counter indicating 267,392 visits to the website over the past decade and the publication about this work has been cited 1,381 times to date (**Figure 1.2c**). Moreover, larger sizes of annotated datasets for algorithm training and benchmarking are reported annually, with a recent



Figure 1.2: Which segmentation collection method will work best in demarcating objects of interest? Popular image editing tools support numerous automatic methods including those shown for (a) Gimp [1], an open-source replacement for Adobe Photoshop and (b) Fiji [80], a version of the widely-used bioimage analysis software ImageJ. (c) LabelMe [79], is a freely-available web-based drawing tool frequently used for crowdsourcing.

publication describing how an extended version of an annotation tool [8] was applied to create hundreds of thousands of segmented objects in images [55]. Until the work described in this thesis (some published elsewhere [35, 36, 37]), publications about crowdsourcing the image segmentation task only discussed studies conducted on “everyday images” showing objects such as a tree or swan captured with visible cameras. Consequently, little was known regarding what to expect when applying such systems for the vast amounts of imagery that show content undetectable to the naked human eye such as microscopic images showing cells or magnetic resonance images showing aortas in a heart (**Figure 1.3; BU-BIL:1-6**). Another gap in the literature was that little work qualifies what to expect in terms of quality from crowdsourced workers compared to experts, despite known mistakes observed from collected crowd drawings.

In parallel with proposing segmentation creation methods, researchers have proposed *quality control methods* to address concerns about the quality of segmentations created by experts, algorithms, and crowdsourced workers. Ensemble methods combine multiple segmentations created by humans, computers, or a mixture of both to obtain a better

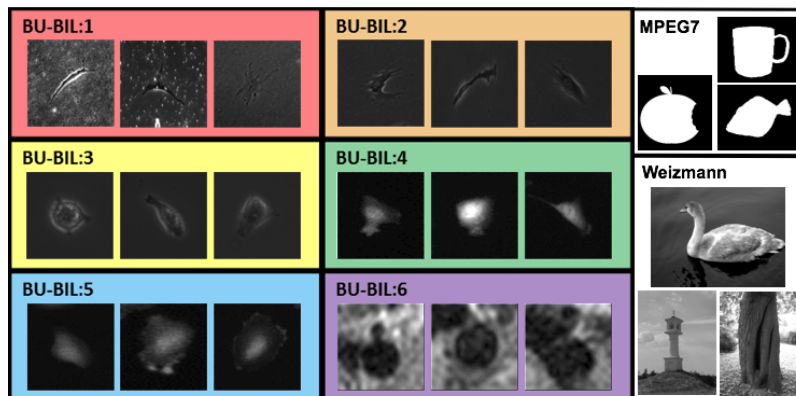


Figure 1.3: This paper examines how to leverage crowd efforts in delineating the boundary of objects in biomedical and everyday images using studies on images taken from three image libraries (BU-BIL, MPEG7, and Weizmann). The focus is on the single primary object in each image. The images shown here exemplify how object appearance can vary significantly with respect to intensity, size, and shape, how edges separating objects from the background can be faint, and how the background can be noisy and cluttered.

segmentation [27, 95]. Crowdsourcing quality control strategies have been proposed to filter workers using training tests or to grade submitted segmentations from the immediately available, yet potentially unreliable crowdsourced workforce [55, 85]. The challenge with applying quality control methods is knowing what is the benefit of each method and so when to apply which method [82].

This work is a contribution to the emerging research field at the intersection of human computation and computer vision that explores how to involve humans to contribute to computing in *hybrid algorithm-crowdsourcing systems*. Recent publications have suggested to engage crowdsourced workers to supply initial bounding regions coarsely hugging object boundaries which the algorithms then evolve to the final desired object boundaries [44]. While such hybrid methods are effective for particular image sets, they only succeed when the algorithm assumptions match the image properties. Another set of hybrid systems show how to pair algorithms with crowdsourced workers in a system workflow to create segmentations of biological structures in microscopy images [28, 42, 45]. These works discuss different hybrid system workflows targeted for specific image sets. Our work, described in this thesis and elsewhere [31, 34, 36], grows the limited body of research on hybrid system

designs for the image segmentation problem by validating system workflows that combine crowdsourced lay people and computer vision algorithms with different workflows on a diversity of image content.

Finally, this work and associated publications [31, 34, 36, 37] are a contribution to crowdsourcing literature that examines how to leverage *non-expert humans to replace domain experts* for extracting information from biomedical images [29]. Exemplar image-based research revealing how to leverage the crowd include classifying galaxies [56], malaria infected red blood cells [60], and colorectal polyps (precursor to malignant cancer) [67] observed in large numbers of infrared, microscopy, and computed tomography images respectively. These citizen science and gamification studies motivate continuing to challenge commonly-held assumptions regarding when expert training is required. We extended existing research by conducting studies [31, 34, 36, 37] using a crowdsourcing internet marketplace, Mechanical Turk, which provides a set of incentives for participation that is different from that of citizen scientists or gamers and so can lead workers to behave significantly differently from volunteers [61]. We chose to conduct experiments using monetary incentives with the paid crowdsourcing platform, Mechanical Turk, because of “easy access to a large, stable, and diverse subject pool, the low cost of doing experiments and faster iteration between developing theory and executing experiments” [62]. Our work demonstrates how to involve paid crowdsourced workers in expert-quality systems.

The key contributions of this dissertation are:

- A principled approach for analyzing segmentation performance that connects annotation collection methods, fusion methods, and evaluation algorithms into a unified framework called SAGE; this approach simplifies the challenge of finding a suitable replacement for an expert by incorporating into the evaluation approach the inconsistencies observed between expert annotators [32].
- Evaluation and comparison of experts, crowdsourced non-experts, and automated segmentation algorithms to find the boundaries of biological structures in biomedical

images [37].

- Analysis of crowdsourcing the image segmentation problem that reveals what may be expected when leveraging crowd workers at different levels of involvement for both familiar (everyday images) and unfamiliar (biomedical images) content: 1) draw only, 2) vote only, or 3) both draw and vote [33, 34].
- Three hybrid system designs and experiments that inform how to utilize the annotation efforts of crowdsourced workers and algorithms together to create object boundaries that are of comparable quality to segmentations created by experts and exceed the performance of pure algorithm and crowdsourcing methods [31, 34, 36].

The remainder of this dissertation is organized in six sections. A methodology and series of three studies that motivate a new segmentation evaluation methodology is described in Chapter 2. These experiments inform how to involve experts to establish references for segmentation evaluation purposes. Then, we describe a comparative analysis of drawings from algorithms, crowd workers, and experts in Chapter 3. Next, in Chapter 4 we discuss crowdsourcing studies suggesting how to more effectively involve crowd workers to gather high quality results. Finally, we conclude in Chapters 5-7 with three hybrid system designs and experiments that highlight ways to combine efforts from algorithms and crowdsourcing to efficiently collect high quality segmentations.

Chapter 2

Evaluation Methodology

An important foundation for image-based applications is demonstrating that the segmentation method consistently provides the desired outcome. Performance analysis of segmentation methods varies depending on the application objectives. Zhang [97] proposed to group evaluation methods into three categories, “analytical methods,” empirical methods based on goodness measures, and empirical methods based on discrepancy measures. Zhang [97] 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. In this work, also discussed in a 2013 publication [32], we focus on performance analysis of segmentations using the current common model, empirical methods based on discrepancy measures.

Prior to this work, there was little discussion about when to use which segmentation analysis method when calculating discrepancy scores. Numerous papers reviewed *evaluation methods* for finding a discrepancy between two segmentations [43, 88, 97]. An active area of research lied in establishing an *annotation collection* process to obtain gold standard segmentations including studies about annotation tools and annotator expertise level [5, 19, 57, 65, 73, 79]. Additionally, *annotation fusion* methods were developed to produce a reliable gold standard segmentation from a collection of annotations for the cases when intra-annotator and inter-annotator variation could be high [12, 17, 85, 95].

Finding the appropriate methodology for analyzing a segmentation method is important for recognizing an effective algorithm or crowdsourcing system design. For example, developers may prematurely dismiss good segmentation systems when their measures in-

dicating 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 is insufficient to approach segmentation analysis by only selecting a discrepancy measure [97], because the chosen gold standard segmentation also impacts the score. Furthermore, access to various segmentation analysis tools and methods is critical for establishing accepted segmentations. 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 published work [32] 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.
- 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.

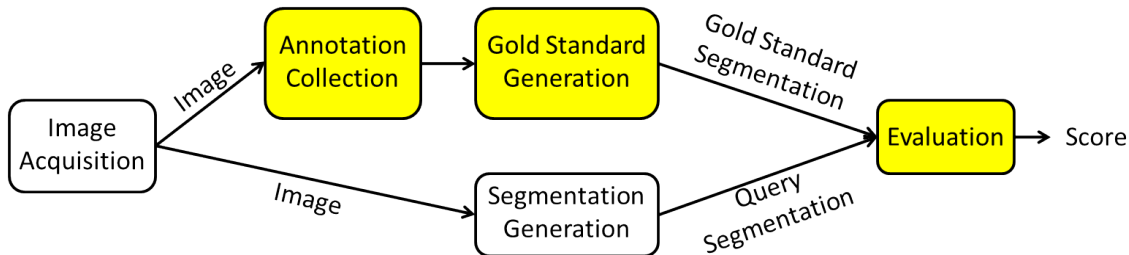


Figure 2.1: Overview of SAGE (yellow boxes) within the context of analyzing a query segmentation.

2.1 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 **C**ollection, **G**old **S**tandard **G**eneration, and **E**valuation (SAGE). We then describe a freely available system implementing this framework.

2.1.1 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 2.1**. 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 current model [97] for analyzing segmentation performance which considers only selecting the appropriate evaluation measure to establish a score.

Since one would use this model 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 2.1**). 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 2.1**). 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.

2.1.2 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. 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 [79], ImageJ [73], and Amira [5]. 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 [65] and Simultaneous Truth and Performance Level Estimation (STAPLE) [95].

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:

$$W_i \equiv f(T_i = 1 | D_i, p, q) = \frac{a_i}{a_i + b_i} \quad (2.1)$$

$$a_i = f(T_i = 1) \prod_{j:D_{ij}=1} p_j \prod_{j:D_{ij}=0} (1 - p_j) \quad (2.2)$$

$$b_i = f(T_i = 0) \prod_{j:D_{ij}=0} q_j \prod_{j:D_{ij}=1} (1 - q_j) \quad (2.3)$$

The EM algorithm uses equation 2.4 to calculate the expected conditional log likelihood in the E -step and equations 2.5-2.6 to update the performance parameters for the M -step.

$$Q(\theta^t | \theta^{t-1}) = \sum_j [\sum_{i:D_{ij}=1} W_i^{(t-1)} \ln p_j + \sum_{i:D_{ij}=1} (1 - W_i^{(t-1)}) \ln(1 - q_j) + \sum_{i:D_{ij}=0} W_i^{(t-1)} \ln(1 - p_j) + \sum_{i:D_{ij}=0} (1 - W_i^{(t-1)}) \ln q_j] \quad (2.4)$$

$$p_j^{(k)} = \frac{\sum_{j:D_{ij}=1} W_i^{(k-1)}}{\sum_i W_i^{(k-1)}} \quad (2.5)$$

$$q_j^{(k)} = \frac{\sum_{j:D_{ij}=0} (1 - W_i^{(k-1)})}{\sum_i (1 - W_i^{(k-1)})} \quad (2.6)$$

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

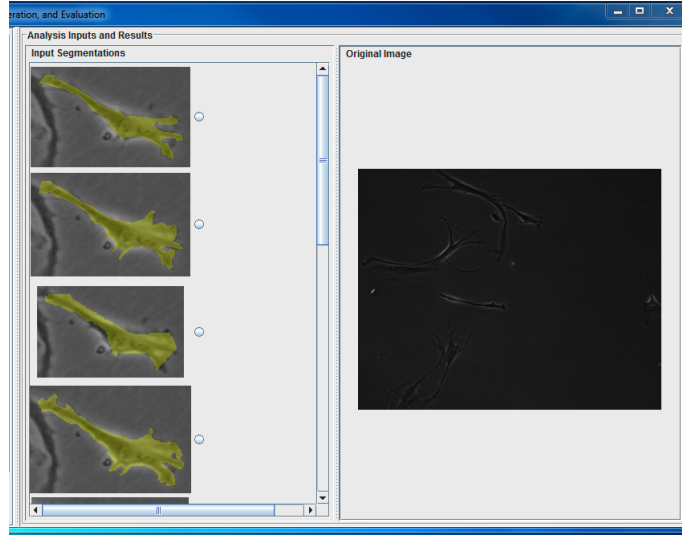


Figure 2.2: Interface of the toolbox for selecting a gold standard from annotations and fused annotation options.

annotations and fusion segmentations is shown in Figure 2.2.

Evaluation Measures: The system supports the following six discrepancy measures commonly used to indicate how far a query segmentation is from a gold standard segmentation - accuracy, intersection over union (IoU), false positive rate, false negative rate, probability of error, and Hausdorff distance [43, 88, 97]. 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 $\frac{|A \cap B|}{|A|}$; **IoU** 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 .

2.2 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.

2.2.1 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% CO₂ on a PAAM hydrogel with embed-

Table 2.1: 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 2.2: Description of annotator experience

ID	Education Level	Worked with cell images	Worked with MRI images	Used ImageJ	Used Amira
A	Undergrad	3 months	None	Yes	No
B	Post-doc	14 years	3 months	Yes	No
C	PhD student	10 years	1 year	Yes	No
D	Post-doc	2 months	None	Yes	No
E	PhD student	3 weeks	1 year	Yes	No

ded fluorescent beads with a size of 0.75 microns. For the second dataset, the cells were cultured at 37°C in 5% CO₂ on a PAAM hydrogel. For the third dataset, the cells were cultured at 37°C in 5% CO₂ on tissue culture plastic. The fourth dataset contains magnetic resonance images (MRIs) 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 2.1**.

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 2.2**.

2.2.2 Studies

Study 1: Impact of Annotation Tool. The five annotators were asked to annotate the first 154 images with two annotation tools, ImageJ [73] and Amira [5], using their own judgement. ImageJ, like LabelMe [79], 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, IoU, 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.

2.3 Results

Study 1: Impact of Annotation Tool. Qualitative results of a set of annotations for an image from each dataset are shown in **Figure 2.3** 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 2.3**.

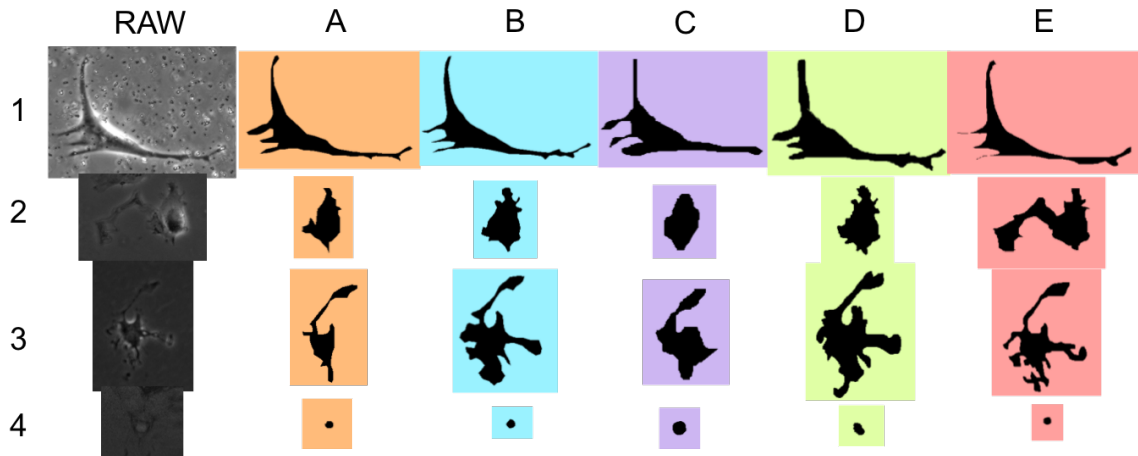


Figure 2.3: Qualitative results showing a set of annotations collected using ImageJ from the five annotators (A-E) for an image from each dataset (1-4).

Study 2: Impact of Annotators. For the post-processed data, the average eval-

Table 2.3: 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	IoU	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

uation 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 2.4**.

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.

Table 2.4: 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 $value \times 10^{-2}$.

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

2.4 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 2.3** and quantitatively in **Table 2.4**. 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 various tuned parameters) while keeping the annotation collection process constant (same collection of input annotations) and evaluation measure constant causes the output score to vary [12]. 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 2.4**.

The results support that SAGE is a principled approach to use when analyzing segmentation quality. The results also suggest that SAGE more accurately describes the factors impacting the performance score than the previous model [97].

Study 1: Impact of Annotation Tool. Images in **Figure 2.3** 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 IoU measures, while lower values are better for the other four measures. In contrast to the findings in Meyer et al’s work [65], 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 2.3** 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 IoU 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, IoU, 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 [12, 95], we investigate whether the fusion methods are perceived to provide improved segmentations over the original annotations. Experimental results indicate a low preference for fusion methods over original annotations by a single expert for our datasets. Results also revealed that a simple pixel majority vote consensus algorithm was perceived by experts as the better option when considered against the widely-accepted expectation maximization consensus algorithm [95] that intelligently weighs the influence of each expert annotation on the final segmentation.

2.5 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 revealed that the annotation tool introduced the greatest amount

of disagreement between experts' annotations among the studied factors annotator education, annotator experience, and annotation tool. Given inconsistencies observed between experts, we suggest as an evaluation methodology to set the performance goal to evaluate query segmentations against a reference segmentation established through a pixel majority vote of multiple expert annotations.

Chapter 3

Comparative Analysis of Segmentations Created by Experts, Algorithms, and Crowd Workers

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*) [42, 47]. 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. 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?”

Often, *domain experts* draw the boundaries of biological structures using annotation software such as ImageJ [73] or Amira [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 integrating *segmentation algorithms* into publicly available image analysis systems and researchers have been designing new algorithms to tackle open seg-

mentation challenges [20, 63, 73, 93]. Older methods including thresholding (e.g., Otsu Thresholding [68]), feature-based (e.g., Hough Transform [6]), and region growing (e.g., Seeded Watershed [91]) 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 [20]. 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 [2, 7, 9, 18, 41, 66].

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 [29]. 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, also discussed in a 2015 publication [37], 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:

- Publicly sharing a library of images collected and used for biomedical research with associated expert annotations
- 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

Table 3.1: Salient properties characterizing each dataset in the image library and the number of objects per dataset. PC represents phase contrast microscopy, FI represents fluorescence microscopy and MRI represents magnetic resonance imaging.

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

of the laboratory protocol for segmenting biomedical images

3.1 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 3.1**). Our image library includes six datasets that represent three imaging modalities and six object types. 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 (**Figure 3.1**). We call this collection the Boston University Biomedical Image Library (BU-BIL) and share it publicly (<http://www.cs.bu.edu/~betke/BiomedicalImageSegmentation>).

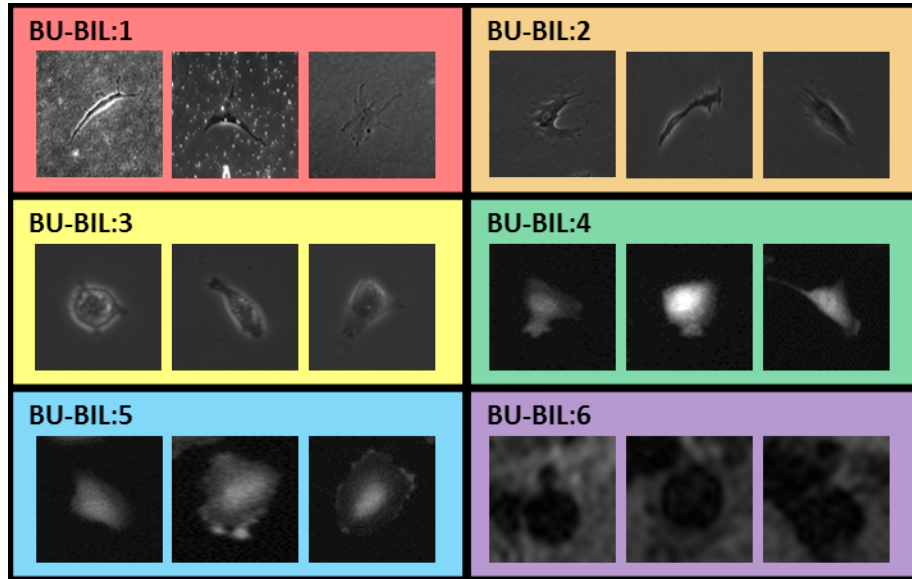


Figure 3.1: Representative images from the six datasets in the image library. Segmentation methods that accurately delineate boundaries of biological structures must handle appearance variations with respect to intensity, size, and shape; faint edges separating structures from the background; and backgrounds with clutter.

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/cm² 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 $0.75\text{-}\mu\text{m}$ fluorescent beads to facilitate imaging of gel deformation, and incubated at 37°C in 5% CO₂ 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% CO₂ 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% CO₂ 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 micro-fabricated polydimethylsiloxane (PDMS) stencil with a 600 micron hole. After 6 hours incubation at 37°C in 5% CO₂, 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 3.1**).

3.2 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.

3.2.1 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 [73], Amira [5], and an iPad touchpad drawing program [24]. 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.

3.2.2 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 [79]. 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 (AMT) internet market-

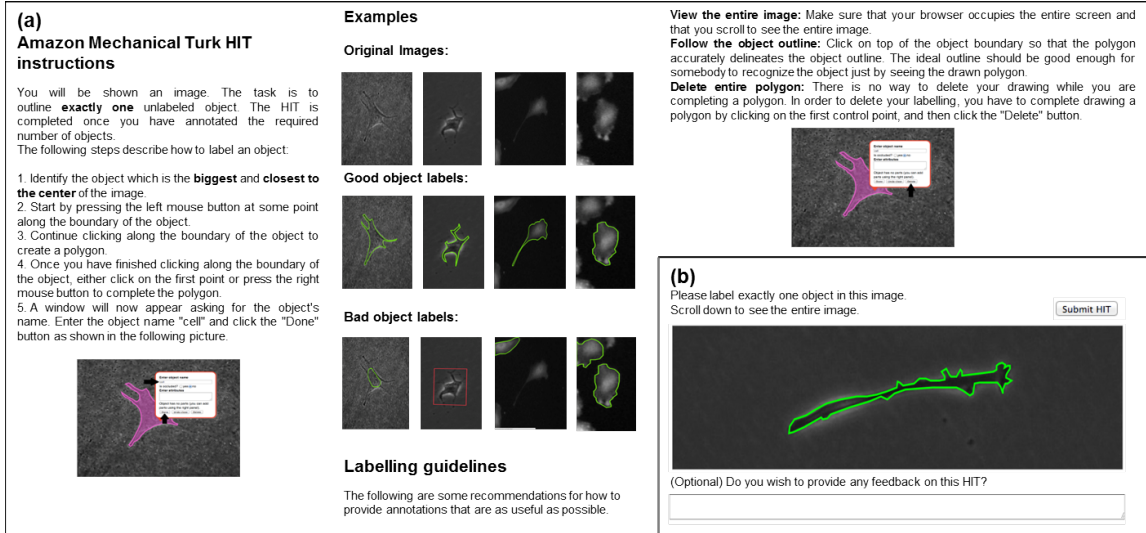


Figure 3.2: Crowdsourcing user interface. An example of (a) the annotation instructions given for datasets 1-5 and (b) a segmentation annotation created using the interface that internet workers use to complete the drawing task, LabelMe.

place. We posted each drawing task for each image to AMT as a human intelligence task (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 (**Figure 3.2a**). After accepting the HIT, a worker was then presented the drawing interface to create the object boundary (**Figure 3.2b**). 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 HITs and received at least a 92% approval rating.

3.2.3 Computer-Drawn Annotations

We evaluated six publicly available algorithms that represent four key classes of algorithms commonly reported in the literature for biomedical images [64]: thresholding (i.e., *Otsu thresholding* [68]), feature-based (i.e., *Hough transform for circles* [6]), region-growing (i.e., *seeded watershed* [91]), and deformable models (i.e., *Chan Vese level set method* [16], *Lank-*

ton region-based level set method [48], and *Shi approximation level set method* [83]).

Otsu thresholding (Otsu) is based on the assumption that biological structures (“foreground”) have different intensity values than the background [68]. 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 [6]. 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 [91]. 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 *HoTr* segmentation for the background marker and the eroded *HoTr* 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 [16]. The *Shi approximation level set method (Shi)* computationally speeds up the evolution process by replacing slow real-valued calculations with faster integer-based calculations [83]. *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 [48]. 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.

Table 3.2: 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 [5]; ImageJ [73]; iPad touchpad drawing program [24]
Non-Expert Annotators (NoEx)	LabelMe [79]
Otsu Thresholding [68] (Otsu)	MATLAB [63]; ImageJ plug-in [73]
Hough Transform for Circles [6] (HoTr)	MATLAB [63]; ImageJ plug-in [73]
Seeded Watershed [91] (SeWa)	MATLAB [63]; ImageJ plug-in [73]
Chan Vese level set method [16] (ChVe)	MATLAB [20]
Shi approximation level set method [83] (Shi)	MATLAB [20]
Lankton region-based level set method [48] (Lank)	MATLAB [20]

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 3.2**).

