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

Danna Gurari¹, Diane Theriault¹, Mehnoosh Sameki¹, Brett Isenberg², Tuan A. Pham², Alberto Purwada², Patricia Solski², Matthew Walker³, Chentian Zhang², Joyce Y. Wong², Margrit Betke¹
¹Department of Computer Science; ²Department of Biomedical Engineering; ³Department of Biology

Motivation

Biology: Relationship between shape and function?
Question: How to extract accurate object boundaries from images?
Key Challenges:
* Which annotation method?
* How to evaluate an annotation method? (no references for “images in the wild”)

Image Library

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Phase Contrast</td>
<td>Rat smooth muscle cells</td>
<td>40x</td>
<td>35,613</td>
<td>0.15</td>
<td>64</td>
<td>61</td>
<td>35</td>
</tr>
<tr>
<td>2</td>
<td>Phase Contrast</td>
<td>Rabbit smooth muscle cells</td>
<td>10x</td>
<td>10,963</td>
<td>0.29</td>
<td>52</td>
<td>50</td>
<td>69</td>
</tr>
<tr>
<td>3</td>
<td>Phase Contrast</td>
<td>Fibroblasts</td>
<td>10x</td>
<td>3,937</td>
<td>0.53</td>
<td>58</td>
<td>50</td>
<td>47</td>
</tr>
<tr>
<td>4</td>
<td>Fluorescence</td>
<td>Lu melanoma cells</td>
<td>10x</td>
<td>836</td>
<td>0.53</td>
<td>48</td>
<td>17</td>
<td>61</td>
</tr>
<tr>
<td>5</td>
<td>Fluorescence</td>
<td>WM993 melanoma cells</td>
<td>10x</td>
<td>1,119</td>
<td>0.71</td>
<td>54</td>
<td>19</td>
<td>58</td>
</tr>
<tr>
<td>6</td>
<td>MRI</td>
<td>Rabbit aorta</td>
<td>10x</td>
<td>216</td>
<td>0.94</td>
<td>25</td>
<td>42</td>
<td>35</td>
</tr>
</tbody>
</table>

Algorithms: Lankton level set, Shi level set, Chan Vese level set, seeded watershed, Hough Transform, Otsu thresholding
Crowdsourcing: LabelMe & Mechanical Turk

Method Evaluation

\[ J(A, B) = \frac{|A \cap B|}{|A \cup B|} \]

Gold Standard (A)
Annotation (B)

Crowd consensus statistically similar to experts!

Expert performance differs, especially for different datasets

Crowd consensus exceeds performance of individual crowd workers

Algorithm performance varies widely, especially for different datasets

Method Comparison

Fused Segmentation: J(A,B) = 0
J(A,B) > 0
J(A,B) --> 1

Acknowledgments
NSF Grants: IIS-1421943, IIS-0910908