EVALUATION AND BENCHMARK FOR BIOLOGICAL IMAGE SEGMENTATION

Elisa Drelie Gelasca, Jiyun Byun, Boguslaw Obara, B.S. Manjunath

Center for Bio-Image Informatics, Electrical and Computer Engineering Department,
University of California, Santa Barbara 93106-9560,
http://www.bioimage.ucsb.edu

ABSTRACT

This paper describes ongoing work on creating a benchmarking and validation dataset for biological image segmentation. While the primary target is biological images, we believe that the dataset would be of help to researchers working in image segmentation and tracking in general. The motivation for creating this resource comes from the observation that while there are a large number of effective segmentation methods available in the research literature, it is difficult for the application scientists to make an informed choice as to what methods would work for her particular problem. No one single tool exists that is effective on a diverse set of application contexts and different methods have their own strengths and limitations. We describe below three different classes of data, ranging in scale from subcellular to cellular to tissue level images, each of which pose their own set of challenges to image analysis. Of particular value to the image processing researchers is that the data comes with associated *ground truth* information that can be used to evaluate the effectiveness of different methods. The analysis and evaluation are also integrated into a database framework that is available online at *http://dough.ece.ucsb.edu*.

Index Terms— Standardized dataset, biological images, segmentation evaluation, ground truth, subcell, cell, tissue.

1. INTRODUCTION

The collection of a standard dataset is a critical first step in evaluating and benchmarking new technologies. This is particularly important in image processing methods that have a wide range of applications ranging from biology to remote sensing. Once the benchmark is set up, the current state-of-the-art image analysis methods can be tested with evaluation measures appropriate to the specific application. This is the main motivation in building the UCSB benchmark dataset for bioimaging application. In particular, we describe in the following datasets at different scales with carefully generated manual ground truths that could be of significant help to not only researchers in biology who have the need for image segmentation/tracking tools but also to the image processing community in general.

This research is supported by NSF ITR-0331697.

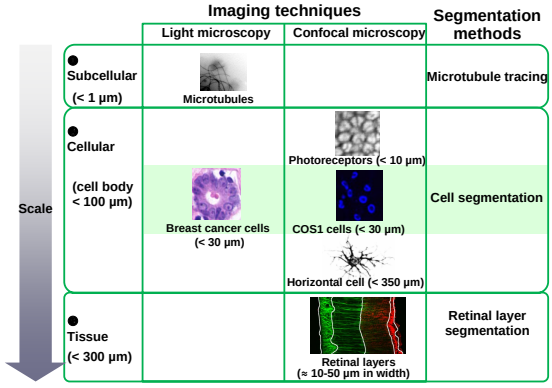


Fig. 1. Example dataset provided in the benchmark.

In recent years, there have been a few successful benchmarking efforts in image analysis and computer vision, such as the face recognition dataset [1], Berkeley segmentation dataset for natural images [2], macrobiological structures such as mammorgams and MRI images [3], and the database collected by the National Institutes of Health[1]. Recently, there have been some efforts in creating microbiological image benchmark such as the cell center database (http://www.ccdb.ucsd.edu) and mouse retina database (http://www.cepko.med.harvard.edu). However, these benchmark do not yet include different scale data, ground truth and/or image analysis tools. This was highlighted at a recent panel involving benchmarking and validation of computer vision methods in biology [2].

In this work, we describe a benchmark for biological images. The datasets include collections with well defined ground truth. We provide representative datasets of microbiological structures whose scales range from a subcellular level (nm) to a tissue level (μm). The collections are obtained through a collaborations with domain scientists in molecular, cellular and developmental biology, and highlight some of the current challenges at these varying spatial scales for image segmentation (Fig. 1).

[1] http://www.nibib.nih.gov/Research/Resources/ImageClinData.
[2] http://www.ece.ucsb.edu/bioimage/workshop2008/program.html

Table 1. Dataset and ground truth in the benchmark.