3.2.4 Fused Annotations

We evaluated 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 then post-processed the segmentation result by filling holes and keeping only the largest object.

3.3 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?

3.3.1 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 IoU metric (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). Recall that 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 resulting segmentations from the majority pixel votes of all expert-drawn segmentations per image.

3.3.2 Analysis of Segmentation Sources

We computed the IoU score 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 3.2.4**) 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 IoU score for all resulting segmentations.

3.3.3 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.

3.4 Results

3.4.1 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 (**Figure 3.3**; 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 standard segmentations for the studied fluorescence images, followed by phase contrast images, and finally magnetic resonance images (**Figure 3.3**; Fluorescence, Phase Contrast, MRI). These observations

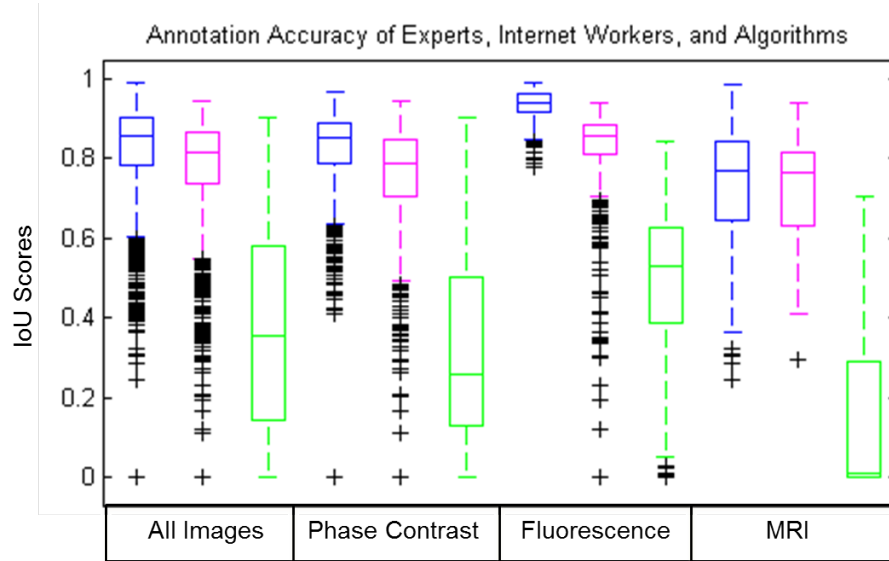


Figure 3.3: IoU scores for segmentations created by experts (blue), non-experts (magenta), and algorithms (green), averaged over all data, and data of each of the three image modalities. For each annotation source, the central mark of the box denotes the median score and the box edges the 25th and 75th 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.

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 3.4**. We found that outliers often stemmed from annotating the incorrect object for humans and identifying no object for algorithms (e.g., **Figure 3.4**; 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 (**Figure 3.3**). 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 (**Figure 3.5a-c**).

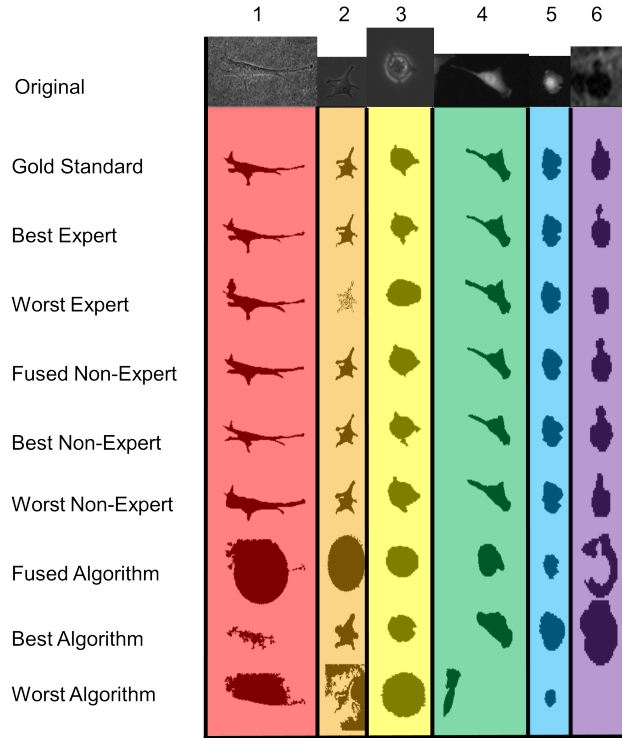


Figure 3.4: Representative segmentation results. Raw images (row 1), followed by fused, highest-scoring, and lowest-scoring segmentations for experts (rows 2–4), non-experts (rows 5–7), and algorithms are shown for a biological structure from each dataset in the image library (cols 1–6).

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 (**Figure 3.5b**). 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 (**Figure 3.5c**).

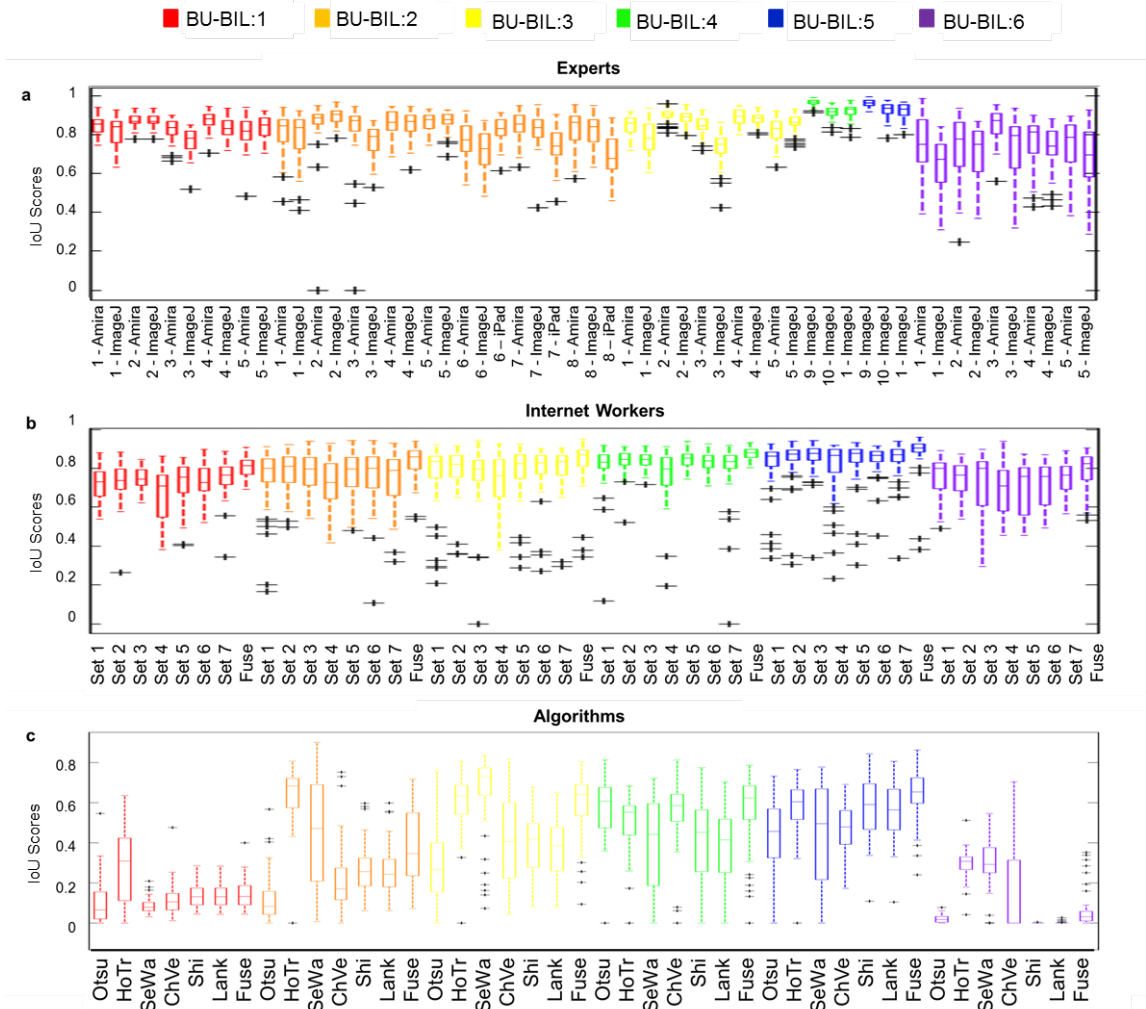


Figure 3.5: Variability within IoU scores obtained for each annotation set for each dataset (phase contrast in red, orange, and yellow; fluorescence in green and blue; MRI in purple. See Figure 3 for the explanation of a box plot visualization). The top plot (a) summarizes scores based on different combinations 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.

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.

3.4.2 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 3.1**). This is exemplified in **Figure 3.4** 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 3.1** reflects what can be observed in **Figure 3.4**, where structures in the fluorescence and magnetic resonance images have a stronger contrast to the background than structures in phase contrast images.

3.5 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 all studied 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 annota-

tion 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.

3.6 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 annotations.

Chapter 4

Crowdsourcing: Domain Expertise Helps & Hurts

In a 2013 study, researchers discarded 33,508 crowdsourced drawings of everyday content, i.e., 32% of collected data, because the results were not “deemed to be good” [8]. Conversely, our study [37] (discussed in the previous chapter) demonstrated that crowdsourced drawings on biomedical image content nearly matched the quality of drawings from domain experts. These contrasting findings are surprising. Could the hidden secret for success on unfamiliar biomedical image content be generalized to familiar everyday image content? Why are there differences in the quality of crowd work reported for the two drawing studies?

Two schools of thought lead to two plausible different ways to explain the poor quality crowd work observed in the 2013 study [8]. The key difference between these ideologies lies in whether or not to trust crowd workers.

One popular approach to deal with poor crowd work is to infer that the problem lies with crowd workers. As posited by *Bernstein et al.* many crowd workers are either “Lazy Turkers” or “Eager Beavers.” So, as “a rule-of-thumb, roughly 30% of the results from open-ended tasks are poor” [11]. According to this interpretation, the 32% of discarded drawings in the 2013 study [8] makes sense and should be expected.

Alternatively, one may infer that poor crowd work is a consequence of an inadequate task design. As discussed by Lease [52], “When we ask users to perform a task that is simple and obvious to us, yet they screw it up, we may infer perhaps that the workers are lazy or deceitful, when in fact it may be our own poor design that is truly to blame.” Based on this interpretation, the large fraction of wasted crowd effort offers a sign that the task design in the 2013 study [8] could have been improved to yield higher quality results.

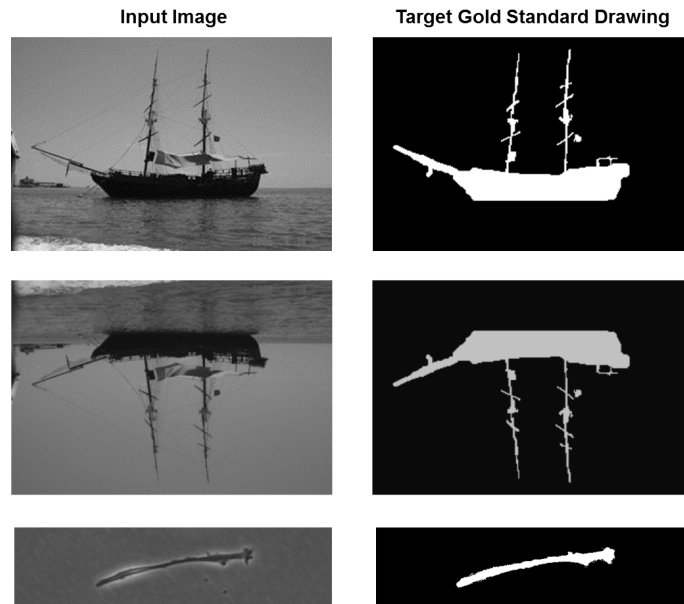


Figure 4.1: When asked to draw the boundary of an object in an image, how are crowd workers influenced by the familiarity of the data? Will crowd workers draw faster and better when the image content is less or more familiar?

In our work [33], we explored the problem of poor quality crowd work by examining how crowd workers’ skills and judgments are influenced by the familiarity of the data. We focused on the open-ended *segmentation problem* of drawing the boundary of a single object in an image (**Figure 4.1**). We were pleasantly surprised to observe that making the data less familiar not only triggered crowd workers to produce significantly better drawings but crowd workers also took significantly less time. Our findings highlight that crowd workers’ recognition of the content can lead to under-utilization of their skills. Our results offer hints that poor crowd performance may be due to workers’ cognitive overload from a complicated task rather than lack of sufficient effort in accomplishing the task. This work offers promising evidence that more efficient, higher quality crowdsourcing system designs can be inspired by applying methodologies to learn one’s own biases.

The remainder of this chapter is organized into six sections. Related work is reviewed in the next section. Then, our crowdsourcing systems and evaluation methodologies are described in the next two sections for the open-ended drawing task and closed-ended voting

task. In the following section, experiments and results for three crowdsourcing studies that explore how data familiarity relates to the drawing skill and judgment of crowd workers are discussed. Finally, we conclude with a discussion and summary about our contributions.

4.1 Related Work

To date, much of the rush towards crowdsourcing the segmentation problem, delineating the border of an object in an image, has been motivated by the desire to build larger annotated datasets, unparalleled in size to those possible from a single, local group. Consequently, crowdsourcing efforts in the computer vision [55, 79] and computer graphics [8] research communities have led to new datasets that consist of hundreds of thousands human drawings. These annotations empower researchers to train machine learning algorithms on more diverse datasets and evaluate automated segmentation algorithms more rigorously. These annotations also support researchers to build search engines that effectively mine images.

Additional interest in crowdsourcing the segmentation problem is also observed with researchers in need of drawings at run-time. For example, crowdsourced drawings serve as computations within state of art computer vision [44] and crowdsourcing [39] systems. In these cases, the use of imperfect crowd drawings instead of (more seriously incorrect) automatically-produced outlines.

Given the widespread interest in crowdsourcing the segmentation problem, there is clear benefit across many communities in improving the design of crowd drawing systems. In particular, researchers commonly report similar warnings about the quality of crowd drawings: “Most workers only produce a coarse outline of the instance resulting in poor segmentations” [55]. While recent research has predominantly focused on developing new web-based tools to more efficiently elicit high quality drawings from the crowd [13, 26, 57, 85], we demonstrate how to collect higher quality drawings based on knowledge about how crowd workers behave with respect to different types of images.

Our work considers how to perform quality control for crowdsourcing. Approaches can be categorized into those that are applied at run-time versus during the design phase.

Run-time quality control approaches introduce additional machinery that reportedly yield higher-quality crowd work. In particular, mechanisms have been designed to filter out workers with insufficient training qualifications [55], edit or validate crowd work [11, 86], or mitigate the influence of poor quality work through redundancy [82]. Filtering workers has the undesirable consequence of limiting the crowd worker pool which, in turn, reduces the degree to which such a crowdsourcing solution can scale. Cleaning poor quality data leads to a loss of money both from collecting the poor quality results and then applying machinery to fix/filter the results. In addition, cleaning poor quality results introduces a delay to acquire results making such approaches less amenable to “real-time” applications.

Alternatively, quality control during the design phase often involves human factors studies to tease out richer information regarding how human behavior relates to various task designs. For example, when choosing how to attract a crowd, one may be influenced by understanding how crowd behavior is related to different incentives (i.e., pay versus volunteer) [61] or cultural biases [71]. A human factors approach is commonly adopted for expert studies. Learned causes of expert disagreement help researchers improve their theories and methodologies [65]. In this work, we chose a human factors approach. We demonstrate that teasing out more detailed information about why crowd behavior is breaking down can lead to the collection of higher quality annotations and a reduced cost.

4.2 Datasets and Annotation Methods

Our goal is to examine how the familiarity of image content influences crowd workers’ drawing skills and perceptions of the difficulty of the drawing task. To do this, we created two on-line crowdsourcing systems that run within the Amazon Mechanical Turk (AMT) internet marketplace. We prepared a crowd drawing environment by configuring a secure web server in an Ubuntu computing environment on the Amazon Elastic Compute Cloud

(EC2). We then installed the open-source LabelMe code [79] and configured scripts that we ran to post our drawing jobs to AMT and record the submitted crowd results. We prepared our crowd perception system by adapting a Human Intelligence Task (HIT) template from AMT, which is HTML code that supports both displaying our instructions and task as well as recording the crowd submitted results. We describe below the datasets and annotation user interfaces used in our studies.

4.2.1 Image Sets - Defining Levels of Familiarity

We used a total of 405 images coming from two public datasets for our crowdsourcing studies. We selected the two image sets because they were intentionally designed to only include images that have a single, dominant object of interest. We also chose the two image sets because they include expert-drawn, pixel-accurate delineations of the object of interest for each image that we could use for evaluation purposes. Finally, we chose content that is both detectable and undetectable to the naked human eye in order to capture images that are more and less likely to be familiar to a lay person. We define three image categories with the 405 images to represent various levels of content familiarity.

4.2.1.1 Unfamiliar Image Content

We included a total of 305 biomedical images from BU-BIL [37] to represent content undetectable to the naked human eye. Ambiguity regarding the object of interest is minimized because images were cropped to only contain the objects of interest.

4.2.1.2 Familiar Image Content

We included 100 images [4] that were collected with visible cameras and so represent content detectable by the naked human eye. The designers of the dataset chose images from royalty free image databases that “avoid potential ambiguities” regarding the object of interest because the objects of interest differ from the “surroundings by either intensity,

texture, or other low level cues.” Images show objects such as animals, trees, buildings, and boats. We infer that crowd workers are likely to be familiar with the objects.

4.2.1.3 Semi-Familiar Image Content

We flipped the 100 familiar images [4] vertically so that we had 100 upside down images. For example, a boat becomes situated such that the water resides above the skyline (**Figure 4.1**). We infer the crowd workers’ drawing performance will be influenced by the familiarity of the content less than when images are upright.

4.2.2 Open-Ended Drawing Task

Our crowd drawing system is a two-step process where workers are first shown instructions and then the interface they use for drawing. We describe both steps below.

4.2.2.1 Annotation Instructions

When a crowd worker on AMT reviews our posted drawing job, he/she is shown the instructions (**Figure 3.2a**). Included are five steps described in English. To minimize concerns regarding the annotation protocol, the instructions emphasize that a worker should annotate the single object which is the largest and closest to the center of the image. Included are also pictures exemplifying correct and poor annotations to clarify that the aim of the task is to create a highly detailed annotation of the single, most prominent object in the image. Examples are intended to address various annotation concerns, such as the common complaint that crowd workers create coarse rather than highly-detailed outlines [55].

4.2.2.2 Annotation Tool

After a worker accepts our HIT, the instructions embedded in the AMT webpage are replaced with the drawing user interface (**Figure 3.2b**). Workers are presented the annotation tool from the computer vision community, LabelMe [79]. After completing the drawing, the worker is prompted with a message allowing the user to delete the drawing, in

a Motivation.






The goal of this task is to help computer scientists build smarter systems to find objects in images. Your responses to our following yes/no questions will help with designing these systems.

Question.

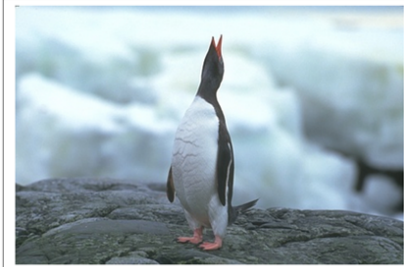
If we asked multiple people to draw the boundary of a **single** object in the given image, do you think all people would pick the same object?

Instructions.


1. Look at each image separately.
2. If you think all people would draw the boundary of the same object when presented the image, choose "yes". Otherwise, choose "no". Look to the right for examples.

Yes	No
	
	
	

b Tasks



Yes
 No



Yes
 No

Figure 4.2: (a) Instructions and (b) user interface for our voting task used by crowd workers in the Amazon Mechanical Turk internet marketplace.

case he/she made a mistake and so wants to redraw the object. Otherwise, the worker must specify a text label naming the object and click “Done” in order to submit the completed drawing.

4.2.3 Closed-Ended Voting Task

Our aim is to learn whether the drawing task is clearly-defined for crowd workers. In particular, one may want to include a voting step to forego the expensive drawing task if workers can reliably deem whether a task seems ambiguous. Our crowd voting system is a one step-process where workers can see the instructions and the user interface on the same webpage before deciding whether to accept the HIT.

4.2.3.1 Annotation Instructions

Inspired by formative studies, we settled on a task header that included the problem motivation, task question, and then two steps instructing how to perform the task (**Figure 4.2a**). We asked workers to answer the following question about an image: “If we asked multiple people to draw the boundary of a single object in the given image, do you think all people would pick the same object?” We intentionally specified criteria that aligns with the drawing task we used in practice. Also, in an effort to help workers feel their contributions are valued, we stated that the long-term aim of the task is to support computer scientists to build systems. Finally, to clarify the aim of the task, we included pictures exemplifying when to label an image with “Yes” versus “No” to indicate well-defined versus ambiguous drawing tasks respectively.

4.2.3.2 Annotation Tool

We presented a set of five images per HIT to increase study efficiency. Each image is shown in a column on the left and the crowd worker casts a vote by selecting one of two radio buttons to the right of each image to indicate “Yes” or “No” (**Figure 4.2b**). Once a worker completes voting on the five images, the worker clicks a button to submit the results. To minimize concerns about worker quality, we used the majority vote answer from five collected answers to assign the image label. To minimize the potential impact of biases related to image clusterings, we chose each set of five consecutive images to pair in the same HIT based on five different randomized orderings of all images per dataset.

4.3 Evaluation Methods

Our goal is to establish whether crowd worker performance is influenced by the level of data familiarity. To do this, we first describe measures we applied to evaluate crowd workers’ efforts and the quality of their submitted results. Then, we discuss a significance test that we adopted to indicate the likelihood that observed differences in crowd performance for

different types of images arose by chance. Finally, we explain our methodology to learn whether a crowd worker’s effort relates to the quality of his/her completed work.

4.3.1 Characterizing Performance of Crowd Workers

We chose four metrics to characterize a crowd worker’s effort and the quality of the result. Our first three metrics have been discussed in previous literature: quality [37, 39, 44], time [13, 90], and number of user clicks [8, 79, 85]. We also introduce a metric that, to the best of our knowledge, has not yet been cited in published crowdsourcing segmentation papers: average drawing time per user click.

4.3.1.1 Drawing Quality

A rigorous methodology to measure quality is to compare crowd drawings against gold standard segmentations established by experts. To do this, we adopt the standard intersection over union (IoU) metric to compute the pixel level similarity of each crowd drawn segmentation against the gold standard segmentation (i.e., $\frac{|A \cap B|}{|A \cup B|}$ where A represents the set of pixels in the crowd drawing and B represents the set of pixels in the gold standard segmentation). We establish a gold standard segmentation for each of the 405 images used in our studies by computing the pixel level majority vote from the multiple expert annotations per image included in the two datasets.

4.3.1.2 Annotation Time

We quantify the amount of time a worker spent completing a drawing HIT using logged metadata shared in the AMT system. In particular, a logged value indicates for each completed HIT the lapsed time between when the crowd worker clicked the “Accept HIT” button through the time the worker clicked the “Submit HIT” button. We use this metadata for both the drawing and voting HITs.

4.3.1.3 Number of User Clicks

We quantify worker drawing effort also based on his/her number of clicks used to delineate the boundary of an object. We compute this value by counting the number of (x,y) image coordinates that make up the closed polygon recorded in the submitted LabelMe result file.

4.3.1.4 Average Drawing Time Per User Click

We additionally quantify worker drawing effort using a metric that accounts for object boundaries that have varying levels of complexity (e.g., a box versus a tree). In particular, we compute for each crowd drawing the average time per user click ($\frac{\text{DrawingTime}}{\text{NumberOfUserClicks}}$).

4.3.2 Measuring Significance of Observed Results

To motivate which observed differences in crowd behavior are related to underlying changes in the crowdsourcing system set-up, we perform significance testing. For instance, does presenting images upright versus flipped upside down trigger significant changes to the time to draw and the quality of resulting drawings from crowd workers? Or are observed differences in drawing results due to the natural variability one would expect from humans performing the drawing task?

Inspired by previous work [84], we chose a paired sample t -test to learn whether observed differences in crowd behavior are likely to arise by chance. Our null hypothesis is that observed differences in crowd behavior for two sets of results are due to inherent noise in our crowdsourcing study, such as from the drawing, voting and evaluation processes. In other words, pairwise differences in crowd behavior for the two sets of results is a normal distribution with zero mean. The significance test returns a p -value which indicates the probability of obtaining the two sets of observed results when the null hypothesis is true. We reject the null hypothesis when the computed two-sided p -value is less than 0.05. Rejecting the null hypothesis means that, with high probability, observed differences are

reflective of a true difference between the two sets of results and so the two crowdsourcing systems.

4.3.3 Correlating Worker Effort to Quality of Work

To motivate whether crowd worker effort is related to the quality of his/her work, we train a predictive model and report its predictive power. Specifically, for a given image and crowd drawing, we are interested in learning whether computed descriptors of time, number of user clicks, and drawing time per click are indicative of the computed IoU score.

4.3.3.1 Predictive Model

We chose a multiple linear regression model to analyze the relationship between a crowd worker’s effort and the quality of a resulting drawing. This model leads to easy-to-interpret linear systems, as the resulting learned prediction system is a weighted linear combination of all chosen predictor values. Generically, this model is represented as follows:

$$y = X\beta + e \tag{4.1}$$

where y denotes a column vector of segmentation quality scores, X denotes a matrix of all observed predictive feature vectors describing all segmentations, β denotes a column vector of model parameters to be learned, and e denotes the vector of random errors between y and predicted values $X\beta$. The regression model is learned by finding the model parameters β that minimize the sum of the squared prediction errors (e). We trained our models using the freely-shared data mining software Weka [38].

4.3.3.2 Model Evaluation

We evaluate the predictive power of our learned model using Pearson’s correlation coefficient (CC). This measure indicates how strongly correlated predicted IoU scores are to observed IoU scores. To collect predictions, we perform 10 fold cross validation. Specifi-

cally, we randomly partition a set of crowd results into 10 independent sets approximately equal in size. For each of 10 iterations, a different set is reserved as the test set and the combination of the remaining sets are the training set. We then use the combination of predictions on the test sets from the 10 iterations. Resulting CC values range between +1 and -1 inclusive, with values further from 0 indicating stronger predictive power of a model.

4.4 Crowdsourcing Studies

We conducted three studies to examine the influence of the familiarity of data on crowd workers. We examined (1) how does data familiarity influence the quality of crowd drawings?, (2) how does making familiar data less recognizable influence crowd work with respect to the quality of results, annotation time, and annotation detail?, and (3) how does data familiarity influence the quality of crowd perception of the drawing task? We accepted as participants in our studies all crowd workers from AMT that had previously completed at least 100 HITs and received at least a 92% approval rating. We accepted all HITs submitted by all crowd workers.

4.4.1 Study 1: Drawing on Everyday and Biomedical Images

We first conducted a study to compare drawing results on familiar and unfamiliar images. We wanted to directly examine the importance of different findings [8, 37] regarding the quality of crowd work for the two types of image content.

4.4.1.1 Experimental Design

For each dataset, we collected five crowd drawings per image. We allotted crowd workers a maximum of ten minutes to complete each HIT and paid \$0.02 per HIT. We evaluated each crowd drawing against the gold standard segmentation. We chose the standard IoU metric to measure pixel level similarity of each crowd drawn segmentation to the gold standard

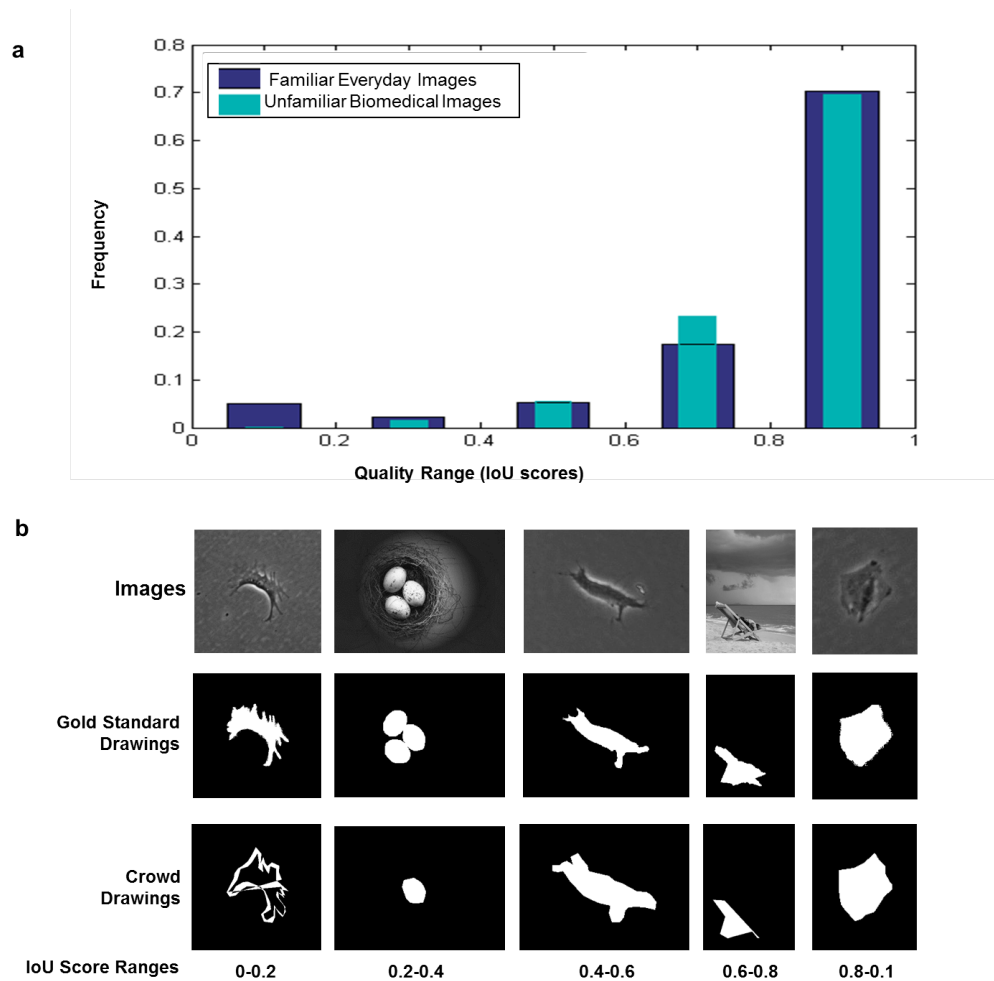


Figure 4.3: Comparison of crowd drawings on unfamiliar biomedical and familiar everyday images. We include both the (a) histogram of IoU scores for each dataset and (b) exemplar drawing results from each histogram bin.

segmentation.

4.4.1.2 Results

Overall, we found that crowd workers created drawings that closely resembled the gold standard drawings for the studied familiar everyday and unfamiliar biomedical images (Figure 4.3a). Following previous work [37], which demonstrates that experts commonly create drawings with scores above 0.6, we observed that approximately 90% of crowd drawings had IoU scores above 0.6. The quality of results above 0.8 (near perfect) is 70%.

In general, the trend looks like a decrease by a factor of two for the percentage of drawings that is in each lower quality bin. The quality of drawings associated with each histogram bin is exemplified in **Figure 4.3b**.

Intriguingly, there tended to be fewer drawings matching the quality of expert drawings for everyday images than biomedical images (**Figure 4.3a**). This difference is evident when comparing the percentage of drawing results with IoU scores from 0 to 0.2 (poorer quality drawings) and from 0.6 to 0.8 (higher quality drawings) for the familiar and unfamiliar datasets. We visually inspected crowd work which had IoU scores between 0 and 0.2. Causes were primarily due to confusions regarding the task aim including the appropriate object to annotate or the appropriate methodology for how to annotate the object (**Figure 4.3b**).

Our findings are based on the work of 83 unique workers. 44 unique workers created the 500 drawings for the familiar everyday images and 40 unique workers created the 1,525 drawings for the unfamiliar biomedical images.

4.4.2 Study 2: Drawing on Images Flipped Upside Down

Our motivation was to use knowledge about how crowd workers draw on upright and flipped images to learn the influence of content familiarity. This experiment is inspired as a compromise between crowdsourcing drawings on unfamiliar and familiar content such as biomedical and everyday images.

4.4.2.1 Experimental Design

We collected a total of 10 crowd drawings per image for the 100 familiar everyday images. We used the five drawings per image collected for the previous study. We also collected five drawings when each image was flipped upside down using the same crowdsourcing set-up as in the previous study.

We then computed four metrics to characterize effort and quality of drawings from crowd workers for each drawing: IoU score, time, number of user clicks, and average

drawing time per user click. We used the 1,000 computed scores for each of the four measures as the foundation for subsequent analyses.

We next performed significance testing to measure whether observed differences in crowd performance for upright and flipped images were significant. We performed four tests to compare the 500 upright and flipped image crowdsourcing results with respect to each of the four computed metrics.

Finally, we evaluated the relationship between the quality of the drawing and each of the three remaining computed metrics as well as the combination of the three metrics. As a result, we evaluated a total of eight prediction models. We learned four models from the 500 crowd results when images were upright and four models from when the 500 crowd results were images were flipped upside down.

4.4.2.2 Results

We found that crowd workers produced higher quality drawings when images were flipped upside down rather than upright (**Figure 4.4a**). The difference in average IoU scores across all crowd drawings was 4.8 percentage points, with average quality for upright images at 0.785 and upside down images at 0.833. Moreover, we found this quality difference was statistically significant. We found that differences in crowd behavior was predominantly isolated to instances where the crowd created “poor quality” drawings. As observed in **Figure 4.4a**, when comparing the two distributions, the delineation for outliers is 10% better and the 75th percentile score is 5% better while median scores and top 25th percentile scores are similar.

We were surprised to find that crowd workers exerted *less* effort to create the higher quality drawings on the flipped images than the lower quality drawings on the upright images (**Figure 4.4b,c**). Crowd workers took 16% less time with an average of 73 seconds for upright images and 61 seconds for flipped images. Crowd workers marked 7% fewer points to create each drawing for upright images in comparison to when images were flipped upside down (i.e., 33.9 and 31.4 number of user clicks respectively). We found that both

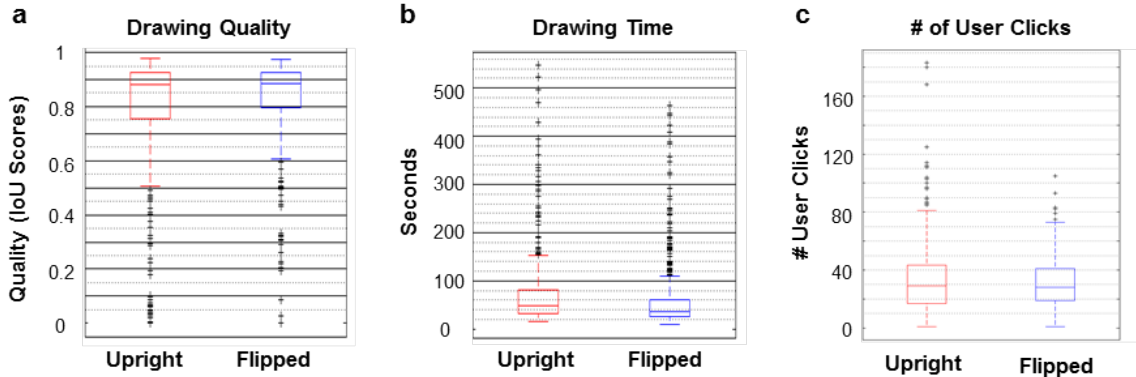


Figure 4.4: Analysis of 1,000 crowdsourced annotations collected on 100 everyday images where five crowdsourced annotations were collected per image when it was upright as well as flipped upside down. For each plot, the central marks of the boxes denote the median values, box edges denote the 25th and 75th percentiles values, whiskers denote the adjacent value to the data point that is greater than one and a half times the size of the inter-quartile range, and black cross-hairs denote outliers. When images were flipped upside down, overall, (a) segmentation quality was higher, (b) crowd workers took less time to annotate, and (c) crowd workers denoted the boundary of objects with more user clicks.

of these effort differences were significant. Overall, flipping images upside down led to a more consistent crowd behavior in terms of drawing time and number of user clicks (i.e., **Figure 4.4b,c**; smaller ranges between all values in each box plot, excluding the outliers).

In contrast, we did not observe a significant difference in terms of the average time a crowd worker took to draw per point when drawing on upright versus flipped images. Crowd workers took on average 3.2 and 3.6 seconds to mark each point for flipped and upright images respectively.

Linear regression analysis provided evidence for how crowd effort relates to segmentation quality (**Table 4.1**). We found from our single variable analyses that, whether crowd workers drew on images that were upright or flipped, segmentation quality was most correlated with the drawing time per point, followed by number of user clicks, and finally drawing time (**Table 4.1, rows 1-3**). Interestingly, segmentation quality tended to be better when a crowd worker took less time to draw each point (**Table 4.1, row 3**). We found that the strongest indication of a higher quality segmentation is when a worker takes

Table 4.1: How is worker effort correlated to the quality of a segmentation when images are upright and flipped upsides down. We evaluate worker effort with respect to four criteria: drawing time, number of user clicks, drawing time per click, and the combination of the three parameters. Segmentation quality is measured as the similarity of a crowd drawing to a gold standard drawing using the IoU evaluation metric. We report the learned linear regression model describing the correlation between worker effort and segmentation quality. We also report the correlation coefficient (i.e., CC), with larger absolute scores indicating greater linear correlation between worker effort and segmentation quality.

Upright Images	Model: IoU =	CC
Time (T)	$0.0002T + 0.7767$	0.03
# User Clicks (C)	$0.0022C + 0.7173$	0.23
Time/Click (TPC)	$-0.0121TPC + 0.8362$	0.33
All	$0.0007T + 0.0006C - 0.0146TPC + 0.7726$	0.40
Flipped Images	Model: IoU =	CC
Time (T)	$-0.0001T + 0.8368$	-0.21
# User Clicks (C)	$0.0024C + 0.7592$	0.22
Time/Click (TPC)	$-0.0052TPC + 0.8498$	0.33
All	$0.0002T + 0.0016C - 0.0051TPC + 0.7847$	0.34

less time to mark each point while drawing more points and taking more time (**Table 4.1, row 4**). Our learned models also reveal that there are slight differences regarding how crowd effort relates to segmentation quality when images are upright versus flipped. For example, segmentation quality reduces over twice as fast when users take more time per point for upright rather than flipped images (**Table 4.1, row 3**).

Our findings are based on the work of 75 unique workers. 44 unique workers created the 500 drawings for the upright images and 34 unique workers created the 500 drawings for the flipped images.

4.4.3 Study 3: Influence of Data Familiarity for Voting Task

We finally conducted a study to learn whether the prevalence of images that were indeed difficult to annotate, as exemplified by egregious drawing disagreements, could be detected as difficult image drawing problems through crowd voting.

4.4.3.1 Experimental Design

For the 100 familiar everyday images and 271 unfamiliar biomedical images (e.g., BU-BIL:1-5), we collected five crowd votes per image and then assigned image labels using the majority vote result. We allotted crowd workers a maximum of two minutes to complete each HIT and paid \$0.02 per HIT. We tallied the number of images that were labeled by majority vote as having a clear object to annotate. We also tallied the time crowd workers took to complete each HIT.

4.4.3.2 Results

Our findings suggest that crowd workers perception of the clarity of a task is influenced by the image content. From the studied 405 images, we found that the crowd perceived the drawing task as more difficult for unfamiliar than familiar image content by a difference of 30 percentage points (**Table 4.2**). The finding that 19% of the familiar image content data was tagged as a difficult drawing problem reflected our findings observed in the first study showing the frequency at which crowd workers created non-expert quality results (e.g., IoU scores < 0.6). However, the finding that 49% of the unfamiliar image content data was tagged as a difficult drawing problem inaccurately reflected our findings observed in the first study. Interestingly, crowd workers take more time on average to make an estimate of the task difficulty for familiar everyday image content than the unfamiliar biomedical image content.

Our findings are based on the work of 25 unique workers who contributed to the 100 voting HITs for the familiar everyday images and 26 unique workers who contributed to the 275 voting HITs for the unfamiliar biomedical images.

Table 4.2: Percentage of images in two datasets that are perceived by the crowd to have a clear segmentation task in the absence of additional instruction.

	# Images	Average Voting Time
Familiar Content	81%	23.4
Unfamiliar Content	51%	20.2

4.5 Discussion

Our results offer promising evidence that recognition of content is an important factor influencing one’s ability to perform a task. We found familiarity of data can detract from the quality of crowd work for open-ended tasks and can be beneficial for collecting higher-quality crowd work for closed-question tasks. In summary, different crowdsourcing tasks are better designed for differing levels of domain knowledge.

Crowdsourcing the Drawing Task. By broadening our analyses of crowd work to include familiar and unfamiliar data, we were inspired to rethink our generally held assumptions about the drawing task design for familiar image content. We suggest when crowdsourcing the drawing task to perform a simple step of flipping images upside down in order to gain great savings during run-time in terms of time and cost. For instance, our observed results at the scale of 100,000 images would translate to an aggregated savings equivalent to over eight forty-hour work weeks by a single person who draws boundaries of objects in images five days per week. Our observed results at the scale of 100,000 images would also translate to approximately 5,000 fewer poor quality results (i.e., IoU score < 0.4) by flipping images upside.

Interestingly, a similarly puzzling finding that flipping images upside down leads to higher quality drawing results has been discovered in the art community. In the New York Times bestselling book “Drawing on the Right Side of the Brain,” beginners are taught how to draw with a pencil on a blank page by flipping images they are trying to replicate upside down so the content is less familiar [21]. Possible future research could compare crowd behavior with respect to drawing on images versus drawing on blank canvases while trying to replicate content in observed images.

Influence of Data Familiarity on Crowd Worker Performance. When crowdsourcing tasks, we found people’s skills and judgments were clouded both for the better and worse by the content type. Workers that performed the open-ended task of drawing the boundary of an object in an image created more drawings that resembled drawings created by experts

when the image content was unfamiliar than familiar. In contrast, workers perceived the drawing task to be more difficult when the content was unfamiliar than familiar. Our results highlight the interesting question of why is popular belief contrasting what is observed in practice regarding the difficulty of a task?

We hypothesize that crowd workers' perceptions of task difficulty relates to their familiarity with the content. When workers are less comfortable with the content, they may be more inclined to perceive that other related tasks are more difficult. In general, it may be desirable to use fast, multiple choice questions to reduce the number of time-consuming, expensive open-ended drawing tasks. However, we infer that this step is best-suited for image content familiar to the crowd.

In contrast, we hypothesize that content familiarity leads to greater task ambiguity and analysis paralysis for crowd workers who perform the open-ended drawing task. For example, when annotating an image of a person, crowd workers that are "experts" on the content may be focused on asking whether they should annotate just the face or also include the person's body hallucinated underneath the clothing. In contrast, when annotating an image of a cell, crowd workers that are "not experts" on the content are less likely to be familiar and so distracted by the intricacies of the nucleus, membrane, and other internal structures that they could annotate within a cell.

Our results demonstrate the value of developing strategies to challenge and learn one's own biases when designing crowdsourcing systems. Analogous experiments in other domains could include examining how crowd workers perform in text-based or audio-based tasks that are in English when their first language is or is not English. Additional related experiments could include investigating how a crowd worker's behavior changes over time as he/she becomes more experienced and perhaps begins to "see" potential task ambiguities.

Open-Ended Tasks. While the crowdsourcing literature is filled with warnings about trusting crowd workers, we found crowd workers were generally highly trustworthy when they spent on average 61 seconds to complete an image drawing task. Our experiments comparing crowd drawings on upright versus flipped images allayed our initial concerns

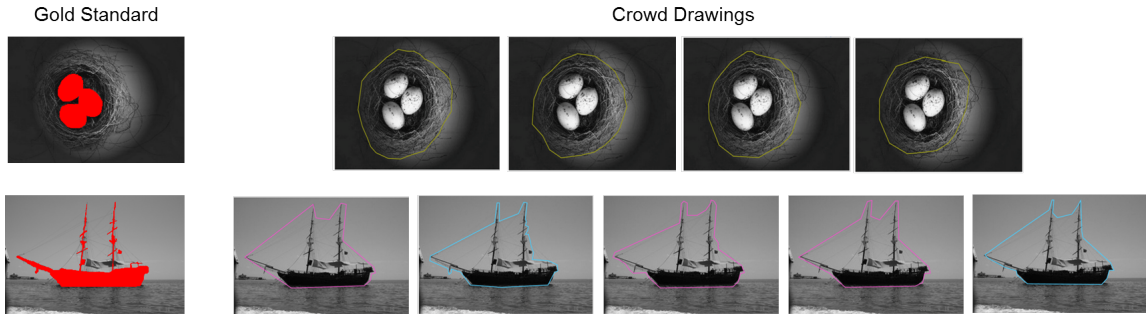


Figure 4.5: Why is crowd work still poor when images are flipped upside down? From visual inspection, we observed that most drawing outliers arose due to the majority of the crowd disagreeing with the gold standard drawing regarding the object of interest. We show a couple of exemplar results presented upright for visualization purposes.

that the greater percentage of poor quality annotations for everyday image content than biomedical image content is due to unreliable crowd workers. In particular, initially, we hypothesized that quality differences for the everyday and biomedical content may be due to crowd morale or boredom. However, we were surprised to observe from the image flipping experiments that workers were exerting more effort in terms of time and number of user clicks for the set of images on which they were producing lower quality annotations. We believe that crowd workers expertise in the content may have caused them to identify the shades of gray for various ways to interpret the drawing task which, in turn, led to divergent opinions and, possibly, analysis paralysis.

While the reliability of crowd workers depends on many factors, we offer our study of the image segmentation problem as a meaningful example for how to uncover possible causes for inefficiencies in leveraging crowd workers. We hope this study encourages others to consider crowdsourcing open-ended tasks who may have been previously deterred.

Drawing Outliers. From visual inspection of drawing outliers on the flipped familiar images, we observed that most poor quality scores arose because crowd workers consistently disagreed with experts regarding the true delineation of the object of interest (**Figure 4.5**). This result highlights an important concern regarding how experts are establishing gold standards. What should the truth be when the majority of the crowd disagree with experts?

In short, another commonly adopted bias when designing crowdsourcing systems is to trust experts over the crowd.

4.6 Conclusions

Our work facilitates the determination of what to expect from crowd workers for drawing and perception tasks on familiar and unfamiliar image content. Experiments on two challenging image sets revealed that crowd workers produced higher quality drawings when content was less familiar. We also found that crowd workers exerted less effort when they produced higher quality results. We recommend flipping images upside down when crowdsourcing the drawing task to gain benefits both in terms of higher quality results and faster collection. We hope our drawing studies will encourage rethinking generally held assumptions that one should expect large fractions of “poor quality” work when crowdsourcing open-ended tasks. Our studies offer promising evidence that researchers can improve designs of crowdsourcing systems by explicitly studying the influence of content familiarity on crowd behavior.

Chapter 5

Hybrid System - Drawing with Quality Control

Crowdsourcing is emerging in many data-rich fields as a promising supplement or substitute for performing data analysis tasks that are too labor intensive for expert practitioners to do themselves, and for which it is unclear which, if any, algorithm will yield accurate results [3, 49, 51, 72, 92, 96]. We address the question “when are efforts from algorithms and crowdsourcing suitable for delineating boundaries of objects in images (segmentation)?” We examined how to bring together the two disparate developments of drawing and quality control methods from the computer vision and crowdsourcing communities into a single framework. This chapter describes the following four key contributions, also discussed in our 2015 paper [34]:

1. Crowd voting as a quality-control step for the image segmentation task. To the best of our knowledge, we are the first to formulate this solution. Our contribution exposes interesting areas for future work, including designing optimal user interfaces to direct human attention to an image region while preserving surrounding image context.
2. Studies evaluating results obtained from crowd workers with respect to skill level and time. They highlight what to expect when leveraging crowd workers for both familiar (everyday images) and unfamiliar (biomedical images) content for drawing and voting tasks.
3. Evaluation and comparison of four implementations of a segmentation workflow based on different combinations of efforts from crowdsourced lay people and computer vision

algorithms for the drawing and voting steps. Analyses demonstrate the benefit of crowdsourcing and computer vision methods for the different steps in a system design.

4. Results revealing that hybrid algorithm-crowdsourcing system designs creates object boundaries that exceed the performance of pure crowdsourcing and algorithm systems; overcome well-known problems of crowdsourcing outliers and algorithm inconsistencies; and can produce segmentations that are of comparable quality (statistically similar) to those created by experts.

The remainder of the chapter is organized in six sections. A series of five formative studies is described in Section 2. These experiments informed us about how to involve crowd workers and build crowdsourcing systems to collect accurate segmentations. They motivated the design of a framework for human computation systems we call “Segmentation Annotation collection, Vote Collection, and Evaluation,” or SAVE, which is described in Section 3. Prototype implementations of this framework are also described in Section 3. A crowdsourcing experiment with SAVE is then described in Section 4; results and analysis are provided in Section 5. Discussions and conclusions follow in Sections 6 and 7.

5.1 Formative Studies

Five formative studies, F1 – F5, motivated how to involve humans for both creating and voting for “best” segmentations. We evaluated prototype tools with experienced and motivated volunteer participants in order to examine human performance when there are no serious concerns about human skill or intentions. The prototype tools are described in Section 5.2.2. For all images analyzed in our formative studies, one object of interest dominates each image, making it clear which object to outline to collect the desired segmentation. For the drawing task, these studies examine user interaction methods that can lead to higher quality human-drawn annotations and whether human-drawn or algorithm-drawn annotation options lead to the most accurate results. For the voting task, these studies investigate how to design the task including the implications behind human disagreement

when voting for a “best” segmentation.

5.1.1 Study F1: Expert-Drawn Segmentations on Black and White Images

We observed expert behavior on outlining objects from black-and-white images [50] to examine how accurately and with which methods an expert creates annotations when there are no perceptual questions of which pixels are part of the object versus background (**Figure 1.3a; MPEG7**). The study provides insight of what one may expect from a crowd worker at best since it reveals what one may expect from a highly qualified expert who dedicates large amounts of time and painstaking attention to the task of accurately capturing object boundaries when there are no challenges associated with annotating real world images, including background noise and ill-defined boundaries separating an object from the background. The study also provides insight into possible causes of drawing error.