Type	Microtubule	Cell Nuclei	Retina
# images	9 stacks	888 images	343 images
size (pixel)	512×600	512×512 (also 768×512)	300×200 (also 768×512)
format	.tiff .stk	.tiff	.bmp .tiff
channels	Rhodamine	TO-PRO	Rod photoreceptors (anti-rod opsin; red) Microglia (isolectin B4; green) Muller cells (anti-GFAP; blue)
condition	Taxol/Docetaxel treated	normal 3-day detached	normal, 1-day, 3-day 7-day, 28-day detached
species	human (HUVEC)	cat	cat
ground truth	1374 traces of microtubules (4 experts)	manual cell count (3 experts) 40 ONL masks	91 layer masks 108 boundary masks

2. DATASET, GROUND TRUTH, ANALYSIS TOOLS

Benchmark images are acquired through two of the most common microscopic imaging techniques: transmitted light microscopy and confocal laser scanning microscopy. The collected images are an effort to standardize the dataset for microbiological structures and present common challenges in segmenting them. The challenges to segmentation include [4]: inhomogeneous illumination across visual fields, occlusion of objects, variation in object shape, size and orientation and considerable variation of the signal intensity of objects from the same class. Table 1 summarizes the image collections at each of these varying spatial resolutions.

The benchmark also includes the analysis tools that are designed to obtain different quantitative measures from the dataset such as microtubule tracing, cell segmentation, and retinal layer segmentation.

Additionally, in the proposed benchmark, *ground truth* is manually created by *experts* from part of each dataset. In the following, we explain the dataset, ground truth and image analysis tools available in the benchmark at different scales (see Table 1). In Sec. 3, we describe evaluation methods provided to assess the performance of the integrated analysis tools.

2.1. Subcellular level

Microtubules are conveyer belts inside the cells. They move vesicles, granules, organelles like mitochondria, and chromosomes via special attachment proteins. Structurally, they are linear polymers of tubulin which is a globular protein. Researchers believe microtubules play a important role in the study of Alzheimer and in certain cancers. To obtain a quantitative description of behavior under different experimental conditions, researchers track individual microtubule traces manually as shown in red in Fig.2. We focus here on microtubule time sequence images obtained by transmitted light microscopes. The challenges at this scale and with this acquisition modality are typical for *in-vivo* cell imaging: high clutter, gaps, and low signal to noise ratio.

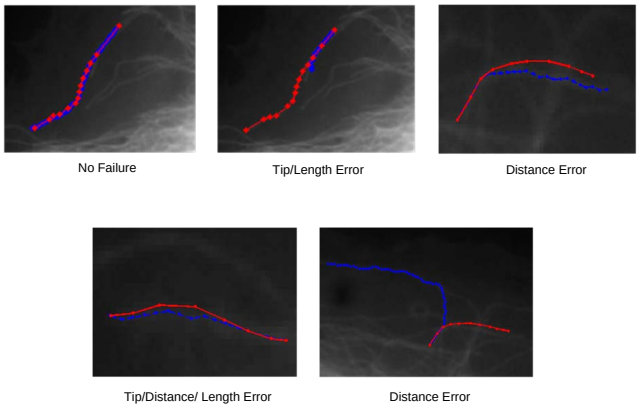


Fig. 2. Microtubule tracing examples. The blue traces are automatically obtained by the algorithm proposed in [5] and the red line are ground truth traces.

The tracking of the microtubule free ends allows biologists to compute the growth and shortening statistics (of the microtubules) which in turn are related to the presence of key proteins such Tau and its interaction with various drugs. Understanding the dynamics of the microtubules under different experimental conditions is important in the study of several neuro-degenerative diseases (e.g., Alzheimer's) as well as cancer. The manual measurements of these microtubules are very labor intensive and time consuming. Due to the limitations in biological sample preparation and fluorescence imaging, typical images in live cell studies exhibit severe noise and considerable clutter and automatic microtubule tracing becomes a hard task. Our benchmark includes an automatic method [5] for extracting curvilinear structures from live cell fluorescence images. The data also include ground truth for microtubule tip location and microtubule bodies that could be useful for evaluating image segmentation and tracking meth-

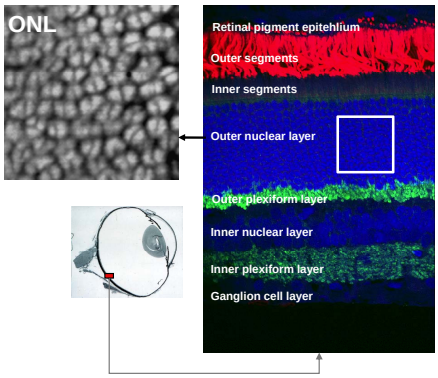


Fig. 3. Retinal layers. Confocal microscopy of a vertical section through a cat retina. Each layer has a different structure which consists of the group of cell bodies or synaptic terminal. The photoreceptor cell bodies comprise the ONL. Left top image shows the ONL at higher magnification (boxed area in right image).

ods.