Annotation Collection. We instructed a professional medical graphics illustrator to draw the object boundary for as many images as possible in one hour. We also informed the illustrator that we were collecting these annotations to compare the quality against algorithm-drawn boundaries. The illustrator created the segmentations using a touchpad with pen in the software Adobe Photoshop.

Results. The illustrator annotated three images. To evaluate segmentation quality, we computed scores that indicate how closely all generated segmentation annotations match the true object boundaries, using the intersection over union (IoU) score (defined in Section 5.2.6). The IoU score for the expert-drawn annotations of images showing an apple, cup, and fish was 0.77, 0.99, and 0.99 respectively. The nearly flawless segmentation annotations for the second and third images were due to the illustrator changing the annotation methodology. The illustrator used a paintbrush tool for the first annotated image and so possible causes of error included the thickness of the drawing tool and hand jitter. For the following two images, the illustrator zoomed in on the image to observe the individual pixels and then marked each pixel along the boundary.

5.1.2 Studies F2 and F3: One Expert Votes for Best Segmentation

We next examined whether domain experts prefer expert-drawn or algorithm-drawn object boundaries for biomedical images. The image library used in formative study F2 contains a sequence of 100 phase contrast microscopy images of a migrating fibroblast population (a large group of cells), and so enables us to study the challenge of annotating images for a single biomedical research experiment. The image library used in formative study F3 represents a more diverse collection of phase contrast microscopy images to observe if trends from formative study F2 generalizes. This image library contains 125 images showing 267 cells and was compiled by a domain expert to represent the variability of fibroblast appearances when captured using various image acquisition parameters in different environmental conditions.

Annotation Collection. We collected one expert-drawn and ten algorithm-drawn annotations per object in both formative studies. We collected the 10 sets of algorithm-drawn segmentations per image using *Algorithms 1-10* (**Table 5.2**). In order to constrain the images to have one dominant object, before applying the algorithms, we cropped each image by using the expert-drawn segmentation to detect the object location and then grew its bounding box by 10 pixels on all sides. The expert-drawn annotations were created in formative study F2 by a biomedical engineering PhD student, Expert A, who spent eight hours to draw a total of 423 cell outlines. The expert-drawn annotations were created in formative study F3 by Expert B, a biomedical engineering PhD student who created the images for quantitative analyses. Both Expert A and B chose to annotate the images using ImageJ, a widely-used biomedical image analysis system, with a computer mouse.

Voting Collection. We then asked a scientist with a PhD in biomedical engineering, Expert C, to vote for the segmentation best representing each cell region from the 11 segmentations (10 algorithm-drawn and one expert-drawn) shown simultaneously for each image. Expert C interacted with our freely-available software SAGE [32] to perform voting. The order of segmentations presented by the user interface was randomized for each each

image to prevent voter bias by possibly learning the algorithmic or manual source of the segmentations.

Results. We tallied the voting results based on the number of Expert C’s votes for the expert-drawn versus algorithm-drawn segmentation options. In formative study F2, we were surprised to find that only approximately 1% of votes (i.e., 4/423) were for the segmentations drawn by Expert A. In formative study F3, we found approximately 43% of votes (i.e., 116/267) were for the segmentations drawn by Expert B.

5.1.3 Study F4: Multiple Experts Vote for Best Segmentation for Everyday Images

To address the concern that the preference for algorithm-drawn over expert-drawn segmentations is due to voting biases from a single expert and is limited to image content unfamiliar to lay people, we conducted a voting experiment on 70 everyday images from a publicly available image library [4]. Objects of interest include, for example, a swan and a tree trunk (**Figure 1.3; Weizmann**).

Annotation Collection. We compiled six segmentations per image. One segmentation was a manual annotation provided with the image library for algorithm evaluation. We created the other five segmentations using *Algorithms* 3, 6, 9-11 (**Table 5.2**).

Voting Collection. For each image, we asked five computer vision PhD students to vote on the segmentation best representing each object region from six segmentation options shown simultaneously. To vote, each voter interacted with our publicly available software SAGE [32], configured with the order of segmentations presented by the user interface randomized for each image. The software was also configured so that the order of the images was the same for all voters.

Results. The voters each spent between 15 to 30 minutes to complete voting. We tallied the voting results based on the number of votes for each of the segmentation sources. Consensus from all five voters occurred for 63% of everyday images (i.e., 44/70). Exactly four voters agreed on the best segmentation for 17% of images (i.e., 12/70), three voters

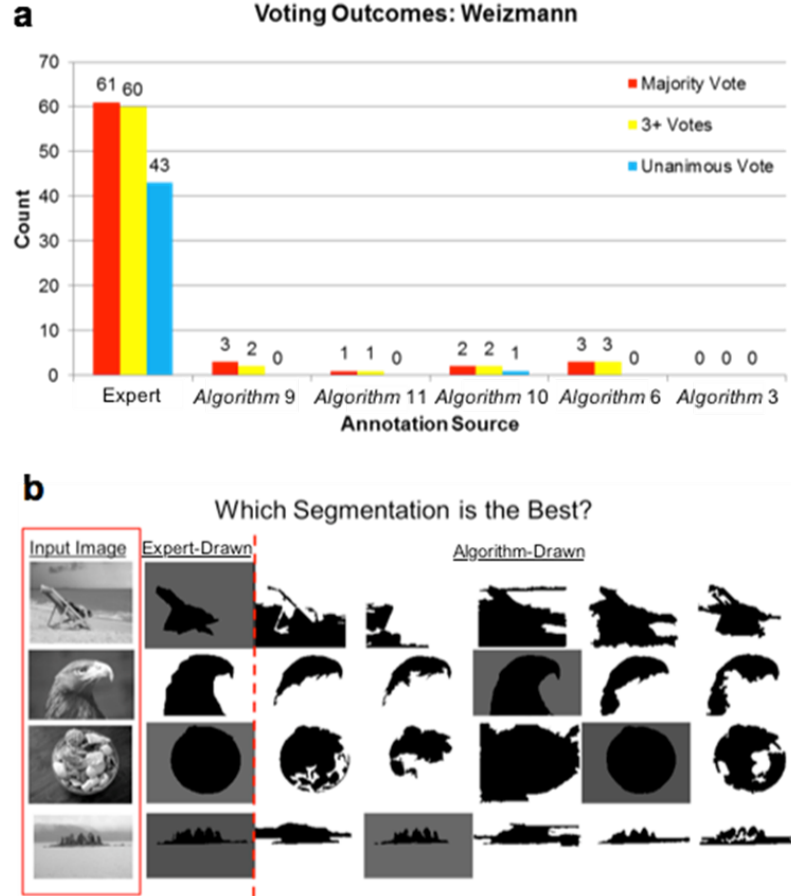


Figure 5.1: Summary of voting outcomes by experts for the best segmentation showing how frequently algorithm-drawn or expert-drawn annotations are preferred as well as the implications of segmentation quality based on the degree of voting agreement between five expert voters. (a) The number of voting outcomes for each of the six sets of annotations for 70 images in the Weizmann image library for three scenarios when considering only results with a majority vote (i.e., 2+ votes), 3+ voting agreements, and agreement between all 5 votes. (b) Each row shows the image followed by segmentations created by an expert and *Algorithms* 9, 11, 10, 6, 3 (Table 2). Darker gray level shadings indicate more votes for the segmentation. The voting outcomes are a unanimous win for the expert in row 1, a unanimous win for the algorithm in row 2, a majority vote win for the algorithm in row 3, and a majority vote win for the expert in row 4. For visualization purposes, segmentations are shown as binary images rather than image overlays.

for 17% of images (i.e., 12/70), and two voters for 3% of images (i.e., 2/70). There were no cases with five differing votes. The distribution of the number of “winning” sources arising from different amounts of voter agreement is shown in **Figure 5.1a**. We found that a small

number of algorithm-drawn annotations were preferred over expert-drawn annotations for each agreement level. Representative cases when expert-drawn or algorithm-drawn segmentations were preferred are shown in **Figure 5.1b**. From visual inspection of the results, we observed that voting disagreement arose when multiple segmentations appeared accurate as well as when multiple segmentations appeared flawed. In the latter case, voters assessed differently which flaws mattered more when identifying a “best” segmentation (**Figure 5.1b**, rows 3 & 4).

5.1.4 Study F5: Multiple Experts Vote for Best Segmentation for Biomedical Images

To address the concern that the preference for algorithm-drawn over expert-drawn segmentations in Studies F1 and F2 on biomedical images may be due to the unbalanced number of segmentation options for algorithms and experts, inadequate drawing performance by experts A and B, and voting biases of Expert C, we set up an experiment in which multiple experts voted on a balanced number of algorithm-drawn and expert-drawn options for images in BU-BIL:1-5 [37] (**Figure 1.3a; BU-BIL**). We conducted the experiment on 274 images that represent three biomedical imaging modalities: phase contrast microscopy, fluorescence microscopy, and magnetic resonance imaging.

Annotation Collection. We compiled six segmentations per image. Three options were the manual annotations provided in the image library [37]. The other three we created using *Algorithms* 2, 8, and 10 (**Table 5.2**).

Vote Collection. For each image, we asked three biomedical engineers (Experts C, D, and E - biomedical engineering PhD and PhD students) to vote on the segmentation best representing each object region from six segmentation options shown simultaneously. As in the previous study, each voter voted using our software SAGE [32] with all images presented in the same order.

Results. The three voters reported that voting took approximately 55 minutes, 2 hours, and 5 hours, respectively. For 18% of images (i.e., 50/274), there was consensus on the best

segmentation. For 59% of images (i.e., 161/274), exactly two voters agreed. There was no consensus for 23% of images (i.e., 63/274). We found that no single annotation source was preferred. Specifically, for the 211 instances of majority agreement, the distribution of “winners” coming from the six sets of annotations is 40% (i.e., 84 wins) for the first set of manual annotations, 32% (i.e., 67 wins) for the second, and 22% (i.e., 47 wins) for the third, and less than 1% (i.e., one win) for *Algorithm 2*, 5% (i.e., 10 wins) for *Algorithm 8*, and 1% (i.e., two wins) for *Algorithm 10*.

5.1.5 Lessons for Crowdsourcing Taken from the Five Formative Studies

Formative studies, F1 – F5, were conducted on 1,037 images that represent numerous object types observed with a variety of imaging modalities (visible, fluorescence microscopy, phase contrast microscopy, magnetic resonance imaging). We limited our study to images for which associated manual segmentations existed that had been created for use as the “gold standard” for algorithm evaluation. This allowed us to analyze the accuracy of segmentations that were established under the assumption that humans draw better outlines than algorithms. Additionally, we limited our study to domain expert voters to avoid concerns about results arising because voters have inadequate skill or even malicious intentions. We infer that the results reveal which observations about human involvement generalize to various image conditions. Several ideas of how to design the crowdsourcing annotation collection and voting collection tasks emerged from these studies (**Table 5.1**).

Annotation Collection. We infer that if highly qualified experts do not create perfect segmentation results, it is likely that less motivated humans using equivalent or less sophisticated annotation equipment will draw imperfect outlines. We observed that different annotators, experts and algorithms, were preferred for different images. Therefore, we infer that overall segmentation quality will improve when considering a collection of options rather than relying on a single algorithm or a single set of human annotations for all images. The results also showed that for generalized image sets it is more convenient to consider a collection of expert-drawn segmentations since algorithm-drawn segmentations

Table 5.1: Key properties of and results from the five formative studies that informed how to use humans in the loop and the final crowdsourcing system design. They motivate collecting multiple votes for the “best” segmentation from a collection of segmentation options.

ID	Objects	Annotators	# Votes	Annotation Results & Discussion	Voting Results & Discussion
F1	Black and white images [50]	1 expert	None	Imperfect annotations created even when experts dedicate lots of time; higher quality segmentations obtained with image zoom	Not applicable
F2	Biomedical images from one experiment	1 expert, 10 algorithms	1	Algorithm-drawn annotations preferred over expert-drawn annotations $\sim 99\%$ of time; expert takes ~ 68 seconds to draw each object boundary	Results may be biased due to a single voter
F3	Biomedical images from multiple experiments	1 expert, 10 algorithms	1	Algorithm-drawn annotations preferred over expert-drawn annotations for $\sim 57\%$ of images	Results may be biased due to unbalanced # of options from algorithms and experts as well as a single voter
F4	Everyday image library [4]	1 expert, 5 algorithms	5	Expert-drawn annotations preferred over algorithm-drawn annotations $\sim 87\%$ of voter consensus results	Voting agreement arose for all images; experts took on average 12 to 26 s to vote for each image
F5	Biomedical image library [37]	3 experts, 3 algorithms	3	Expert-drawn annotations preferred over algorithm-drawn annotations $\sim 94\%$ of voter consensus results; no single annotator preferred	Voting agreement arose for 77% of images; experts took on average 12 to 66 s to vote for each image

are rarely preferred however, for particular image sets, there can be great benefit when also considering algorithm-drawn segmentations or even only algorithm-drawn segmentations. We also infer, from observing expert behavior in study one, that a system design for human annotation should leverage image zoom to yield higher quality annotations.

Voting Collection. We found that experts have different priorities or interpretations when collapsing the possibly many observed imperfections or differences for each segmentation into a single assessment and then choosing the best option from each of these assessments. To make it easier for voters to agree on the segmentations that the experiment designer considers high quality, we suggest clearly motivating the criteria that will be used

to evaluate “quality” by incorporating into training or instructions how to handle the cases that are the common causes of undesired interpretations or voting disagreements. We also suggest collecting multiple votes for the “best” annotation. Finally, inspired by the restricted user interface in SAGE which presents annotation options in a vertical column (**Figure 5.3a**), we suggest examining an appropriate grid layout and number of options to present to the user in order to still yield accurate voting results while minimizing or eliminating user scrolling.

5.2 Methods

We propose a segmentation collection methodology that decomposes the task into a series of three micro-tasks that could be distributed and completed by any combination of humans and computers. We first describe this framework we call “**S**egmentation **A**nnotation collection, **V**ote Collection, and **E**valuation,” or SAVE. We then describe implementations of this framework used for our studies that support these tasks to be completed by experts, algorithms, and crowd workers (**Table 5.2**). Lessons learned from our formative studies motivated the choice and design of the crowdsourcing tools. We finally describe four system implementations of this framework that combine crowdsourcing and algorithm efforts in different ways to support studies to learn when to distribute the tasks to crowd workers or algorithms in practice.

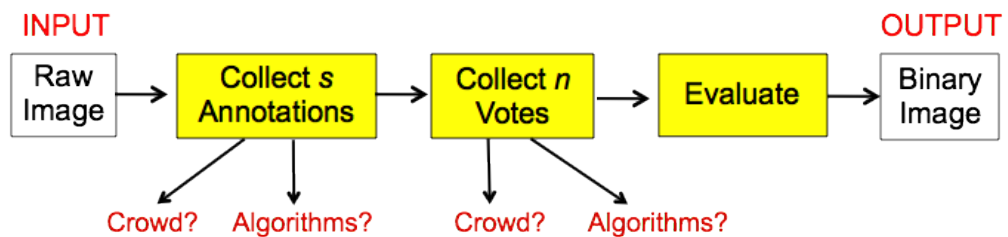


Figure 5.2: SAVE (Segmentation Annotation collection, Vote collection, and Evaluation), the proposed annotation collection methodology. A user collects s annotations, then collects n votes indicating which pixels/regions/annotations best reflect that of the true segmentation, and finally evaluates to establish a final segmentation.

5.2.1 SAVE Framework

SAVE takes as input an image to be annotated and outputs a single segmentation. SAVE involves performing a series of three steps for each image (**Figure 5.2**): (1) *Annotation Step*: The image is annotated by s algorithms or humans. (2) *Voting Step*: n votes at the pixel level or image level are collected from either humans or an algorithm to determine the “best” annotation from the s annotations (3) *Evaluation Step*: A decision mechanism interprets the votes to establish a final annotation to use. The key design decisions for implementing this pipeline are to determine (1) which annotation collection methods?, (2) which voters?, and (3) what annotation recommendation decision mechanism?

5.2.2 Annotation Collection Implementation

Expert-Based System: We utilized freely-available ImageJ [73] and Photoshop to collect expert segmentation drawings. ImageJ takes as input user specified points and connects them sequentially with straight lines to create the boundary of an object. Photoshop collects user brush strokes to produce a boundary or binary mask including all pixels in an object.

Algorithms: We compiled a comprehensive set of 11 image segmentation algorithms that together span four categories of algorithms commonly reported in the literature for biomedical images [64] (**Table 5.2**). The set consists of thresholding methods (algorithms 1 and 2 [68]), feature-based methods (algorithms 3 [6] and 4 [35]), region growing methods (algorithm 5 [91]), and deformable model based methods [10, 15, 16, 48, 53, 83]. We utilized freely-available implementations of each algorithm. We initialized *Algorithm 5* with two initial markers using the convex hull of the algorithm 3 segmentation for the background marker and the eroded segmentation from algorithm 3 as the foreground marker. We initialized *Algorithms 6-11* with *Algorithm 4* because of its reported success [35].

Crowd Worker System: To collect crowd-drawn segmentations, we set up the freely-available source code for the on-line image annotation tool LabelMe [79] in an Ubuntu

Table 5.2: List of segmentation drawing and voting tools used in studies for different annotator options. We built the crowdsourcing voting tool since no web-based tool exists for the segmentation problem and utilized existing tools/algorithms for the remaining five tasks.

	Annotation Collection Tool(s)	Voting Tool
Experts	Photoshop, ImageJ	SAGE
Algorithms	1: Adaptive thresholding 2: Otsu thresholding 3: Hough transform for circles 4: variance maps 5: seeded watershed 6: geodesic active contours 7: active contours without edges 8: localized region-based active contours 9: Bernard level set algorithm 10: Shi level set algorithm 11: Li level set algorithm	Pixel voting
Crowd Workers	LabelMe (web-based)	New tool (web-based)

computing environment on the Amazon Elastic Compute Cloud (EC2).

5.2.3 Vote Collection Implementation

Expert-Based System: We utilized our freely-available tool SAGE [32] to perform image level voting by experts (**Figure 5.3a**). This system shows an original image in a column on the left and each segmented region overlaid transparently on the original image in a vertical column on the right. To vote, the user selects a radio button next to the desired segmentation and then clicks a button to submit the vote. To prevent biases from segmentation ordering, the system randomizes the order of the set of segmentations presented by the user interface for each image.

Algorithm: We implemented an algorithm that performs pixel level voting to create a final segmentation. Specifically, the algorithm 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.

Crowd Worker System: Since existing web-based voting tools did not address challenges

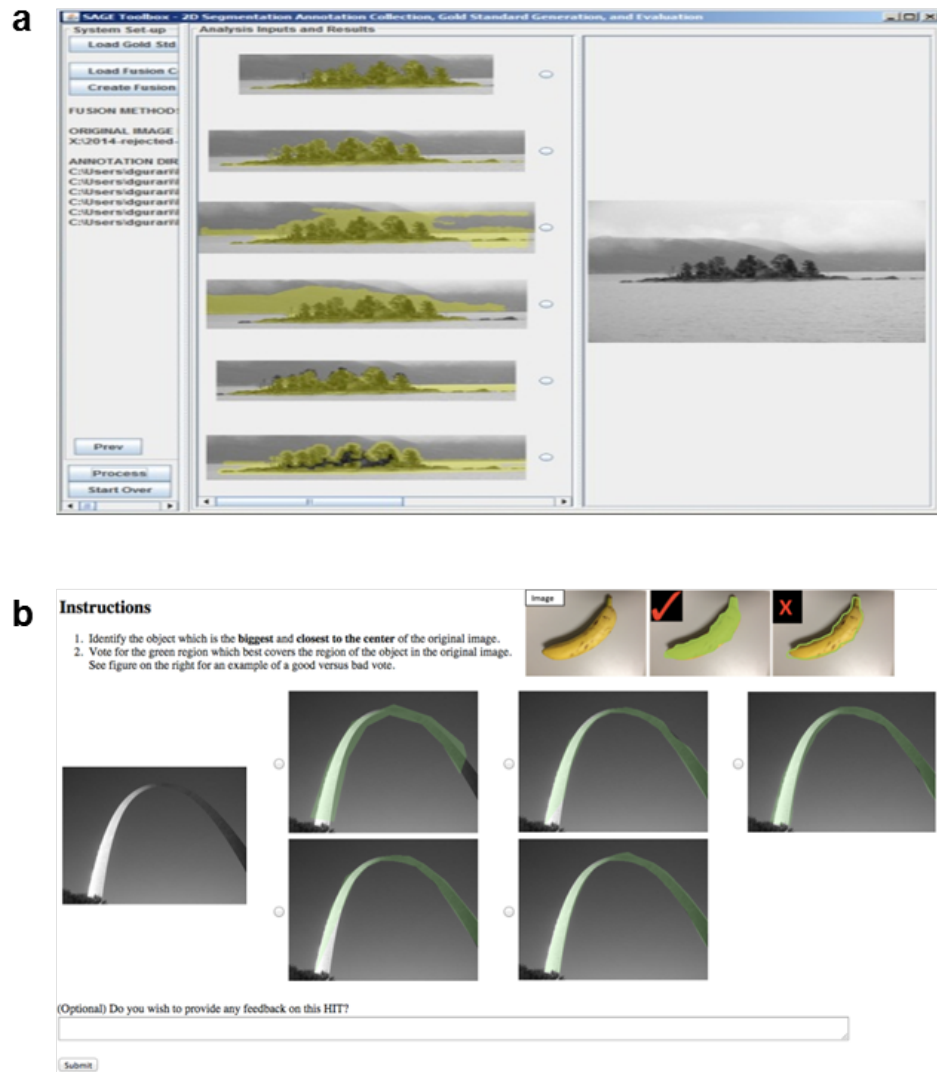


Figure 5.3: User interfaces for human-interaction voting tools that participants in the studies used to vote for a best segmentation among multiple options: (a) Freely-available SAGE system used by experts; (b) Web-based interface we created for use by internet workers. Key design choices were to present segmentation options with transparent green image overlays and to show all options in two rows to prevent user scrolling.

specific to the segmentation problem, we created a segmentation voting tool. One challenge is how to display each segmentation option. We overlay each segmentation option on the original image rather than presenting the segmentation as a binary image in order to encourage users to choose the option that is pixel perfect rather than semantically meaningful.

Selecting an overlay color/texture/transparency that works well in general for a variety of image characteristics (intensity contrasts, intensity textures, grayscale versus color images) is challenging. Preliminary studies motivate overlays used for our experiments, however a generalized solution is an open area for future research. Another challenge relevant to the image segmentation problem is choosing an appropriate grid layout. Motivated by observations from the formative studies, we designed the webpage to display all segmentation options in two columns and we scaled images to span the maximum width and/or height of the allotted grid cell in the web page to minimize user scrolling while supporting accurate voting. As a result, the webpage presents the original image on the left and a grid layout consisting of two rows that shows each segmented region overlaid transparently on the original image on the right (**Figure 5.3b**). In addition, to minimize the internet worker’s subjectivity in interpreting the criteria that he or she needs to optimize when voting, we also pasted short, step-by-step instructions at the top of the user interface with exemplar images to visually show undesired votes. To vote, a user selects a radio button next to the desired segmentation and then clicks a submit button. To prevent biases from segmentation ordering, the tool randomizes the order of the set of segmentations presented by the user interface for each image.

5.2.4 Evaluation Implementation

We separate “Evaluation” as its own task to enable one to make educated decisions regarding whether to trust or be “suspicious” of voting outcomes. For example, humans may not necessarily be trustworthy and so one may not want to trust all majority vote outcomes. One may want to instead weight the influence of different voters (human voting) or pixels (algorithm voting) differently. For each study in this paper, we specify what number of votes are used to determine the final segmentation and, when there is a tie, we select the first segmentation result that accrues the most votes.

5.2.5 Four SAVE Systems

We developed four implementations of SAVE that represents each of the four possible combinations of using crowdsourced workers and algorithms to perform the annotation collection and voting tasks. For each image, SAVE collects s segmentations of the image, collects n votes at the image or pixel level to establish the best segmentation, and saves the segmentation resulting from the majority vote. For annotation collection, the system supports using any combination of crowdsourced segmentations and *Algorithms* 1-11 (Table 5.2).

5.2.6 Quantitative Segmentation Performance Analysis

To evaluate segmentation quality, we used the intersection over union (IoU), a standard evaluation metric (i.e., $\frac{|A \cap B|}{|A \cup B|}$, where A represents the set of pixels in the true region and B represents the set of pixels in the annotated region). We then used significance testing to compare the quantitative results for different methods and establish if performance differences between methods are negligible for a collection images. In particular, we conducted a one-way analysis of variance (ANOVA), followed by a multiple comparison test with Tukey’s honestly significant difference criterion to perform pairwise comparisons of annotation performance. Statistically significant results are deemed those where the significance level p is less than 0.05.

5.3 Experiments

We conducted studies to evaluate and compare the four proposed SAVE implementations against standalone algorithmic and crowdsourcing methods. We examined (1) which among two pure crowdsourcing methods, two pure algorithmic methods, and two hybrid algorithm-crowdsourcing methods yield expert accuracy and perform the best?, (2) what are the benefits of using hybrid system designs over pure crowdsourcing and algorithmic methods?, (3) What are the benefits of using crowd workers versus algorithms for annotation and

voting respectively to solve the image segmentation problem?, (4) how do crowd workers perform for both the segmentation drawing and voting tasks with respect to skill level and time?, and (5) how does image content impact crowd worker behavior?

5.3.1 Image Libraries

We analyzed all segmentation methods on a total of 405 images from freely-available image libraries of everyday images [4] and biomedical images [32] that were also used in formative studies F3 and F4. We chose these image libraries because they each include multiple expert-drawn segmentations. For each image, we applied the *pixel majority vote algorithm* to fuse the multiple experts' annotations into a final gold standard segmentation to reduce the impact of biases and mistakes from a single expert on performance analyses.

5.3.2 Crowdsourcing Platform and Participants

We used the Amazon Mechanical Turk internet marketplace to recruit crowdsourced workers. We accepted all Mechanical Turk workers that had previously completed at least 100 HITs and received at least a 92% approval rating. For all HITs, we allotted a maximum of ten minutes to complete the task. We approved all submitted HITs.

Annotation Collection. We created external HITs by applying scripts provided with LabelMe [79] to post HITs to Mechanical Turk and record the submitted results in our cloud computing environment. When Workers reviewed the HITs, they were redirected to a webpage that contains a five step set of instructions followed by pictures exemplifying good and bad annotations. After accepting a drawing HIT, the Worker was presented the user interface to complete and submit the task (**Figure 3.2**). We paid each worker \$0.02 to complete the drawing task.

Vote Collection. We created internal HITs by adapting Mechanical Turk templates to both post our voting HITs as an embedded webpage and record the submitted results. When Workers reviewed the voting HITs, they were first shown a two step set of instructions with pictures exemplifying a good and bad vote followed by the original image on the

left and segmentation options on the right (**Figure 5.3b**). Motivated by preliminary crowdsourcing experiments, we use exemplar images that demonstrate that the task is to choose the segmentation which has the largest number of pixels overlapping the object of interest rather than selecting segmentations for which the object could be best recognized. Also motivated by preliminary analyses, we present each segmentation option for all images using a green overlay of the segmentation on the original image. We paid each worker \$0.01 to complete a voting task.

5.3.3 Performance Analysis for Four SAVE Implementations

We evaluated six segmentation options based on different combinations of efforts from crowdsourced workers and computer vision algorithms (**Table 5.3**). Two of these configurations are pure crowdsourcing methods: crowd-drawn segmentations (*C1*) and segmentations chosen by crowd voting on multiple crowd-drawn segmentations (*C2*). Another two of these configurations are pure automated methods: algorithm-drawn segmentations (*A1*) and segmentations created by algorithm voting on multiple algorithm-drawn segmentations (*A2*). The final two configurations are hybrid human-computer methods: segmentations created by algorithm voting on multiple crowd-drawn segmentations (*CA*) and segmentations chosen by crowd voting on multiple algorithm-drawn segmentations (*AC*).

We collected five annotations and five votes for the annotation collection and vote col-

Table 5.3: Description of seven annotation methods we evaluated and compared in the studies. Each annotation method is described in terms of the SAVE pipeline and a ranking shows how these methods compare across the studied 405 everyday and biomedical images with respect to median IoU scores.

ID	Method Type	Annotations Per Image	Votes	Rank
Ex	Expert	3 expert-drawn	None	1
C1	Crowd	5 crowd-drawn	None	4
C2	Crowd	5 crowd-drawn	5 crowd per image	3
A1	Algorithm	1 algorithm-drawn	None	7
A2	Algorithm	5 algorithm-drawn	5 pixels per pixel	6
CA	Hybrid	5 crowd-drawn	5 pixels per pixel	2
AC	Hybrid	5 algorithm-drawn	5 crowd per image	5

lection tasks respectively from both the crowd and algorithms. For crowd annotation collection, we posted for each image library five batches of HITs for all images simultaneously. For the algorithm annotation collection methods, we collected five sets of segmentations for each image using five different algorithms for the two image libraries. Motivated by demonstrated success of algorithms in the formative studies, we applied *Algorithms 3, 6, 9, 10, and 11* (**Table 5.2**) for the “Everyday Images” and *Algorithms 2, 3, 5, 8, and 10* (**Table 5.2**) for the “Biomedical Images.” For the crowd vote collection methods, we posted for each image library five batches of HITs for all images simultaneously. We ran this experiment twice using as input the five sets of crowd-drawn segmentations and algorithm-drawn segmentations respectively. For algorithm voting method, we fused five segmentations per image into a single segmentation that represents the pixel-level majority vote (e.g., assign a pixel as “object” only when at least three input segmentations assign the pixel as “object”). We ran algorithm voting twice, using as input the five sets of crowd-drawn segmentations and algorithm-drawn segmentations respectively.

To establish whether any of the six studied segmentation methods can compete with expert accuracy, we computed the IoU scores for three expert-drawn annotations per image for all 405 everyday and biomedical images (*Ex*). We then computed the IoU score for every segmentation produced by all pure crowd-based (*C1, C2*), pure algorithm-based (*A1, A2*), and hybrid crowd-algorithm based methods (*CA, AC*). Preliminary experiments motivated our algorithm selection for *A1* where we chose among the 11 algorithm options the algorithms that yielded the highest median score for each image library: *method 6* for the “Everyday Images” and *method 10* for the “Biomedical Images.” In total, 5,265 computed scores characterizing the seven annotation sources (experts + 2 standalone + 4 SAVE systems) served as the foundation for our subsequent analyses.

5.3.4 Characterizing Crowd Behavior

We conducted studies to highlight what to expect when leveraging crowd workers for both drawing and voting tasks for familiar (everyday images) and unfamiliar (biomedical images)

content.

Annotation Collection Task. We counted the number of unique workers that contributed to creating the 2,025 drawings and computed statistics to characterize the time they took to draw each object boundary. To identify how the quality of drawings from the crowd compares to drawings created by experts, we then compared IoU scores computed for the crowd-drawn segmentations and expert-drawn segmentations.

Voting Task. We counted the number of unique workers that contributed to creating the 4,050 votes and computed statistics to characterize the time they took for the voting task. We also computed the number of majority vote outcomes that led to a segmentation with an IoU score that was within 5% of the top-scoring segmentation option available.

5.3.5 Characterizing Successes of Image Segmentation Algorithms

We counted the number of majority voting outcomes for each of the five algorithm options summed over all studied images. We analyzed these results independently for the everyday and biomedical images.

5.4 Results

We analyzed segmentation results coming from a total of 6,075 HITs completed by a total of 208 unique workers and a total of 2,835 segmentations created by algorithms (stand-alone and majority-pixel vote).

5.4.1 Performance Analysis for Four SAVE Implementations

Top Performer(s). We found that the best segmentation option overall is a hybrid approach that applies algorithm voting to fuse crowdsourced drawings (*CA*), when evaluating by comparing the median scores of the segmentation options (**Figure 5.4; All Images**). Significance testing revealed that crowd voting on crowd drawings (*C2*) is a comparable top-performing option, when evaluating for “All Images.”

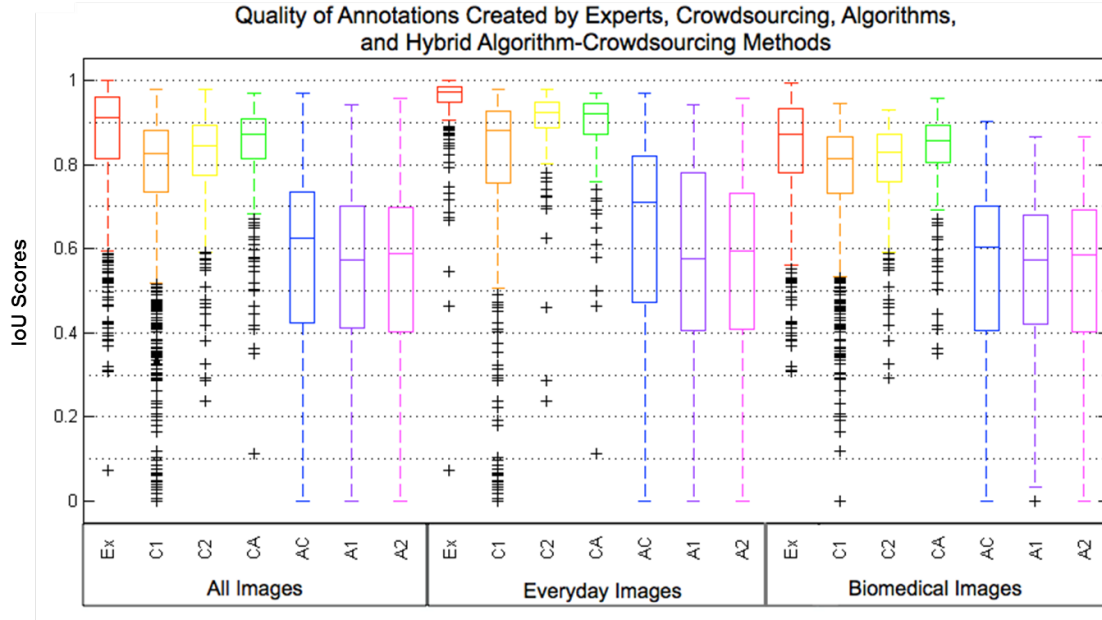


Figure 5.4: Summary of IoU scores shown for all 405 images, as well as based on the image content (i.e., everyday images only and biomedical images only), when applying six segmentations systems per each object and evaluating expert annotations: experts (red); two crowdsourcing methods - C1 (orange) and C2 (yellow); two hybrid algorithm-crowdsourcing methods - CA (green) and AC (blue); and two algorithm methods - A1 (indigo) and A2 (magenta). 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). We found that crowdsourcing approaches led to segmentations of comparable accuracy to that of experts and better accuracy than fully-automated methods. We observed that the best method for involving crowdsourced workers was a hybrid approach of algorithm voting to fuse crowdsourced drawings. This method yielded expert-quality annotations for the biomedical images. We observed that crowd-drawn annotations resulted in a larger percentage of egregious outliers for everyday images than biomedical images (i.e., IoU < 0.3).

Expert Equivalent Options. Significance testing revealed segmentations created by two of the SAVE implementations (*C2*, *CA*) are comparable to expert-drawn annotations (*Ex*) for the “Biomedical Images.” In contrast, significance testing revealed that none of the six studied methods performed comparably to experts for the “Everyday Images.”

SAVE vs Stand-Alone Options. The results highlight performance gains from applying the SAVE system implementations (**Figure 5.4**; *A2, C2, AC, CA*) in place of standalone annotation collection methods (**Figure 5.4**; *A1, C1*). We found that combining multiple crowd-drawn segmentations with crowd or algorithm voting improved overall quality while eliminating most of the egregious outliers (**Figure 5.4**; *A2* and *AC* vs *A1*). We found that combining algorithm-drawn segmentations with crowd or algorithm voting could lead to higher quality (i.e., median score) results than relying solely on the studied top performing computer vision algorithm (**Figure 5.4**; *A2* and *AC* vs *A1*).

Hybrid Algorithm-Crowdsourcing Systems. Overall, we found that hybrid methods yielded superior performance to pure automated and pure crowdsourcing methods (**Figure 5.4**; *All Images*). Specifically, significance testing demonstrated significant quality improvement when pairing algorithm-drawn annotations with crowd voting over the two studied pure algorithm-drawing methods (**Figure 5.4**; *All Images - AC* vs *A1* and *A2*) and pairing crowd-drawn annotations with algorithm voting over the pure crowd-drawing method (**Figure 5.4**; *All Images - CA* vs *C1*). Moreover, algorithm voting on crowd drawing performs better than crowd voting on crowd drawings, when comparing median scores (**Figure 5.4**; *CA* vs *C2*).

5.4.2 Characterizing Crowd Behavior

Annotation Collection Task. 90 unique crowd workers created the 2,025 drawings. The time that crowd workers spent on completing a HIT was on average 30 seconds (i.e., median time), 25th percentile and 75th percentile times ranged from approximately 20 to 55 seconds, and was at most close to three minutes (**Figure 5.5**; *All Images - C1*). We observed that the main distinguishing factor between crowd-drawn and expert-drawn annotations is that more egregious outliers are created by crowd workers than experts (**Figure 5.4**; *All Images - C1* vs *Ex*). We visually inspected the 45 crowd-drawings that received a score below 0.3 (score distinguishing where majority of outliers lie between the crowd and

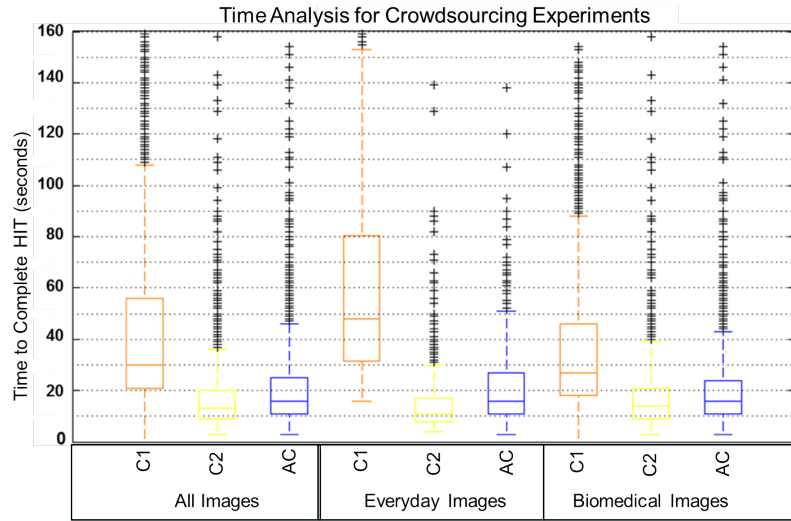


Figure 5.5: How much time do crowd workers spend on drawing and voting tasks? Summary of time taken for crowdsourced drawings (orange), crowdsourced voting for the best among five crowdsourced drawing options (yellow), and crowdsourced voting for the best among the five algorithm drawings (blue) is shown for all images as well as only for the “Everyday” and “Biomedical Images.” See Figure 5.4 for the explanation of a box plot visualization.

experts). We categorized these outliers into five types of observed crowd behaviors which are exemplified in **Figure 5.6**: 1) multiple closed contours drawn, 2) points marked on the image along with a text label indicating what the object is, 3) wrong object annotated, 4) object annotated at incorrect granularity, and 5) spam.

Voting Task. 129 unique crowd workers contributed to the 4,050 voting tasks. When voting between five crowd-drawn options, the crowd workers took on average (i.e., median time) 12 seconds, 25th percentile and 75th percentile times ranged from approximately 9 to 20 seconds, and took at most close to three minutes (**Figure 5.5**; *All Images - C2*). When voting between five algorithm-drawn options, the crowd workers took on average (i.e., median time) approximately 16 seconds, 25th percentile and 75th percentile times ranged from approximately 11 to 25 seconds, and took at most close to three minutes (**Figure 5.5**; *All Images - C2*). We found the voting outcome led to IoU scores within 5% difference



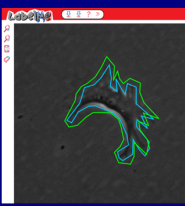
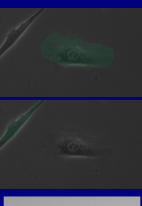


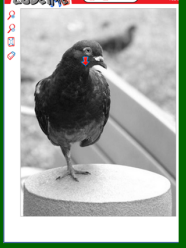
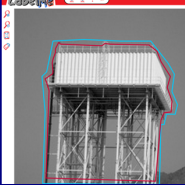


	<u>Spam</u>	<u>Point(s) with Labels (Recognition Task)</u>	<u>Multiple Closed Contours</u>	<u>Wrong Object</u>	<u>Ambiguous Object Granularity</u>
Categories of Crowd Drawing "Mistakes"					
					
Fix?	Filter Workers	Modify or Create New Web-Based Drawing Tool			Improve Instructions

Figure 5.6: Why do crowd workers make drawing mistakes? We categorized egregious drawing outliers (i.e., IoU scores below 0.3) into five categories which we show in five columns with representative crowd-created results. Two categories were only observed on “Everyday Images” (highlighted in dark green) and the other three categories were observed for all studied images (highlighted in navy blue). In the bottom row, we suggest ways that we hypothesize would prevent these observed outliers.

of the highest scoring segmentation option for 68% (i.e., 277/405) of voting outcomes on crowd-drawings ($C2$) and 57% (i.e., 229/405) of voting outcomes on algorithm-drawings (AC).

Impact of Image Content. We found differences between crowd behavior for different image content in terms of quality, time, and worker recruitment (**Figure 5.4**, **Figure 5.5**, **Table 5.4**).

For annotation collection, the total elapsed time from posting to submission of all HITs was proportionally about three times more for “Biomedical Images” (i.e., 1,955 minutes for 2,025 drawings) than for “Everyday Images” (i.e., 226 minutes for 500 drawings). In terms of quality, we found that the percentage of outliers with an overlap score below 0.3

Table 5.4: Statistics characterizing the six independently run crowdsourcing experiments in terms of task cost, elapsed time to collect all submitted tasks, and number of unique workers for all tasks.

	Everyday (100 images)			Biomedical (305 images)		
	Draw (C1)	Vote (AC)	Vote (C2)	Draw (C1)	Vote (AC)	Vote (C2)
US Dollars Paid Per HIT	\$0.02	\$0.01	\$0.01	\$0.02	\$0.01	\$0.01
Elapsed Time to Collect Five Sets of HITs (Normalized to min/500 HITs)	226	56	35	641	26	38
Unique Worker Count (Avg. # Tasks Per Worker)	44 (11)	36 (14)	13 (38)	40 (51)	44 (46)	45 (45)

accounted for approximately 6.6% of annotations (i.e., 33/500) observed on 22 unique “Everyday Images” and 0.6% of annotations (i.e., 12/2025) observed on 11 unique “Biomedical Images.” From visual inspection of these outliers, we learned that two of the five types of outliers were only observed for “Everyday Images”: six instances of identifying image point(s) with associated text labels to perform a “recognition” task and two instances of spam annotations.

For voting, we found that the majority vote agreement yields higher quality voting outcomes for everyday images than biomedical images. It led to selecting a segmentation with an IoU score within a 5% difference from the score of the top-performing segmentation option for 81% of outcomes for the everyday images (i.e., 87/100 crowd-drawn and 74/100 algorithm-drawn) and 49% of outcomes for the biomedical images (i.e., 190/305 crowd-drawn and 155/305 algorithm-drawn). This better voting performance for everyday images is also observed when comparing how the median score in practice compares against the maximum and minimum possible median scores that could arise if voters always chose either the highest scoring segmentation or lowest scoring segmentation respectively from the five segmentation options for each image (**Table 5.5**).

5.4.3 Characterizing Successes of Image Segmentation Algorithms

We found that no single algorithm was preferred and all algorithms were perceived as an optimal option for some number of images. We report the distribution of voted high

Table 5.5: How do majority vote outcomes from the crowdsourcing experiments compare to the best and worst possible voting outcomes? Median scores are shown for the crowdsourcing experiments as well as scenarios when voting outcomes identified the maximum and minimum possible scoring segmentation for all images.

	Crowd-Drawn			Algorithm-Drawn		
	Actual (C2)	Max	Min	Actual (AC)	Max	Min
Everyday Images	0.92	0.93	0.68	0.71	0.74	0.2
Biomedical Images	0.83	0.86	0.71	0.6	0.68	0.14

quality segmentation sources for both image sets based on majority vote winners. For the 305 biomedical images, we tallied that the winners were 29, 87, 56, 56, and 10 instances for *Algorithms* 2, 3, 5, 8, and 10 respectively. For the 100 everyday images, we tallied that the winners were 25, 23, 28, 19, and 5 instances for *Algorithms* 3, 6, 9, 10 and 11 respectively.

5.5 Discussion

Intersection of Computer Vision and Human Computation. We proposed SAVE, a modular image segmentation framework which users can plug in different annotation collection and voting quality control methods, algorithm-based and human-based. While the inherent tradeoffs of using crowdsourcing for accuracy and algorithms for efficiency are well known, the strengths of both approaches are shown to successfully emerge when combining their efforts in hybrid SAVE system implementations (*CA*, *AC*). Pairing existing crowd drawing tools with algorithm voting quickly and inexpensively filters out the occasional segmentation errors while improving the accuracy of results overall. Pairing existing algorithm drawing methods with crowd voting empowers users to quickly offload the task of identifying the best-suited algorithm among several options for different images at a low cost. These results provide exciting evidence that this simple framework is effective for a diversity of images to bring out the strengths from the disparate computer vision and human computation communities and efficiently create higher-quality segmentations.

Crowdsourced Voting. Our work reveals that crowd voting for the best segmentation can vastly improve the quality of segmentations. While we were surprised to observe that this

quality control method has never been applied before, we hypothesize the reason may be because this seemingly simple voting task for the “best” annotation comes with numerous open challenges unique to the image segmentation problem. Challenges include establishing the optimal method to display segmentation options (e.g., green overlays on the original image) and the optimal grid layout with number of segmentation options (e.g., in our web interface, two rows with fives options). Experiments with our proposed user interface design revealed that users can trust crowd workers to judge what is a best segmentation among several options for both everyday and biomedical images. Moreover, we found that users could expect a quick turn-around time for such experiments given that, for our four crowdsourced voting experiments, we collected 535-1,153 completed voting tasks per hour.

Crowd Behavior for Different Image Content. Surprisingly, we observed that unfamiliar (biomedical) images elicited fewer egregious drawing errors while, less surprisingly, crowd workers performed better for crowd voting on the familiar (everyday) images. The egregious drawing errors (IoU scores below 0.3) occurred approximately 11 times more frequently for everyday images than biomedical images. We observed that recognition of the image content led to more drawing mistakes because crowd workers would misinterpret the task as annotating objects at smaller granularities that they recognize rather than the largest “blob” in the center of the image (e.g., stone versus basket) or indicating the identity and location of observed objects.

Expert-Quality System. Our results demonstrate the potential of the proposed methodology, SAVE, to produce expert-quality annotations with crowdsourced workers and algorithms. The immediate practical importance of this finding is clear for biomedical image analysis studies when comparing results from the formative studies and SAVE study. Specifically, whereas the domain expert in study F2 spent eight hours to produce 423 outlines of biological structures in biomedical images, we have demonstrated that a domain expert can instead spend around \$30 and wait for approximately 31 hours to have a hybrid SAVE implementation (CA) collect 305 expert-quality outlines of biological structures in biomedical images.

Future Work. Many possible directions for future research emerge from our work. For human computation, interesting future research directions include examining what to expect from the crowd when using different incentive structures (e.g., gamification or citizen science) in terms of quality for both the crowd drawing and voting tasks for familiar and unfamiliar image content. Another open research question is how to improve instructions or worker training to improve crowd voting results for unfamiliar image content. For computer vision, an interesting future research direction would be to find the smallest collection of image segmentation algorithms that together yield at least one algorithm that will work well for a diverse set of images. Finally, at the intersection of human computation and computer vision lies an opportunity to grow crowd-based voting studies to learn on a large-scale which algorithms work best for which images. Our ultimate research goal is to use voting outcomes to train classifiers to automatically predict the best-suited algorithm based only on the image information. Success with this work will lead to a fully-automated system that works well for the diversity of images and so empower users to collect accurate segmentations at scale relatively quickly and inexpensively.

5.6 Conclusions

The proposed image segmentation framework SAVE allows mixing and matching of crowdsourcing and algorithm methods for segmentation creation and voting quality control in a single framework. Experiments with different combinations of algorithm and crowd efforts in the SAVE framework revealed that hybrid systems designs outperformed pure algorithmic and crowdsourcing approaches. We were excited to find that one of the hybrid algorithm-crowdsourcing system designs created expert-quality segmentations for biomedical images. This finding highlights that a new question may be realistically explored within the biomedical community of “What is possible if we could efficiently and inexpensively gather thousands of expert-quality segmented images?” We also found that, overall, crowd workers could be trusted to perform both the drawing and voting tasks for both familiar

and unfamiliar image content. Finally, we found that introducing the new crowd voting quality control method for the segmentation problem is a powerful starting point towards automatically learning on a large scale how to pair data with appropriate algorithms.

Chapter 6

Hybrid System - Human Initializes Algorithm

Level set methods are widely used to automate finding accurate boundaries of biological structures in biomedical images and videos. In general, level set methods deform an initial contour to a final contour that separates image foreground from background so that some method-specific image partition condition is enforced. While new energy functionals controlling how to partition an image continue to be proposed to address the spectrum of possible image conditions, the continued development and wide-spread sharing of new options is making it difficult for both non-experts and experts to know which method to use when. A further challenge for applying such methods is knowing which type of initial contour will be sufficiently close to the desired boundary since they often produce locally optimal segmentation results which may not match the desired globally optimal segmentations. As a result, a common question asked by individuals trying to apply level set methods is “which method with which initial contour will produce the desired boundary in my images?”

To address this question, we evaluated level set algorithms that currently have a potential widespread practical impact due to their inclusion in freely-shared bioimage analysis systems [20, 23, 80]. *Geodesic active contours* evolve the initial contour to end up in regions with strong edges (high contrast) [15]. *Active contours without edges* evolve the initial contour to try to separate the image into two homogeneous regions [16]. Both *Lankton region-based active contours* [48] and the *Li level set* algorithm [53] evolve 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 regions. The *Shi approximation method*

computationally speeds up the evolution process by replacing slow real-valued calculations with faster integer-based calculations [83]. The method by *Bernard et al.* uses a linear combination of B-spline basis functions for process speedup [10]. Currently, there is no work comparing these freely shared algorithms on biomedical image sets.