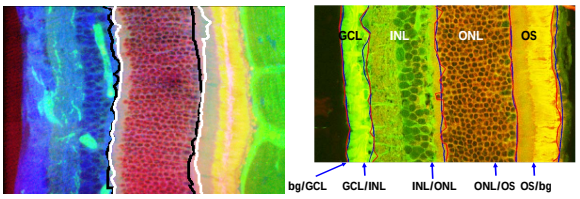
2.2. Cell level

Challenges at the cellular level include large variations in cell shape and staining, intensity variation within and across the images, and clustered objects. In biology, cell addition and cell death are important occurrences in the study of diseases or injuries. Therefore, one of the common tasks is to count the number of cells and nuclei, particularly in histological sections, and characterize various cell attributes such as shape, size, smoothness of the boundary, etc. For example, in retinal images, the number of photoreceptor nuclei in the outer nuclear layer (ONL), depicted in Fig.3, is one of the important measurements of the death and degeneration of the retina. Our retinal dataset consists of 40 confocal images of normal and 3-day detached feline retinas (20 normal and 20 3-day detached). The detached retinal samples are obtained by surgically detaching a retina and leaving the animal in the detached retinal state for one to three days before imaging the tissue samples. Images were collected using a laser scanning confocal microscope from tissue sections. For each image, the ground truth, available in the benchmark, consists of an ONL binary mask and the corresponding manually cell count in ONL by three different experts. A nucleus detector based on a Laplacian of Gaussian filter is integrated into the benchmark (for more detail see [6]).

In addition, there are about 50 histopathology images used in breast cancer cell detection with associated ground truth data available. There are, however, no benchmark methods currently available for performance evaluation on these histopathology images.

2.3. Tissue level

Confocal microscope images of retinas taken during detachment experiments are critical components for understanding the structural and cellular changes of a retina in response to disease and injury. As the first step of any other analysis (e.g.



(a) (b)

Fig. 4. Example of layer segmentation results compared to ground truth for feline retinal images normal condition. (a) ONL segmentation: white boundaries detected by [7]'s method compared to black one (ground truth); (b) layer segmentation: blue boundaries detected by [8]'s method compared to red one (ground truth).

before cell counting), it is crucial to have a reliable map of the retinal layers. Hundreds of retinal images and layer ground truth are part of the benchmark (see Tab.1). Four major layers of the retina are segmented manually: the ganglion cell layer (GCL), the inner nuclear layer (INL), the outer nuclear layer (ONL), and the outer segments (OS).

Two retinal layer segmentation algorithms are integrated in the benchmark. One is a variational segmentation approaches based on pixel pairwise similarities [7]. This method exploits prior information (reference image) and a dissimilarity measure between the pixels of the reference and the pixels of the image that has to be segmented. The other segmentation algorithm integrated in the benchmark uses parametric active contour to detect the boundaries between layers [8]. An example of the segmentation results and ground truth are respectively shown in Fig.4 (a) and (b).

3. EVALUATION

The tasks associated with the above data analysis are very labor intensive. The benchmark segmentation/tracking methods associated with these data are an initial attempt at automating some of these manual tasks. The ground truth data provide a critical reference point to evaluate these automated methods. In some cases, for e.g.-microtubule tracing and tracking, automation not only helps in performing the original computations efficiently but also facilitates quantitative measurements (full body tracking of the microtubules–a task that is very difficult, if not impossible, to do manually) that are otherwise not possible. In order for the scientists to use such methods, however, one needs to provide evaluation metrics that would make a reasonable comparison between the automated results and the "ground truth" data obtained manually. Such performance metrics would also enable comparison of different image processing methods on standardized datasets.

3.1. Microtubule (MT) tracing

We propose the following three measurements to evaluate MT tracing. 1) MT tip distance: Tip distance error is the Eu-

clidean distance between the ground truth tip to the trace tip (i.e. the tip found by the algorithm). 2) MT trace distance: Trace distance error is the average distance from all the points on the ground truth to all the points on the trace. 3) MT length errors: Length difference is simply the difference between the length of the ground truth and the trace. Acceptable threshold for these error measures are set by biologists and tracing algorithm failures occur when: 1) tip distance larger than 0.792 μm; 2) length difference is larger than 0.792 μm ; 3) trace distance (mean) is larger than 0.396 μm;

When we run the integrated tracing algorithm on the microtubule dataset, the failures are on average 0.09%. Examples of failure and no failure are shown in Fig. 2.