Domain experts planning to use level set methods on their biomedical images encounter an additional overhead of creating initial contours. With freely available image analysis software [20, 80], they create initial contours, clicking on images to create simple geometric shapes, points connected into polygons/splines, or free-hand tracings, and then typically wait for seconds or minutes per image for the input contour to finish evolving to a segmentation [20]. While recent as well as foundational publications reported that simple initial contour shapes such as bounding squares, rectangles, circles, ellipses, and triangles led to accurate segmentations [10, 15, 16, 48, 53, 83], other recent publications suggested these initial contours can be insufficient. As an example, specialized contour initialization methods have been proposed for phase contrast image sets [22, 54] to avoid common curve evolution failures. It can be faster for domain experts to manually trace boundaries themselves than to run an algorithm and possibly risk running it repeatedly until finding an initial contour that returns an accurate segmentation (**Figure 6.1**).

To provide practical guidance for obtaining accurate segmentations with level set methods, we conducted an extensive comparison study of six publicly-available level set methods paired with popular initial-contour shapes which we discussed in a 2014 publication [36]. We analyzed when to use which method and how to use the methods effectively on fluorescence and phase contrast images. To further minimize the overhead for domain experts of creating initial contours for their biomedical images, we also proposed to use crowdsourcing to create them. Finally, to facilitate extensions of this study, we publicly share all code (<http://www.cs.bu.edu/~betke/BiomedicalImageSegmentation>).

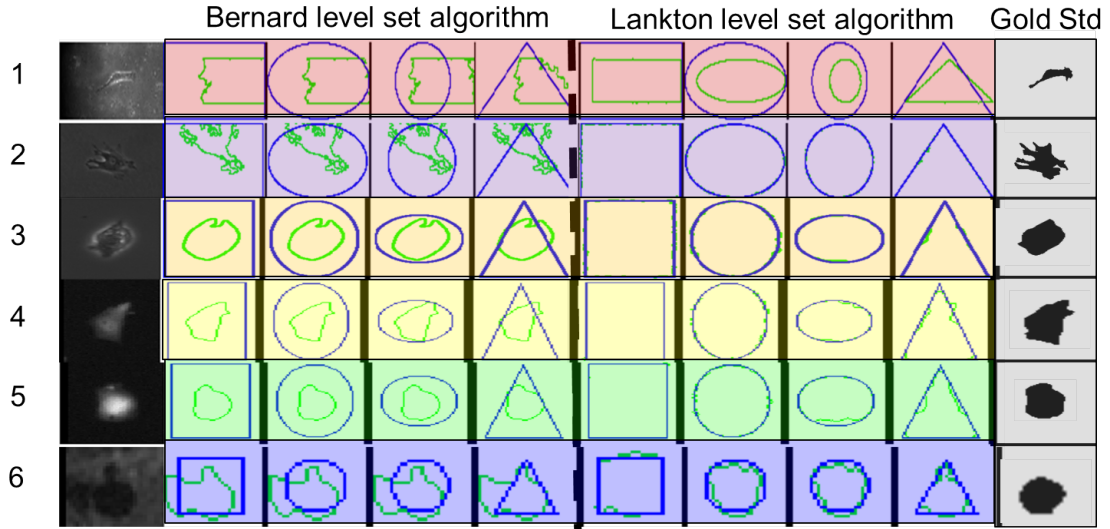


Figure 6.1: Representative segmentation results exemplifying that a trial-and-error effort to find a contour initialization may or may not lead to successful use of a level set method. Raw images (column 1), shown for a biological object from each dataset in the image library, were processed with the “Bernard level set algorithm” [10] (columns 2–5), the “Lankton level set algorithm” [48] (columns 6–9), and manually (column 10). Blue lines show initial contours, green lines the final segmentation.

6.1 Methods

To find a contour initialization method that works well in general, we designed and implemented a system that supports trial-and-error analyses by applying all combinations of chosen initial-contour shapes and level-set algorithms to all image sets in an image library. A user runs the system with one command and can configure the system to apply a collection of level set algorithms initial-contour pairings to a collection of image sets.

6.1.1 Segmentation System

Images are processed sequentially. For each image, the system applies the segmentation method with the associated curve initialization method. Different segmentation and curve initialization configurations with different parameter settings can be applied for different image sets (described below). Next, the segmentation result is post-processed by filling holes and keeping only the largest object. Finally, the system saves the resulting binary

Table 6.1: List of algorithms analyzed in comparison study, their inclusion in biomedical image analysis toolboxes, and initialization methods reported in the initial publications.

Tool	Software Options	Published Curve Initializations
Geodesic Active Contours [15]	Fiji[80], ITK[23], Creaseg [20]	Rectangle, Circle
Chan Vese level set method [16]	Creaseg [20]	Square, Circle
Lankton level set method [48]	Creaseg [20]	Rectangle, Square, Circle, Ellipse, Merged Rectangles
Li level set method [53]	Creaseg [20]	Square, Circle, Ellipse, Triangle
Shi level set method [83]	Creaseg [20]	Circle, Merged Circles
Bernard level set method [10]	Creaseg [20]	Square, Circle, Ellipse, Merged Circles

segmentation as an image.

Segmentation Modules. Each of six publicly available implementations [20] of level set algorithms are wrapped into a single module that the user may use interchangeably in the system: geodesic active contours [15], Chan Vese level set method [16], Lankton level set method [48], Shi level set method [83], Bernard level sets [10], and Li level set method [53] (**Table 6.1**). Each segmentation module is decoupled from the initial contour by being linked to an *Initial Contour Module* option that, at run time, creates an initial contour.

Initial Contour Modules. Each initial contour module shares the same interface. Given an input image, it returns a binary mask of the same dimensions. The system supports four initial contour methods the user may use interchangeably: rectangle, ellipse, circle, and a triangle. To create the contour, the *rectangle* module uses the boundary of the rectangle drawn by removing n pixels from all sides of the image region, the *ellipse* module uses the boundary of an ellipse drawn to span the image region downsized by n pixels on all sides, the *circle* module uses the boundary of a circle drawn at the center of the image region with a radius of half of the smallest image region dimension minus n pixels, and the *triangle* module uses the boundary of a triangle drawn to span the image region downsized by n pixels on all sides using two corners of the bounding box on the bottom image side

and the midpoint between both corners on the top image side. The user can configure parameter n to control the size of the shape.

6.1.2 Crowdsourced Initial Contour Module

When basic geometric shapes are insufficient, as reported for phase contrast image sets [22, 54], a user can instead use a crowdsourced initial contour to create an estimate of the object boundary that is closer to the true boundary. We incorporate the publicly available on-line annotation software, LabelMe [79], into the system usage pipeline to collect the initial contours. To address concerns about trusting annotations from a single annotator, whether due to weaker skills or even malicious motivations, we incorporated into the pipeline the *Probability Maps (p-map)* algorithm so that the user can combine multiple crowdsourced segmentations for each image. This algorithm takes as input N segmentations and outputs a single segmentation where a pixel is labeled as foreground when at least M of the crowdsourced segmentations label it as foreground and background otherwise. Finally, the segmentation result is post-processed to fill holes and keep only the largest object.

6.2 Experiments

We conducted three studies using the proposed system on biomedical images to examine which among the six freely-available level set methods yield the most accurate segmentations for various image modalities, what is the impact of contour initialization on segmentation quality, and whether paid crowdsourced workers can be leveraged to expedite successful use of level set methods for biomedical images.

We analyzed the algorithms on a total of 271 images from BU-BIL:1-5. We computed scores that indicate how closely algorithm-generated segmentations match gold standard segmentations provided with the image library. We quantitatively analyzed each algorithm for all images using *IoU*, a standard evaluation metric (i.e., $\frac{|A \cap B|}{|A \cup B|}$ where A and B represent the set of pixels in the gold standard and algorithm-generated segmentations respectively).

Study 1: Impact of Initial Contour. We applied our system to all images in the library to collect segmentations using all six algorithms. We did this 12 times to analyze the impact of the shape and size of the initial contour by setting $n = 5, 15,$ and 25 pixels for the rectangle, ellipse, circle, and triangle.

Study 2: Comparison of Level Set Methods. We applied our system to all images in the library to collect segmentations using all six algorithms. We set the initial contour to the gold standard segmentation mask. We also compared algorithms using as the initial contour a circle with $n = 15$ since we found in Study 2 this pair performed well for both phase contrast and fluorescence images.

Study 3: Analysis of Using Crowdsourced Initial Contours. We applied our system to all images in the library to collect segmentations using all six algorithms paired with the initial contours created by crowdsourced workers. To create the initial contours, we collected five crowdsourced annotations per image and fused the segmentations into a single binary mask with the *p-map* algorithm setting $M = 3$ (i.e., a pixel is part of the object only if at least three annotators included it as part of the object). To minimize concerns about work quality, we only accepted workers that had previously completed at least 100 HITs and received at least a 92% approval rating. Workers receive a five step set of instructions detailing how to submit a HIT followed by pictures exemplifying good and bad annotations. All submitted HITs were accepted and workers were paid \$0.02 for completing each drawing task.

6.3 Results

Study 1: Impact of Initial Contour. We found that the shape and size of the initial contour can impact algorithm performance for both phase contrast and fluorescence images (**Figures 6.1, 6.2**). For *fluorescence* images, we found a noticeable difference in algorithm performance based on initial contour shape and size for all but the Bernard level set algorithm. For initial contour shape, the ellipse and circle led to the best performance for

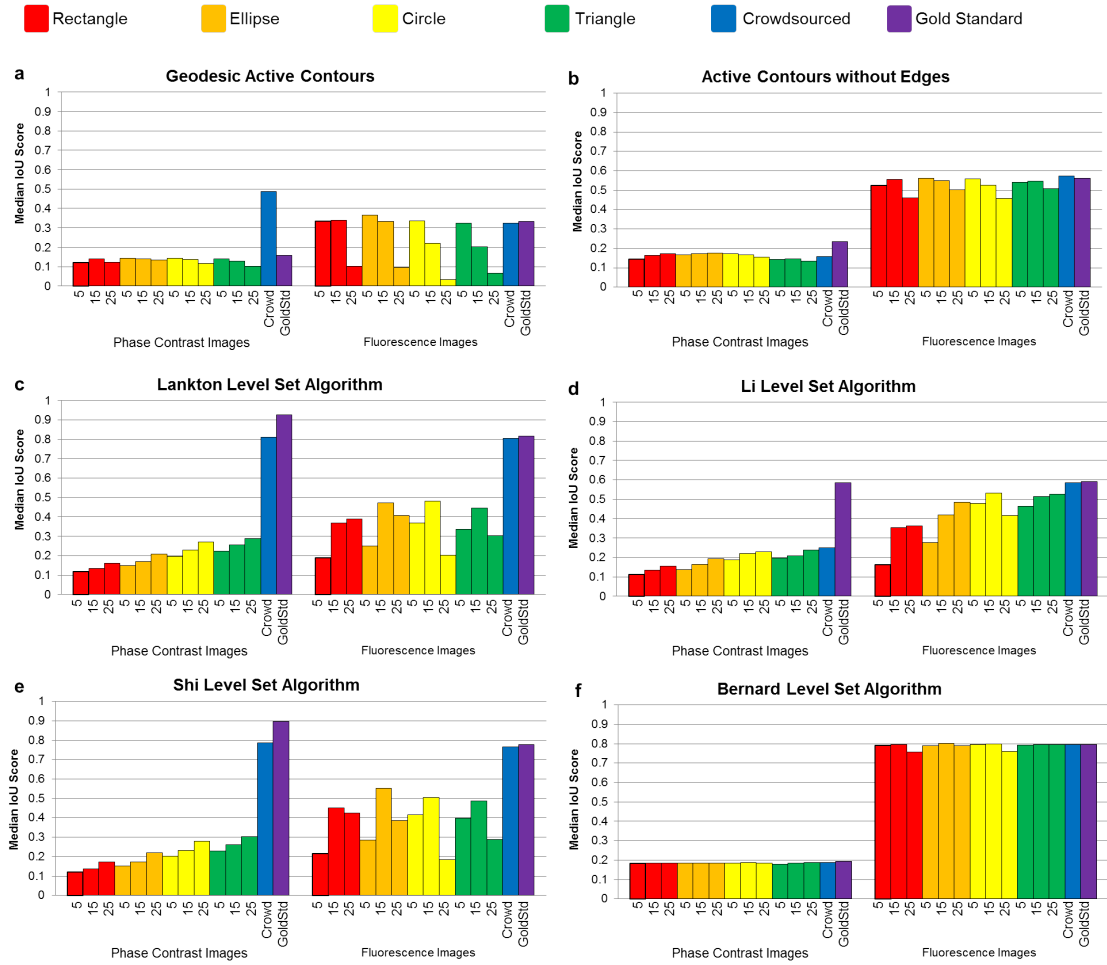


Figure 6.2: Results showing the performance of six level set methods paired with 14 unique contour initializations. Each plot shows the median IoU score for all phase contrast images and fluorescence images independently when using as the initial contour four geometric shapes at three different sizes, the crowdsourced segmentation boundary, and the gold standard boundary.

most of the algorithms. For initial contour size, for most algorithms, the medium-sized bounding region led to the best performance ($n = 15$). For *phase contrast* images, we found a slight difference in algorithm performance based on initial contour shape and size for the Lankton, Li, and Shi level set based algorithms. For initial contour shape, the ellipse, circle, and triangle each led to better performance for different algorithms and only the rectangle consistently led to inferior or equal performance. For initial contour size,

for most algorithms, a smaller bounding region led to the best performance regardless of initial contour shape (larger n value).

Study 2: Comparison of Level Set Methods. We found, when the initial contours were set to the boundaries of the gold standard segmentations, that only the Lankton, Shi, and Bernard level set algorithms performed well (**Figure 6.2**). For the *phase contrast* images, the Lankton and Shi level set algorithms yielded the best performance. For the *fluorescence* images, the Bernard and Shi level set algorithms yielded the best performance. We found that the top-performing algorithms resulted in scores over 10% higher for *phase contrast* images than for *fluorescence* images.

We found, when comparing algorithms using the circle as an initial contour, that different algorithms performed well for different image modalities (**Figure 6.2**). For *phase contrast* images, we found that the Lankton and Shi level set algorithms led to the best performance. For *fluorescence* images, we found that the Bernard level set algorithms led to the best performance. We found that the top-performing algorithms resulted in scores over 50 percent points higher for *fluorescence* images than for *phase contrast* images.

Study 3: Analysis of Using Crowdsourced Initial Contours. We found that pairing segmentation algorithms with our proposed initial contour method yielded over 50 percent points performance improvement for *phase contrast* images and comparable performance for *fluorescence* images in comparison to the top-performing algorithm initial-contour pairings found in study 2.

6.4 Discussion

We analyzed freely-available level set algorithms to report about algorithms with immediate wide-spread relevance. We were surprised that most of the algorithms yielded low-quality segmentations when evolving the gold standard boundary to a final boundary. We infer from these results that the algorithm energy functionals most closely matching the inherent properties of the studied image modalities and biological structures are Lankton and Shi

level set algorithms for the phase contrast images and Bernard and Shi level set algorithms for the fluorescence images. We infer from our results that, when applying these algorithms in practice, all the studied initial contour shapes and sizes yield high quality segmentations when paired with the Bernard level set algorithm, while the other three level set algorithms should be paired with an initial contour that closely hugs the true object boundary. Lastly, we infer from our results that non-expert paid crowdsourced workers can replace domain expert involvement to create initial contours for biomedical images.

6.5 Conclusions

Greater wide-spread use of algorithms to successfully collect high quality segmentation annotations relies on knowing which algorithm to choose and then how best to use it. We found that only a few of the studied freely-available level set algorithms are designed with assumptions that are well-suited for the studied phase contrast and fluorescence images of cells. For the well-suited algorithms, we found that one simple detail, the initial contour, can trigger over a 50 percent point improvement for phase contrast images. Finally, our results show the potential of using paid crowdsourced workers without domain-specific training to reliably and inexpensively replace domain experts in creating initial contours that are needed to use these algorithms effectively. Our study may be a start point towards a larger community effort to empower those applying level set methods to make an informed choice about which algorithm to use, how to use it effectively, and how to replace their efforts with non-experts. We encourage the reader to leverage our system so that the number of comparison studies of this sort can grow to address a wider range of biomedical problems important to the research community. Future work will explore how to more efficiently utilize crowdsourcing by analyzing the reliability of crowdsourced workers and what number of annotations are necessary. Possible future research directions also include running the study on a larger image set and quantitatively analyzing the causes in images that influence the successes and failures of the different algorithms and initial contours.

Chapter 7

Hybrid System - Predicting Computing Source

A common question individuals ask when needing to annotate images in practice is whether, for a given image, available automated options are sufficient for their purposes or they should instead bring humans in the loop to create accurate annotations. The knowledge of which segmentation algorithm(s), if any, will succeed is a highly-specialized skill often resigned to computer vision PhDs or applications specialists who have spent years studying the variety of options. We explore the problem of automatically predicting when to use available algorithm options versus humans for the task of demarcating object regions, i.e., creating segmentations.

In our work [31], we focus on intelligently recruiting human annotation work by leveraging predicted performance of segmentation algorithms in the absence of ground truth segmentations. This is of interest for many applications where coarse or fine-grained segmentations are needed for algorithm input or as a final result. We examine both interests for the tasks of creating input for interactive segmentation algorithms and evaluating resulting segmentations.

Specifically, one valuable application is to distribute the efforts between humans and computers to create coarse initial outlines needed as input for interactive segmentation algorithms [16, 48, 78]. These algorithms refine user supplied coarse segmentations in an attempt to produce higher quality annotations which incorporate missing pieces and trim off excess pieces. Initialization is a critical factor that can drive the success or failure of interactive segmentation algorithms, and a one size fits all solution remains to be found. Some researchers have suggested offloading the time-consuming, labor intensive task of

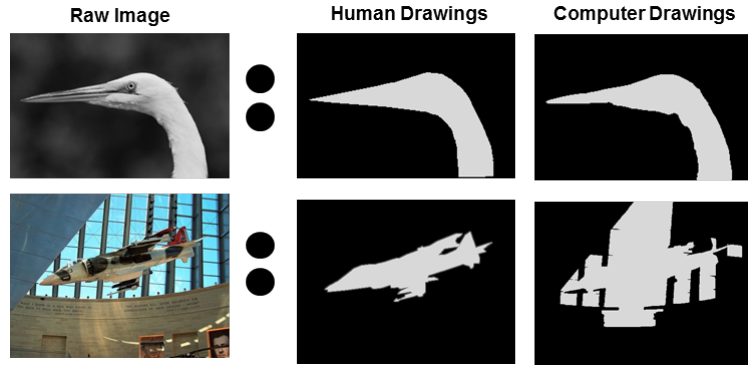


Figure 7.1: Use a human or computer drawn segmentation?

creating coarse segmentations to crowdsourced workers [36, 44]. Other researchers have proposed using fully-automated methods to create coarse segmentation estimates by relying on simple geometric shapes such as bounding boxes [16, 48] or more sophisticated segmentation methods [22]. We explore the plausibility of combining strengths of both approaches by distributing the annotation task to computers when they are successful and relying on human input otherwise.

Collecting high-quality, fine-grained segmentations is another task that can be supported by distributing annotation work between humans and computers. Segmentation of objects in everyday, biomedical, and medical images at the fine-grained level has been addressed by numerous segmentation systems discussed in the mainstream literature [16, 48, 78], which purportedly have the potential for widespread impact. These algorithms differ in the computational assumptions they embed that determine how to separate an object from the background for a given image. For example, some methods assume there should be two homogeneous intensity regions either globally [16] or locally [48]. Many researchers agree that there is not a one-size-fits-all segmentation solution. The challenge for users to efficiently exploit these algorithms is to know when each algorithm will succeed. Our work examines how to automatically select a best-suited segmentation tool or recommend human involvement when it is believed no suitable automated options are available.

A natural question is what prediction framework should one apply to decide whether

to rely on algorithms, or pull the plug on them and use human annotations? Two types of approaches relevant for predicting the likelihood of segmentation algorithm success have been proposed both in the computer vision [14, 58, 75] and medical imaging [46] communities. Our work was inspired by the desire to develop a single prediction system that proves applicable across domains, handling large variations in image content and modality.

Predicting directly from an image the probability an algorithm will succeed is one plausible approach [58]. A crucial question, however, is whether one can detangle image segmentation difficulty from knowledge of one’s available segmentation tools. Moreover, a quick analysis of running multiple segmentation algorithms on multiple images often reveals that each algorithm will produce dramatically different results and work well in different contexts.

Another plausible approach is to predict, using an algorithm-generated segmentation, the quality of the segmentation in absolute terms [14, 46, 75]. In general, these methods use supervised learning to build prediction systems. Specifically, for a set of images, the system runs the segmentation algorithms, extracts features characterizing the images and segmentation results, and then computes scores indicating the similarity of algorithm generated segmentations to ground truth segmentations. Then, one uses machine learning tools to learn whether some weighted combination of computed features can be combined to predict the observed scores. Previous work used intensity based features which we found did not generalize well in our study.

Our work is a contribution to the emerging research field at the intersection of human computation and computer vision. Developments in crowdsourcing systems reveal it is possible to rapidly collect large amounts of human annotation [55, 79]. We explore how to involve humans to contribute to computing in hybrid algorithm-crowdsourcing systems.

Broadly speaking, the aim of this work is to minimize human involvement while collecting accurate segmentations. Successful solutions may be applicable when one needs to capture highly detailed, fine-grained information for shape analysis, which includes characterize tumors in medical images, automatically analyze product quality in factories, and

visualize 3D structures. Successful solutions also may be applicable for creating coarse segmentations which are highly valuable starting points for solving many downstream image analysis tasks such as object detection [40], recognition, and tracking. The key contributions of this work are:

- A method to predict the quality of candidate object segmentations in the absence of ground truth.
- A system that predicts when to delegate the task of creating coarse segmentations used by interactive segmentation algorithms to computers or humans.
- A hierarchical prediction system for interactive segmentation algorithms that automatically identifies a best-suited initialization and then evaluates the resulting segmentation.

7.1 Predicting Segmentation Quality

Our motivation is to build a reliable prediction model that indicates when a segmentation algorithm produces a reasonably accurate object region. Given that a segmentation algorithm can produce results that transition from “miserably-poor quality” to “nearly-perfect quality” in a continuous manner, we chose a regression rather than classification tool. A regression approach enables flexibility for different applications by not locking into a single definition regarding what defines a “sufficient” versus “insufficient” segmentation.

7.1.1 Prediction Model

Our system uses a multiple linear regression model, a supervised learning tool. This prediction model leads to easy to interpret, intuitive systems. The model can be rewritten as $y = X\beta + e$ where \mathbf{y} denotes a column vector of segmentation quality scores, X denotes the model specification matrix that specifies all observed predictor values, β denotes a column vector of model parameters, and e denotes the vector of random errors between y and predicted values $X\beta$. The objective is to learn β so that e is minimized.

7.1.2 Training Data Generation

We want to capture in our training data the variability between good and bad segmentations that can arise in practice. Towards this goal, we have our system collect 11 binary segmentation masks per training image. Then, our system represents each training instance with nine features treated as predictor variables and a segmentation quality measure for the response variable.

To ensure there are positive examples, our system creates three binary masks based on the ground truth segmentations. The system uses the ground truth directly. Our system also dilates and erodes the ground truth binary mask by three pixels to simulate a slightly under-segmented and over-segmented segmentation respectively. These examples highlight object appearances when fine details either get smoothed out or chopped off.

For negative examples, we derive a variety of binary masks from segmentation algorithms to reveal their diverse failure behaviors. There are a wide array of segmentation algorithms one could use to generate training data. We chose three fully-automated, computationally fast segmentation algorithms to generate binary masks for training data that have widespread applicability given their simplicity and availability in many image processing tool kits. Our system applies two algorithms: Otsu thresholding and adaptive thresholding method using the local median from a window size of 45 pixels (1-4). We use the result and its complement. Our system also applies a third variant of adaptive thresholding method using the local mean from a window size of 45 pixels (5). Finally, our system applies three variants of the Hough Transform with circles method using a circle radius of 3, 5, and 10 respectively (6, 7 & 8). For each binary mask, our system then post-processes the results to contain exactly one object by filling all holes to address that our chosen algorithms tend to have lots of holes in resulting segmentations. Then, our system post-processes the results to contain exactly one object to keep only the largest object.

Finally, to create the labeled data, for each of the 11 candidate segmentations, our

system computes an input feature vector and an output label. To create each input feature vector, the system computes the nine features discussed in Section 7.1.3. To create each output label, the system computes a score indicating the quality of the algorithm-drawn segmentation. We use the standard IoU score (i.e., $\frac{|A \cap B|}{|A \cup B|}$).

7.1.3 Prediction Features

Our motivation is to use knowledge about algorithm behavior on everyday and biomedical images to choose predictive features (**Figure 7.2**, top row). We observe that when algorithms “mess up” they do so amazingly well with characteristics unlike what one would expect from widely meaningful object shapes (**Figure 7.2**, middle row). We propose nine features derived from the binary segmentation mask. We hypothesize that, in aggregation, these features may account for objects of different shapes and sizes.

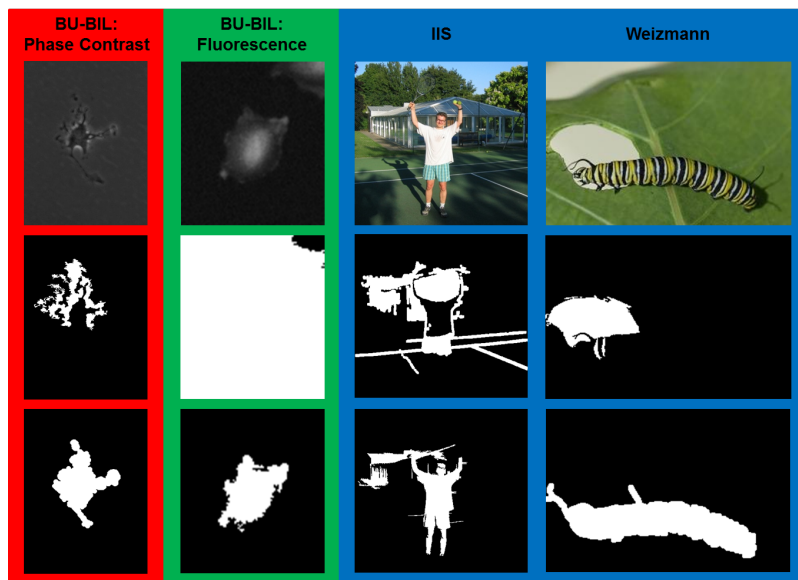


Figure 7.2: We propose a method to automatically evaluate candidate segmentations. Our design was motivated by observations of algorithm behavior when demarcating everyday and microscopic objects captured with three fundamentally different image acquisition systems (top row). We chose nine predictive features describing the segmentation binary mask that characterize algorithm failure behavior (middle row) which typically is in stark contrast to what is observed when algorithms accurately capture object regions (bottom row).

Segmentation Boundary. When algorithms fail, resulting segmentations often have boundaries characterized by an abnormally large proportion of highly-jagged edges. We implement two boundary-based features to capture this observation. We compute the *mean* and *standard deviation of the Euclidean distance of every point on the segmentation boundary to the centroid*. The boundary is defined as all pixels on the exterior of the object in a binary mask using an 8-connected neighborhood. The centroid is defined as the center of mass of the segmentation in the binary mask.

Segmentation Compactness. When algorithms succeed, proposed segmentations are often compact, meaning that included pixels typically lie within a small distance from each other. This region compactness is not commonly observed when algorithms fail. We implement three features to capture this observation. Two measures compute the coverage of segmentation pixels within a bounding region. *Extent* is defined as the ratio of the number of pixels in the segmentation proposal to the number of pixels in the area of the bounding box. *Solidity* is defined as the ratio of the number of pixels in the segmentation proposal to the number of pixels in the area of the convex hull. We also compute the *shape factor* to capture the circularity of the region proposal since a pure circle is a good measure to indicate highly compact objects. It is defined as the ratio of region area A to a circle with the same perimeter P : $\frac{4\pi A}{P^2}$.

Location of Segmentation in Image. When algorithms fail, resulting segmentation regions often lie closer to the edges of images. We compute the *normalized x* and *y centroid coordinates* of the segmentation centroid in the image to capture this observation. Specifically, we compute the x value of the center of mass divided by the image width and y value of the center of mass divided by the image height.

Coverage of Segmentation in Image. When algorithms succeed, resulting segmentations often do not cover abnormally large or small areas in the image. We implement two features to capture this observation. First, we compute the *fraction of pixels in the image that belong to the segmentation*. Second, we compute the *fraction of pixels in the image that belong to the bounding box of the segmentation*.

We use these features, together with the training masks discussed in Section 7.1.2, to train the regression model.

7.2 Segmentations by Humans or Computers?

We aim to meet demands for high quality segmentations while minimizing human involvement. Towards this goal, we predict when to rely on humans versus automated methods to create segmentations. We propose a system based on a budgeted approach. Specifically, if one can collect human annotations for only $N\%$ of images, we aim to best to use that human effort. First, we focus on the problem of creating coarse segmentations that an interactive segmentation subsequently refines. Then, we additionally consider the budget problem of distributing the work between algorithms and humans to create final segmentations.

Interactive Segmentation Algorithms. We include in our system three options for interactive segmentation algorithms that are important both in the computer vision and medical imaging communities - Grab-Cut [78], Chan Vese level sets [16], and Lankton level sets [48]. These algorithms, from the graph cuts and level set families, represent a set of optimization-based approaches that deform a user-provided initial segmentation, which we call a "coarse" segmentation in the following sections. Grab Cuts enforces color homogeneity and spatial proximity. Chan Vese level set method uses global image information to try to separate an image into two homogeneous intensity regions while enforcing smoothness of the object boundary. The Lankton level set method uses local neighborhood statistics for each pixel to separate an object from the background so that there are two homogeneous intensity regions within a band containing the object boundary.

Coarse Segmentation: Computer or Human? Our aim is to collect exactly one input coarse region per image to maximize overall quality while minimizing the cost asso-

ciated with repeatedly running an interactive segmentation algorithm. This is particularly important for methods that take on the order of minutes or more per image, which is commonly the case for level set based algorithms. There are two key questions underlying building such a system for a human allocated budget: 1) how to create the initialization for each image? and 2) which images get human input?

While ideally one could rely on a single fully-automated algorithm to create all coarse segmentations, in practice different algorithms often work well for different conditions. We therefore propose a system where, for each image, the decision is to either rely on one of eight automated methods or humans to create a coarse segmentation. We apply the eight computationally fast automated methods described in Section 7.1.2 to create the eight candidate segmentations. Then, we apply our prediction framework to indicate the quality of each candidate segmentation. Next, we choose the highest scoring segmentation as our automated candidate. Finally, we sort all images from highest to lowest predicted scores based on all selected automated candidates. We enlist human involvement for the allotted percentage of images where predicted algorithm scores are lowest.

Fine-Grained Segmentation: Computer or Human? We propose an alternative two stage prediction system to create high quality segmentations. In the first stage, the prediction framework is applied to choose the best-suited algorithm to create a coarse segmentation for every image. Then, each coarse segmentation is refined by an interactive segmentation algorithm. In the second stage, the prediction framework is applied to all resulting segmentations from the interactive segmentation algorithm to estimate the quality of each result. Again, all images are sorted based on highest to lowest predicted scores characterizing the quality of the segmentations. Finally, humans are recruited for the allotted percentage of images where predicted scores are lowest.

7.3 Experiments and Results

We conducted studies to both analyze the power of our prediction framework to accurately evaluate candidate object segmentations and estimate the value of our system to more efficiently allocate human resources when using interactive segmentation algorithms.

We conduct our studies using three public datasets representing three imaging modalities: Boston University Biomedical Image Library (BU-BIL:1-5) [37] includes 271 grayscale images coming from three fluorescence microscopy image sets two phase contrast microscopy image sets, Weizmann [4] consists of 100 grayscale images showing a variety of everyday objects, and finally Interactive Image Segmentation [30] (IIS) includes 151 RGB images showing a variety of everyday objects. Each of these datasets come with pixel-accurate ground truth segmentations for evaluation purposes.

7.3.1 Predicting Quality of Candidate Segmentation

We analyzed the predictive power of the proposed model to evaluate a given segmentation in the absence of human input. We evaluated and compared prediction models using Pearson’s correlation coefficient (CC) and mean absolute error (MAE). CC indicates how strongly correlated predicted scores are to observed scores. Values range between +1 and -1 inclusive, with values further from 0 indicating stronger predictive power of a model. MAE is a linear measure that indicates the average size of prediction errors when negative signs are ignored. It is the mean from all computed absolute values of the differences between the predicted and observed IoU scores.

To train and test our model, we used the code from the freely-shared data mining system Weka [38] to solve for the model parameters $\beta_0, \beta_1, \beta_2, \dots, \beta_k$. It takes as input n tuples, each consisting of k observed predictor variables followed by the observed response. We used the M5 greedy feature selection option to iteratively evaluate each model parameter and eliminate any parameters that do not yield prediction improvements. We created 11 candidate segmentations per image using the methods described in Section 7.1.2 to create

Table 7.1: Evaluation and comparison of our model and the baseline method CPMC at predicting the absolute IoU score indicating the quality of a candidate segmentation in the absence of human input. We evaluate with respect to two evaluation metrics on three datasets: correlation coefficient (i.e., CC) and mean absolute error (i.e., MAE). We report performance scores for our method both for training and testing exclusively on one set of images (“Ours: Single-Dataset”) as well as when testing on one dataset and training on the other two datasets (“Ours: Cross-Dataset”). Higher CC scores are better and lower MAE scores are better.

Image Library (# segs):	BU-BIL (2981)		Weizmann (1100)		IIS (1661)		All (5742)	
Evaluation Metric:	CC	MAE	CC	MAE	CC	MAE	CC	MAE
Ours: Single-Dataset	0.69	0.18	0.69	0.2	0.78	0.18	0.68	0.2
Ours: Cross-Dataset	0.61	0.31	0.64	0.24	0.68	0.22	NA	NA
CPMC	0.36	0.33	0.61	0.32	0.67	0.31	0.53	0.32

a total of 5,742 segmentations which we characterized for training and testing.

Ours: Single-Dataset Predictions. We analyzed the proposed prediction framework per dataset. We evaluated three models that were dataset-specific (i.e., Weizmann, IIS, BU-BIL) as well as one model built using the combination of images from all datasets. For each of the four models, we trained and tested our linear regression model using 10-fold cross-validation. We used the predictions for all images collected from the ten partitions to compute the correlation coefficient and mean absolute error scores (**Table 7.1**, row 1). Our approach performed well with high correlation coefficients and low mean absolute scores.

Ours: Cross-Dataset Generalization. To analyze whether the success of the prediction models is due to over-fitting to statistics from a particular dataset, we evaluated how well a prediction model trained on two of the datasets performs on the third dataset (**Table 7.1**, row 2). Surprisingly, we found the models continued to be very effective, even when trained on two everyday image sets and applied to biomedical images representing two imaging modalities not observed during training. This is possibly because resulting binary masks when algorithms fail tend to be consistent across datasets.

Baseline. We compare our model to that used in the CPMC system [14], which also predicts a IoU score indicating the quality of a given segmentation for a given image. We used the publicly available code shared by the author. The CPMC prediction system was

trained on everyday images using a non-linear random forests regression model and a mix of 34 intensity-based and shape-based features. We applied the system for each algorithm-generated segmentation and used all the predictions to compute the CC and MAE scores (**Table 7.1**, row 3).

While the CPMC method was designed to generalize across different object types, it was less precise than our model on all studied datasets. One possible reason for this performance difference is due to our advantage in having trained and tested our system with segmentations created by the same algorithms. This suggests a possible value in learning the statistics of specific tools one intends to use in systems rather than relying on one size fits all approaches. The results also reveal a plausible limitation that the CPMC method does not generalize well for objects observed in images captured with different image acquisition technologies (e.g., phase contrast and fluorescence microscopy imaging).

7.3.2 Interaction Tools - Human vs Computer Input

With interactive image segmentation tools, the user is asked to provide an initial segmentation which will subsequently be refined. Currently, users either create the initial segmentations by relying exclusively on automation [16, 22, 48] or human involvement [36, 44]. We examined the trade-off between quality and human effort when combining both approaches. We ran our study on all 522 images in the three image sets (i.e., BU-BIL, IIS, Weizmann). We evaluated the impact of our prediction framework with three interactive segmentation tools: grab cuts, Lankton level set algorithm, and Chan Vese level set algorithm.

For each interactive segmentation tool, we compared our method with four other methods. We made comparisons by evaluating how the allocated amount of human annotations relates to the quality of segmentations created by the interactive segmentation tools. In particular, better methods would yield higher quality with less human effort.

- *Our Predictor:* We used the method discussed in Section 7.2 with the cross-dataset predictions discussed in Section 7.3.1 in order to avoid biasing our system to learn the

statistics for a particular image set.

- *Perfect Predictor*: We replaced the predicted score for our method (Section 7.2) with the actual IoU score that indicates the similarity of each candidate segmentation with the ground truth segmentation. This influences both the input segmentation chosen per image as well as which images get allocated human input. This predictor demonstrates the best one can expect with our system.

- *Chance Predictor*: We randomized the selection of an algorithm among the eight segmentation options (Section 7.2) and the order of images. This influences both the input segmentation chosen per image and images allocated human input. This predictor demonstrates what one may expect from random distribution of annotation efforts to available human and computer resources.

- *Bounding Box*: We used a bounding box as the initial segmentation for all images. To do this, we remove n pixels from all sides of the image region. We set n for each image to be 5% of the number of pixels in the minimum image resolution dimension. We randomly selected images for human involvement.

Following Jain and Grauman [44], we simulated coarse human input from computer generated segmentation by dilating the ground truth segmentations by 20 pixels.

In total, we evaluated 7,830 resulting segmentations created using five initialization methods with three interactive segmentation tools. We evaluated overall segmentation quality when we used human input for the following percentages of images for each of the three interactive segmentation algorithms: $N = 0, 5, 10, \dots, 100\%$. Results demonstrate that a single prediction model which instructs how to distribute annotation efforts between humans and computers successfully led three interaction tools to produce higher quality

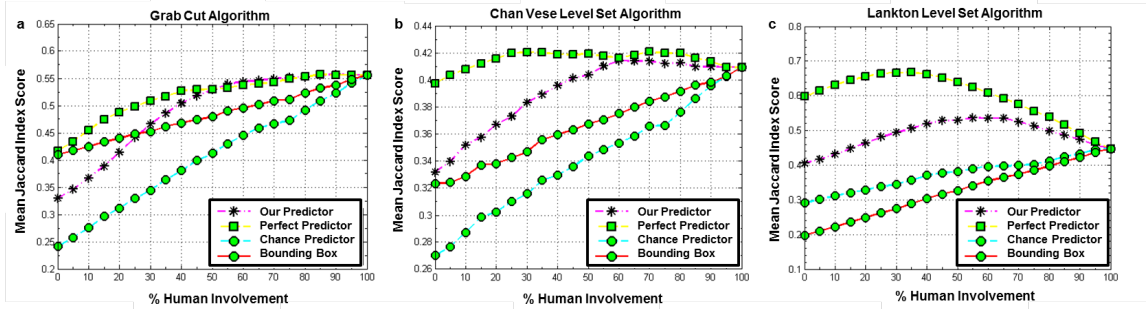


Figure 7.3: Different methods for distributing the effort to create coarse segmentation input between humans and computers leads to different results for three interactive segmentation algorithms (a-c). Each plot shows the mean IoU score indicating the overall quality for 522 segmentations created for everyday and biomedical images from three datasets as a function of varying levels of human involvement. Boundary conditions include exclusively choosing the segmentations created by automated options (0% human involvement) and human input (100% human involvement).

results at significantly reduced human costs than observed using random prediction schemes (i.e. chance predictor, bounding box) (Figure 7.3). Given that the best practitioners can achieve today is chance prediction, these findings can lead to immediate, practical implications for more effective use of interactive segmentation tools today.

Impact of Initialization. Using a good initialization as input is clearly important for interactive algorithms to produce the best segmentations they can (Figure 7.3). As observed in the three plots, each algorithm performed the worst when only relying on algorithm input (i.e., 0% human involvement) and steadily improved in performance as the allocated human input budget increased. However, some algorithms performed best when relying strictly on human input (i.e., 100% human involvement) while other algorithms performed best when relying on a combination of algorithm and human input. Two of the three algorithms always performed better when initialized with the collection of eight candidate segmentation options studied in this paper (i.e., perfect predictor) than with the commonly used bounding box [16, 48]. We hypothesize Grab Cuts initially performs better when initialized with bounding box because this algorithm always shrinks the initial segmentation which may be a poor behavior for some of the predicted input. In practice,

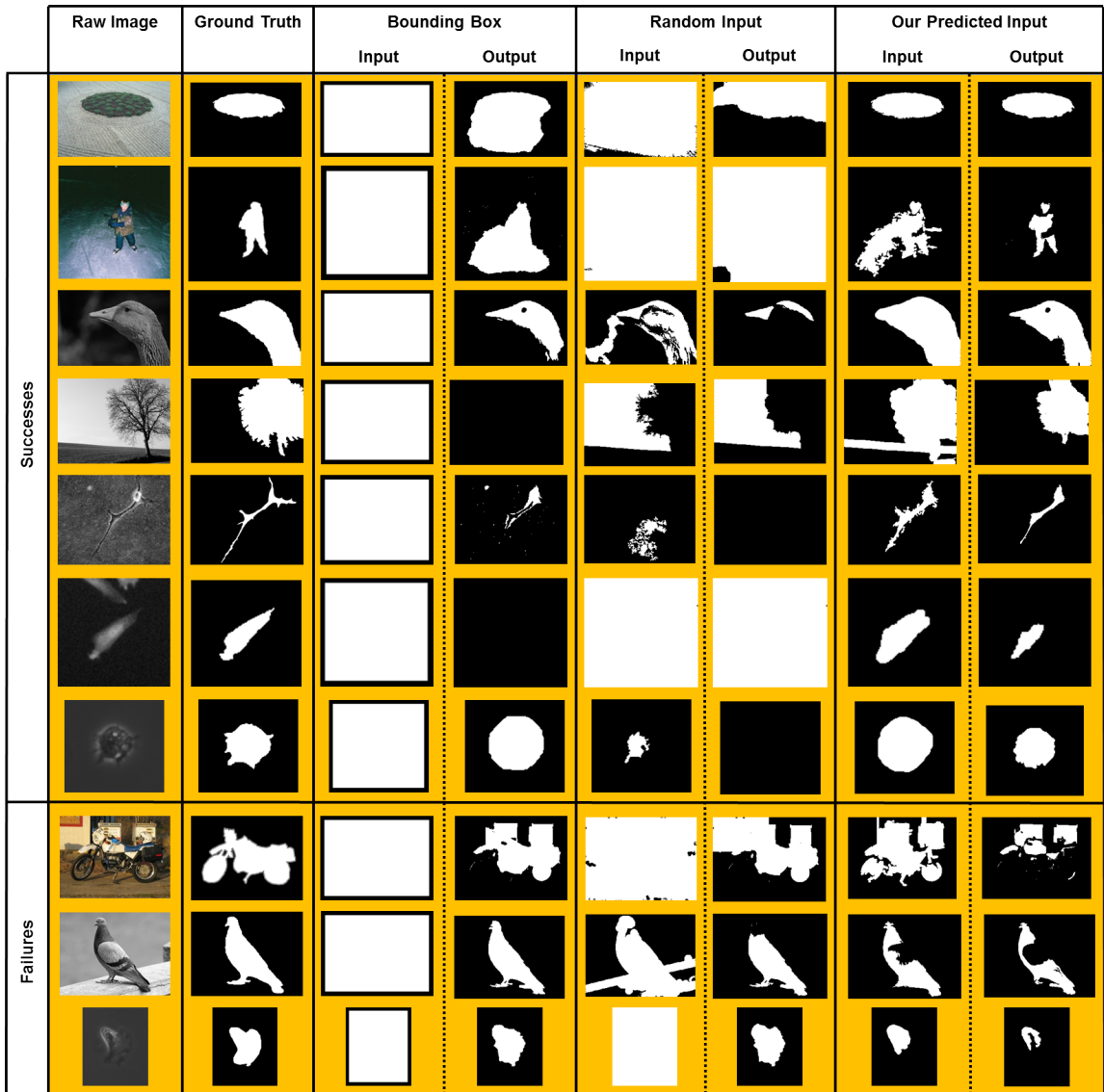


Figure 7.4: Sample results from the Grab Cuts algorithm when it is initialized using three fully automated methods: bounding box, randomly chosen method among eight automated options, and predicted best choice among eight automated options. As observed in the "Successes", the quality of segmentation results are higher when using well-chosen initial segmentation estimates (Our Predicted Input) rather than arbitrarily chosen initial segmentation estimates (Random Input, Bounding Box). As observed in the "Failures," an initial segmentation estimate that does not fully contain the object of interest can lead to poor segmentation results.

our prediction system, which determines how to distribute work between humans and computers, significantly outperformed random decisions (i.e., chance predictor, bounding

box) for all budget levels for two of the three studied interactive segmentation algorithms. Exemplar results illustrate the performance of the Grab Cut algorithm when initialized using the three fully-automated methods: our predicted segmentation option, a randomly selected segmentation option, and a bounding box (**Figure 7.4**).

Comparison of Interactive Segmentation Tools. The results demonstrate that picking a good initialization is not sufficient to guarantee high quality segmentations from interactive segmentation tools (**Figure 7.4**). All algorithms performed poorly when relying exclusively on human input. Only Lankton level sets demonstrated high potential as a one size fits all segmentation tool. We hypothesize this difference is because the Lankton level set algorithm relies on local information to refine boundaries which is in contrast to Chan Vese and grab cut algorithms which rely on global image information. Also surprising is the observation that different algorithms responded very differently to bounding box input. While it was insufficient for Lankton level sets which predominantly fails to propagate the shape to the true segmentation due to the inadequacy of the local information, it regularly was a reasonable input for the grab cut algorithm which always shrinks the initialization using global information. Performance of the three interactive segmentation algorithms initialized with the same input are illustrated in **Figures 7.5** and **7.6**.

7.3.3 Interaction Tools - Human vs Computer Output

Given that optimally initialized interactive segmentation algorithms often fail to yield high quality segmentations, users continue to face the challenge of how to exploit these algorithms only in the contexts they will succeed. We analyzed the effectiveness of our prediction framework to automatically decide when to recruit crowdsourced annotations to replace computer-drawn segmentations. We conducted a case study with the Weizmann dataset to evaluate the value of our two stage hierarchical prediction system (see Section 7.2, last paragraph) in practice. We conducted this study three times with the three aforementioned interactive segmentation tools.

Implementation. We used our prediction framework to both select input and output



Figure 7.5: Sample results for the everyday images (i.e., Weizmann and IIS). See the previous figure for the explanation of the image layout. In some cases, an interactive segmentation algorithm can perform well when using a low quality segmentation as input, as observed for the image of the sheep (row 6, Grab Cut algorithm). In other cases, none of the interactive segmentation algorithms perform well when initialized with a low quality segmentation, as observed for the image of the person (row 4).

from interactive segmentation algorithms. In stage 1, to initialize interactive segmentation tools, we used our prediction system to choose the highest quality automatically

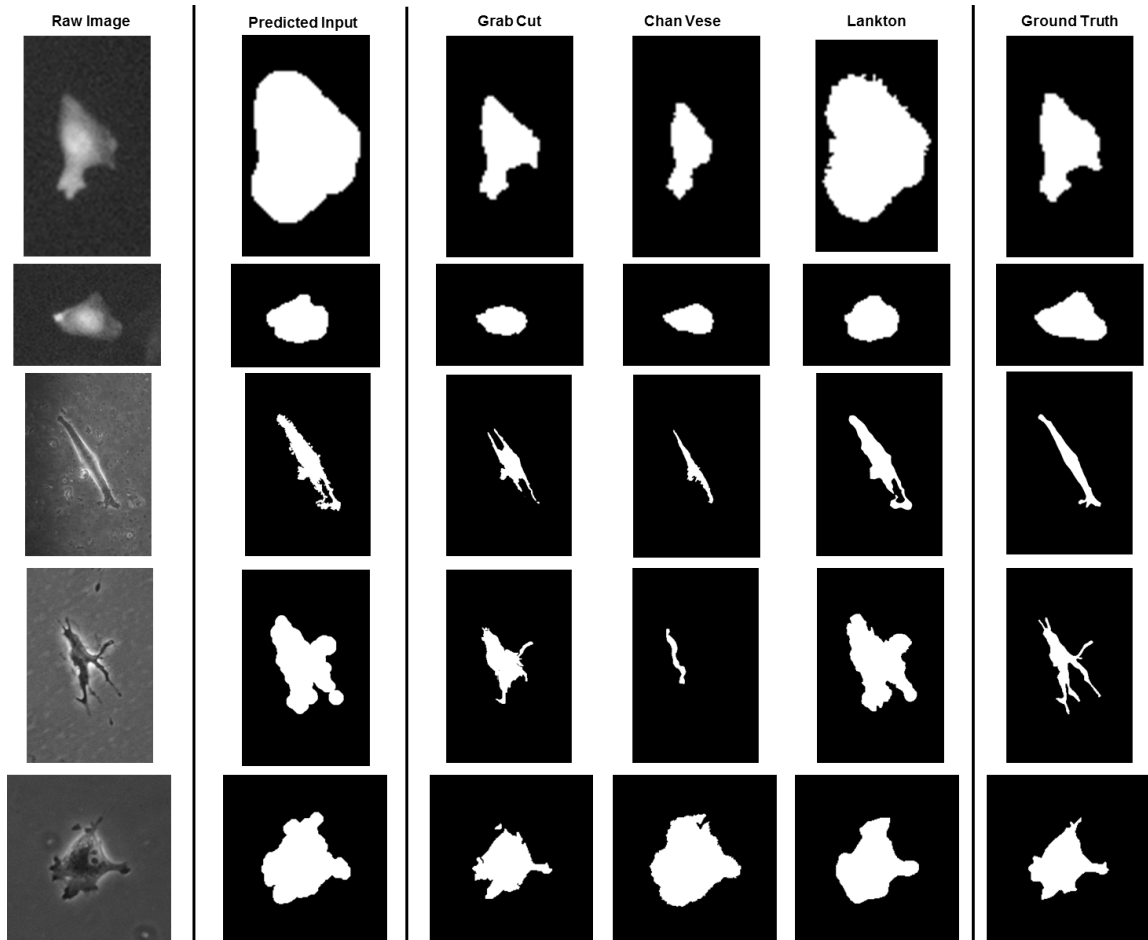


Figure 7.6: Sample results for the biomedical images (i.e., BU-BIL). Raw images (col 1) and the predicted input option from eight automatically generated options (col 2), followed by the resulting segmentation from the grab cut algorithm (col 3), Chan Vese level set algorithm (col 4), Lankton level set algorithm (col 5). The ground truth segmentation is shown in column 6. In order to produce segmentations that resemble the ground truth, interactive segmentation algorithms require sufficiently accurate segmentation input as well as suitable mathematical assumptions that match properties of each given image. All interactive algorithms can produce similar results, as observed in row 2. Each interactive algorithm can also produce dramatically different results from each other, as observed in row 4.

generated segmentation (i.e., largest expected IoU score) among the aforementioned eight fully-automated options, as described in Section 7.3.2. We trained our prediction model on images from IIS and BU-BIL to avoid over-fitting our model to statistics of the Weizmann dataset. In stage 2, after initializing the interactive segmentation algorithm, the system

then predicts the expected quality, in terms of an IoU score, of the resulting segmentation. Our prediction model to evaluate resulting segmentations came from the cross dataset training across all images discussed in Section 7.3.1.

For human input, we relied on crowdsourcing to collect human annotations. We recruited crowdsourced workers from the Amazon Mechanical Turk internet marketplace to create annotations using the on-line image annotation tool LabelMe [79]. We accepted all Mechanical Turk workers that had previously completed at least 100 jobs and received at least a 92% approval rating. We paid each worker \$0.02 to complete the drawing task for a single image. To overcome concerns about trusting annotations from a single annotator, we collected five drawings per image and then fused these annotations into a single annotation by labeling pixels as foreground only when the majority of images mark it as foreground. The segmentation was then post-processed to have a single object by filling holes and keeping the largest object.

System Evaluation. As done in the previous study, we compared our prediction framework to perfect predictions as well as chance predictions to decide when to use humans versus computers. We found that our prediction model, without training for the statistics of the interactive segmentation algorithms, still could lead to higher quality predictions

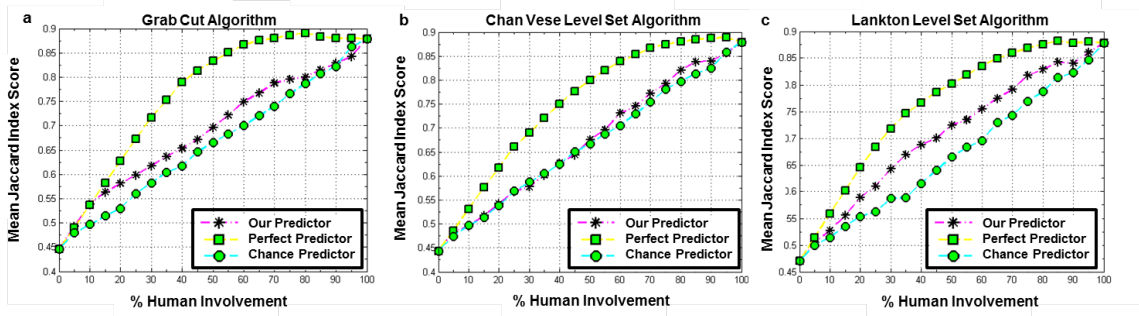


Figure 7.7: Predicting when to replace segmentations created by three optimized interactive algorithms (a-c) with annotations created by crowdsourcing improves overall quality for 100 everyday images (i.e., Weizmann dataset). Boundary conditions include exclusively choosing the segmentations created by the single algorithm (0% human involvement) and crowdsourcing (100% human involvement).

than by chance (**Figure 7.7**). We hypothesize our system performed poorly for Chan Vese results because the algorithm strongly enforces a smooth boundary which, our prediction model uses to assess segmentation quality. Overall, a simple step yielded higher quality annotations.

7.4 Conclusions

We sought to build systems that utilize the expertise of available computer and human resources to efficiently produce high quality segmentations. Our proposed prediction framework successfully evaluated the quality of candidate segmentations for our datasets, with stronger predictive capabilities than existing widely-used methods. We demonstrated this framework could successfully be leveraged to solve two novel tasks that involve intelligently distributing the annotation effort between algorithms and humans. While our work demonstrates clear benefits for applying our prediction framework as is to solve these segmentation tasks, our ultimate aim is to build a prediction system that is agnostic to the segmentation method, imaging modality, and object type.

Chapter 8

Closing Remarks

While in a perfect world image segmentation would be fully-automated, the unfortunate reality is that many segmentation tasks remain open problems today, despite decades of research from the computer vision community. We demonstrated the effectiveness of three hybrid system designs to produce superior results for the segmentation task compared to widely adopted stand-alone algorithm and crowdsourcing methods. Moreover, we demonstrated it is possible to achieve expert-grade annotations on biomedical and medical images with a hybrid system. While the merit of this research has already been recognized by two Best Paper awards [32, 36], the findings and recognition to date underscore the enormous potential for algorithmic-crowdsourcing approaches to benefit image and video analysis more widely. We hope that this work will encourage other researchers to explore hybrid system designs that may more effectively combine the strengths of crowd workers and algorithms to replace expert annotation efforts. Furthermore, we especially hope that this work will inspire future research that addresses challenges related to annotating biomedical and medical images, given that such improvements stand to benefit society at-large by addressing health problems.

Bibliography

- [1] Gnu image manipulation platform (Gimp). <http://www.gimp.org/>, 2014.
- [2] R. Adollah, M. Y. Mashor, N. F. M. Nasir, H. Rosline, H. Mahsin, and H. Adilah. Blood cell image segmentation: a review. In *4th Kuala Lumpur International Conference on Biomedical Engineering*, pages 141–144, 2008.
- [3] O. Alonso and R. Baeza-Yates. Design and implementation of relevance assessments using crowdsourcing. *Advances in information retrieval*, pages 153–164, 2011.
- [4] S. Alpert, M. Galun, R. Basri, and A. Brandt. Image segmentation by probabilistic bottom-up aggregation and cue integration. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8. IEEE, 2007.
- [5] Amira, software platform for visualizing, manipulating, and understanding, life science and bio-medical data. Retrieved January, 2014, from <http://amira.com>.
- [6] D. Ballard. Generalizing the Hough transform to detect arbitrary shapes. *Pattern Recognition*, 13(2):111–122, 1981.
- [7] P. Bamford. Empirical comparison of cell segmentation algorithms using an annotated dataset. In *Proceedings of the 2003 IEEE International Conference on Image Processing (ICIP), Vol. 2*, pages 1073–1076, September 2003.
- [8] S. Bell, P. Upchurch, N. Snavely, and K. Bala. OPENSURFACES: A richly annotated catalog of surface appearance. *ACM Transactions on Graphics (TOG)*, 32(4):111, 2013.
- [9] E. Bengtsson, C. Wahlby, and J. Lindblad. Robust cell image segmentation methods. *Pattern Recognition and Image Analysis*, 14(2):157–157, 2004.
- [10] O. Bernard, D. Friboulet, P. Thevenaz, and M. Unser. Variational b-spline level-set: A linear filtering approach for fast, deformable model evolution. *IEEE Transactions on Image Processing*, 18(6):1179–1191, 2009.
- [11] M. S. Bernstein, G. Little, R. C. Miller, B. Hartmann, M. S. Ackerman, D. R. Karger, D. Crowell, and K. Panovich. SoyLent: A word processor with a crowd inside. In *Proceedings of the 23rd annual ACM symposium on User interface software and technology*, pages 313–322, 2010.
- [12] 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.

- [13] A. Carlier, V. Charvillat, A. Salvador, X. Giro i Nieto, and O. Marques. Click'n'Cut: Crowdsourced interactive segmentation with object candidates. In *Proceedings of the 2014 International ACM Workshop on Crowdsourcing for Multimedia*, pages 53–56. ACM, 2014.
- [14] J. Carreira and C. Sminchisescu. Constrained parametric min-cuts for automatic object segmentation. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3241–3248, 2010.
- [15] V. Caselles, R. Kimmel, and G. Sapiro. Geodesic active contours. *IEEE Transactions on Image Processing*, 22(1):61–79, 1997.
- [16] T. Chan and L. Vese. Active contours without edges. *IEEE Transactions on Image Processing*, 10(2):266–277, 2001.
- [17] S. R. Cholleti, S. A. Goldman, A. Blum, D. G. Politte, S. Don, K. Smith, and F. Prior. Veritas: combining expert opinions without labeled data. *International Journal on Artificial Intelligence Tools*, 18(5):633–651, 2009.
- [18] L. P. Coelho, A. Shariff, and R. F. Murphy. Nuclear segmentation in microscope cell images: A hand segmented dataset and comparison of algorithms. In *IEEE International Symposium on Biomedical Imaging: From Nano to Macro (ISBI)*, pages 518–521, 2009.
- [19] S. Dasiopoulou, E. Giannakidou, G. Litos, P. Malasioti, and Y. Kompatsiaris. A survey of semantic image and video annotation tools. In *Knowledge-driven Multimedia Information Extraction and Ontology Evolution*, pages 196–239. Springer, 2011.
- [20] T. Dietenbeck, M. Alessandrini, D. Friboulet, and O. Bernard. Creaseg: A free software for the evaluation of image segmentation algorithms based on level-set. In *IEEE International Conference on Image Processing (ICIP)*, pages 665–668, 2010.
- [21] B. Edwards. *Drawing on the Right Side of the Brain*. ACM, 1997.
- [22] Ilker Ersoy, Filiz Bunyak, Kannappan Palaniappan, Mingzhai Sun, and Gabor Forgacs. Cell spreading analysis with directed edge profile-guided level set active contours. 2008.
- [23] L. Ibanez et al. The itk software guide, 2003.
- [24] S. Kim et al. I'm cell: A touch pad tool for annotating cell images. *Proceedings of the 1st Biomedical Signal Analysis Conference*, 2014.
- [25] A. Fedorov, K. Tuncali, F. M. Fennessy, J. Tokuda, N. Hata, W. M. Wells, R. Kikinis, and C. M. Tempany. Image registration for targeted mri-guided transperineal prostate biopsy. *Journal of Magnetic Resonance Imaging*, 36(4):987–992, 2012.
- [26] L. Galli, P. Fraternali, D. Martinenghi, M. Tagliasacchi, and J. Novak. A draw-and-guess game to segment images. In *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Conference on Social Computing (SocialCom)*, pages 914–917, 2012.

- [27] S. Ghosh, J. J. Pfeiffer, and J. Mulligan. A general framework for reconciling multiple weak segmentations of an image. In *IEEE Workshop on Applications of Computer Vision (WACV)*, 2009. 8 pp.
- [28] R. J. Giuly, K. Kim, and M. H. Ellisman. DP2: Distributed 3D image segmentation using micro-labor workforce. *Bioinformatics*, 29(10):1359–1360, 2013.
- [29] B. M. Good and A. I. Su. Crowdsourcing for bioinformatics. In *Bioinformatics*, volume 29, pages 1925–1933, 2013.
- [30] V. Gulshan, C. Rother, A. Criminisi, A. Blake, and A. Zisserman. Geodesic star convexity for interactive image segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3129–3136, 2010.
- [31] D. Gurari, S. Jain, K. Grauman, and M. Betke. Pull the plug? predicting if computers or humans should segment images. In *IEEE International Conference on Computer Vision (ICCV)*, 2015 (Under Review).
- [32] D. Gurari, S. Kim, E. Yang, B. Isenberg, T. Pham, A. Purwada, P. Solski, M. Walker, J. Y. Wong, and M. Betke. SAGE: An approach and implementation empowering quick and reliable quantitative analysis of segmentation quality. In *Proceedings of the IEEE Workshop on Applications in Computer Vision (WACV)*, pages 475–481, January 2013. 7 pp.
- [33] D. Gurari, M. Sameki, and M. Betke. Crowdsourcing tasks: Domain expertise helps and hurts. In *Computer Human Interaction (CHI)*, 2016 (In Preparation).
- [34] D. Gurari, M. Sameki, Z. Wu, and M. Betke. Utilizing crowdsourcing and algorithms to find boundaries of objects in biomedical and everyday images. *Computer Human Interaction (CHI)*, 2016 (In Preparation).
- [35] D. Gurari, D. Theriault, and M. Betke. Informed segmentation: A framework for using context to select an algorithm and a case study using humans in the loop. *Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI): Interactive Medical Image Computation (IMIC) Workshop*, page 9 pp., 2014.
- [36] D. Gurari, D. Theriault, M. Sameki, and M. Betke. 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. *Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI): Interactive Medical Image Computation (IMIC) Workshop*, page 9 pp., 2014.
- [37] D. Gurari, D. Theriault, M. Sameki, B. Isenberg, T. A. Pham, A. Purwada, P. Solski, M. Walker, C. Zhang, J. Y. Wong, and M. Betke. How to collect segmentations for biomedical images? A benchmark evaluating the performance of experts, crowd-sourced non-experts, and algorithms. *IEEE Winter conference on Applications in Computer Vision (WACV)*, page 8 pp., 2015.

- [38] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The weka data mining software: an update. *11(1):10–18*, 2009.
- [39] K. Hara, V. Le, and J. Froehlich. Combining crowdsourcing and Google street view to identify street-level accessibility problems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 631–640. ACM, 2013.
- [40] B. Hariharan, P. Arbel aez, R. Girshick, and J. Malik. Simultaneous detection and segmentation. In *Computer Vision–ECCV*, pages 297–312, 2014.
- [41] L. He and et al. A comparative study of deformable contour methods on medical image segmentation. *Image Vision Comput*, 26(2):141–163, 2008.
- [42] M. Helmstaedter, K. L. Briggman, S. C. Turaga, V. Jain, H. S. Seung, and W. Denk. Connectomic reconstruction of the inner plexiform layer in the mouse retina. *Nature*, 500:168–174, 2013.
- [43] D. P. Huttenlocher, G. A. Klanderman, and W. J. Rucklidge. Comparing images using the Hausdorff distance. *IEEE Transactions on Pattern Analysis and Machine Learning*, 15(9):850–863, 1993.
- [44] S. D. Jain and K. Grauman. Predicting sufficient annotation strength for interactive foreground segmentation. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1313–1320. IEEE, 2013.
- [45] Jinseop S Kim, Matthew J Greene, Aleksandar Zlateski, Kisuk Lee, Mark Richardson, Srinivas C Turaga, Michael Purcaro, Matthew Balkam, Amy Robinson, Bardia F Behabadi, et al. Space-time wiring specificity supports direction selectivity in the retina. *Nature*, 509(7500):331–336, 2014.
- [46] T. Kohlberger, V. Singh, C. Alvino, C. Bahlmann, and L. Grady. Evaluating segmentation error without ground truth. *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pages 528–536, 2012.
- [47] A. S. Krupnick, V. K. Tidwell, J. A. Engelbach, V. V. Alli, A. Nehorai, M. You, H. G. Vikis, A. E. Gelman, D. Kreisel, and J. R. Garbow. Quantitative monitoring of mouse lung tumors by magnetic resonance imaging. *Nature Protocols*, 7(1):128–142, 2012.
- [48] S. Lankton and A. Tannenbaum. Localizing region-based active contours. *IEEE Transactions on Image Processing*, 17(11):2029–2039, 2008.
- [49] M. Larson, M. Melenhorst, M. Menéndez, and P. Xu. Using crowdsourcing to capture complexity in human interpretations of multimedia content. In *Fusion in Computer Vision*, pages 229–269. Springer, 2014.
- [50] L. J. Latecki. Mpeg-7 core experiment ce-shape-1 dataset. <http://www.dabi.temple.edu/~shape/MPEG7/dataset.html>, January 2015.

- [51] E. Law and H. Zhang. Towards large-scale collaborative planning: Answering high-level search queries using human computation. In *AAAI Conference on Artificial Intelligence*, page 6 pp., 2011.
- [52] Matthew Lease. On quality control and machine learning in crowdsourcing. In *Proceedings of the AAAI Workshop on Human Computation (HCOMP)*, 2011.
- [53] C. Li, C. Y. Kao, J. C. Gore, and Z. Ding. Minimization of region-scalable fitting energy for image segmentation. *IEEE Transactions on Image Processing*, 17(10):1940–1949, 2008.
- [54] K. Li, E. D. Miller, M. Chen, T. Kanade, L. E. Weiss, and P. G. Campbell. Cell population tracking and lineage construction with spatiotemporal context. *Medical Image Analysis*, 12(5):546–566, 2008.
- [55] T. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, and C. L. Zitnick. Microsoft COCO: Common objects in context. *IEEE European Conference on Computer Vision (ECCV)*, pages 740–755, 2014.
- [56] C. J. Lintott, K. Schawinski, A. Slosar, K. Land, S. Bamford, D. Thomas, M. J. Raddick, R. C. Nichol, A. Szalay, D. Andreescu, et al. Galaxy zoo: Morphologies derived from visual inspection of galaxies from the sloan digital sky survey. *Monthly Notices of the Royal Astronomical Society*, 389(3):1179–1189, 2008.
- [57] J. Little, A. Abrams, and R. Pless. Tools for richer crowd source image annotations. In *Applications of Computer Vision (WACV), 2012 IEEE Workshop on*, pages 369–374, 2012.
- [58] D. Liu, Y. Xiong, K. Pulli, and L. Shapiro. Estimating image segmentation difficulty. In *Machine Learning and Data Mining in Pattern Recognition*, pages 484–495, 2011.
- [59] L. Liu and S. Sclaroff. Medical image segmentation and retrieval via deformable models. *International Conference on Image Processing*, 3:1071–1074, 2001.
- [60] M. A. Luengo-Oroz, A. Arranz, and J. Frean. Crowdsourcing malaria parasite quantification: An online game for analyzing images of infected thick blood smears. *Journal of Medical Internet Research*, 14(6), 2012.
- [61] A. Mao, E. Kamar, Y. Chen, E. Horvitz, M. E. Schwamb, C. J. Lintott, and A. M. Smith. Volunteering versus work for pay: Incentives and tradeoffs in crowdsourcing. In *First AAAI Conference on Human Computation and Crowdsourcing*, pages 94–102, 2013.
- [62] W. Mason and S. Suri. Conducting behavioral research on Amazon’s Mechanical Turk. *Behavior research methods*, 44(1):1–23, 2012.
- [63] MATLAB. The Mathworks, Inc., Natick, MA.
- [64] E. Meijering. Cell segmentation: 50 years down the road. *IEEE Signal Processing Magazine*, 29(5):140–145, 2012.

- [65] C. R. Meyer, T. D. Johnson, G. McLennan, D. R. Aberle, E. A. Kazerooni, H. MacMahon, B. F. Mullan, D. F. Yakelevitz, E. J. R. van Beek, S. G. Armato, M. F. McNitt-Gray, A. P. Reeves, D. Gur, C. I. Henschke, E. A. Hoffman, P. H. Bland, G. Laderach, R. P. D. Qing, C. Piker, J. Guo, A. Starkey, D. Max, B. Y. Croft, and L. P. Clarke. Evaluation of lung MDCT nodule annotation across radiologists and methods. *Academic Radiology*, 13(10):1254–1265, 2006.
- [66] B. Moller and S. Posch. Comparing active contours for the segmentation of biomedical images. *International Symposium on Biomedical Imaging*, pages 736–739, 2012.
- [67] T. B. Nguyen, S. Wang, V. Anugu, N. Rose, M. McKenna, N. Petrick, J. E. Burns, and R. M. Summers. Distributed human intelligence for colonic polyp classification in computer-aided detection for ct colonography. *Radiology*, 262(3):824–833, 2012.
- [68] N. Otsu. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1):62–66, 1979.
- [69] S. Park and H. Kautz. Privacy-preserving recognition of activities in daily living from multi-view silhouettes and rfid-based training. *AAAI Fall Symposium: AI in Eldercare: New Solutions to Old Problems*, pages 70–77, 2008.
- [70] A. Pinidiyaarachchi and C. Wahlby. Seeded watersheds for combined segmentation and tracking of cells. *Image Analysis and Processing - ICIAP*, pages 336–343, 2005.
- [71] G. Quattrone, L. Capra, and P. D. Meo. There’s no such thing as the perfect map: Quantifying bias in spatial crowd-sourcing datasets. In *Proceedings of CSCW*. ACM, 2015.
- [72] A. J. Quinn and B. B. Bederson. Human computation: A survey and taxonomy of a growing field. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1403–1412, 2011.
- [73] W. Rasband. ImageJ, U.S. National Institutes of Health, Bethesda, Maryland, USA. <http://imagej.nih.gov/ij/>, January 2015.
- [74] T. R. Raviv, Y. Gao, J. J. Levitt, and S. Bouix. Statistical shape analysis of neuroanatomical structures via level-set-based shape morphing. *SIAM Journal on Imaging Sciences*, 7(3):1645–1668, 2014.
- [75] X. Ren and J. Malik. Learning a classification model for segmentation. In *Ninth IEEE International Conference on Computer Vision*, pages 10–17, 2003.
- [76] T. Riklin-Raviv, N. Kiryati, and N. Sochen. Prior-based segmentation and shape registration in the presence of perspective distortion. *International Journal of Computer Vision*, 72(3):309–328, 2007.
- [77] A. Rizk, G. Paul, P. Incardona, M. Bugarski, M. Mansouri, A. Niemann, U. Ziegler, P. Berger, and I. F. Sbalzarini. Segmentation and quantification of subcellular structures in fluorescence microscopy images using squassh. *Nature Methods*, 9(3):586–596, 2014.

- [78] C. Rother, V. Kolmogorov, and A. Blake. Grabcut: Interactive foreground extraction using iterated graph cuts. *ACM Transactions on Graphics (TOG)*, 3(309–314), 2004.
- [79] 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.
- [80] J. Schindelin, I. Arganda-Carreras, E. Frise, V. Kaynig., M. Longair, T. Pietzsch, S. Preibisch, C. Rueden, S. Saalfeld, B. Schmid, J. Tinevez, D. J. White, V. Hartenstein, K. Eliceiri, P. Tomancak, and A. Cardona. Fiji: an open-source platform for biological-image analysis. *Nature Methods*, 9(7):676–682, 2012.
- [81] S. Sclaroff. Deformable prototypes for encoding shape categories in image databases. *Pattern Recognition*, 30(4):627–641, 1997.
- [82] A. Sheshadri and M. Lease. Square: A benchmark for research on computing crowd consensus. *AAAI Conference on Human Computation and Crowdsourcing (HCOMP)*, 2013.
- [83] 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.
- [84] M. D. Smucker, J. Allan, and B. Carterette. A comparison of statistical significance tests for information retrieval evaluation. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 623–632, 2007.
- [85] A. Sorokin and D. Forsyth. Utility data annotation with Amazon Mechanical Turk. *Urbana*, 51(61):820, 2008.
- [86] H. Su, J. Deng, and L. Fei-Fei. Crowdsourcing annotations for visual object detection. In *Workshops at the Twenty-Sixth AAAI Conference on Artificial Intelligence*, 2012.
- [87] D. H. Theriault, M. Walker, J. Y. Wong, and M. Betke. Cell morphology classification and clutter mitigation in phase-contrast microscopy images using machine learning. *Machine Vision and Applications*, 23(4):659–673, 2012.
- [88] 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. *Computerized Medical Imaging and Graphics*, 30(2):75–87, 2006.
- [89] E. R. Velazquez, C. Parmar, M. Jermoumi, R. H. Mak, A. van Baardwijk, F. M. Fennessy, J. H. Lewis, D. De Ruyscher, R. Kikinis, P. Lambin, et al. Volumetric ct-based segmentation of nsclc using 3d-slicer. *Scientific Reports*, 3, 2013.
- [90] S. Vijayanarasimhan and K. Grauman. What’s it going to cost you?: Predicting effort vs. informativeness for multi-label image annotations. In *Conference on Computer Vision and Pattern Recognition*, pages 2262–2269, 2009.

- [91] 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.
- [92] S. Vittayakorn, K. Yamaguchi, A. C. Berg, and T. L. Berg. Runway to realway: Visual analysis of fashion. *IEEE Winter conference on Applications in Computer Vision (WACV)*, page 8 pp., 2015.
- [93] C. Wählby, L. Kamentsky, Z. H. Liu, T. Riklin-Raviv, A. L. Conery, E. J. O’Rourke, K. L. Sokolnicki, O. Visvikis, V. Ljosa, J. E. Irazoqui, P. Golland, G. Ruvkun, F. M. Ausubel, and A. E. Carpenter. An image analysis toolbox for high-throughput c. elegans assays. *Nature Methods*, 9(7):714–716, 2012.
- [94] T. Walter, D.W. Shattuck, R. Baldock, M. Bastin, A. E. Carpenter, S. Duce, J. Ellenberg, A. Fraser, N. Hamilton, S. Pieper, M. A. Ragan, J. E. Schneider, P. Tomancak, and J. Heriche. Visualization of image data from cells to organisms. *Nature Methods Supplement*, 7(S26–S41), 2010.
- [95] 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.
- [96] M. Yuen, I. King, and K. Leung. A survey of crowdsourcing systems. In *Privacy, security, risk and trust (passat), 2011 IEEE third international conference on and 2011 IEEE third international conference on social computing (socialcom)*, pages 766–773, 2011.
- [97] Y. J. Zhang. A survey on evaluation methods for image segmentation. *Pattern Recognition*, 29(8):1335–1346, 1996.

Curriculum Vitae