3.2. Cell/nucleus detection

A simple evaluation method is integrated to evaluate the approaches that count cells, nuclei, or other objects in sectioned materials. The error, E, in cell counting is measured by the percentage error between manual counts (obtained from three experts) and the result of a nucleus detector as follows:

$$E = \frac{1}{N} \frac{|ND - \overline{GT}|}{\overline{GT}}$$

where N is the number of images in the dataset, ND and $\overline{GT}$ is the number of nuclei detected by the nucleus detector and by the average of manual counting, respectively. When the nucleus detector is applied on this dataset, it correctly counts nuclei within the ONL with an average error 3.52%.

3.3. Retinal layer segmentation

Both common and new evaluation measures are integrated to test the performance of automatic methods for layer and boundary retinal segmentation: 1) distance$_{layer}$ = distance between ground truth and segmented boundaries for each layer obtained using Fast Marching; 2) Precision = the ratio between true positive and automatically detected pixels; 3) Recall (sensitivity) = the ratio between true positive pixels and ground truth 4) 1-sensitivity = the fraction of false positives; 5) F measure= harmonic mean between precision and recall for each layer; 6) weighted F measure = it scores F-measure layers in proportion to their percentage of the total area and sum them all up in order to weight more segmentation errors in larger layers. For example, when applied on the dataset the method [7] gives a F-measure around 0.88%.

4. DISCUSSION AND CONCLUSION

The proposed benchmark provides a unique, publicly available, datasets as well as image analysis tools and evaluation methods for bioimages. The benchmark will help researchers to validate, test and improve their algorithms, and provide biologists a guidance of algorithms' limitations and capabilities. The benchmark is integrated into the Bisque bioimage database infrastructure (*http://dough.ece.ucsb.edu*) at UCSB and all the tools described above can be applied on the proposed dataset. Users are encouraged to upload their bioimages and ground truth, test the analysis tools and perform the evaluation. Moreover, new (user contributed) segmentation algorithms and evaluation methods can be integrated upon request. Dataset and ground truth can be downloaded from our website *http://www.bioimage.ucsb.edu*.

5. ACKNOWLEDGMENTS

We would like to thank Emre Sargin and Jose Freire for their valuable help and Camil Caizon, Eden Haven, Stephanie Perez, Corey Cox and Nicholas Secchini Drelie for providing ground truth. The bioimage datasets are contributed by the Fisher lab (retinal images) and Feinstein-Wilson labs (microtubule data) at UCSB.

6. REFERENCES

[1] P. J. Rauss P. J. Phillips, H. Moon and S. Rizvi, "The FERET evaluation methodology for face recognition algorithms," *IEEE Transactions on PAMI*, vol. 22, no. 10, October 2000.

[2] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *Proc. 8th ICCV*, July 2001, vol. 2, pp. 416–423.

[3] T. W. Nattkemper, T. Twellmann, W. Schubert, and H. Ritter, "Human vs. machine: Evaluation of fluorescence micrographs," *Computers in Biology and Medicine*, vol. 33, no. 1, pp. 31–43, 2003.

[4] T. W. Nattkemper, "Automatic segmentation of digital micrographs: A survey," in *MEDINFO*, San Francisco, USA, 2004, AMIA/IMIA.

[5] M. E. Sargin, A. Altinok, K. Rose, and B.S. Manjunath, "Deformable trellis: Open contour tracking in bio-image sequences," in *ICASSP*, April 2008.

[6] J. Byun, M. R. Verardo, B. Sumengen, G. P. Lewis, B. S. Manjunath, and S. K. Fisher, "Automated tool for the detection of cell nuclei in digital microscopic images: Application to retinal images," *Molecular Vision*, vol. 12, pp. 949–960, Aug 2006.

[7] L. Bertelli, J. Byun, and B.S. Manjunath, "A variational approach to exploit prior information in object-background segregation: Application to retinal images," in *ICIP*, Sep 2007, pp. VI–61–VI–64.

[8] N. Vu, , P. Ghosh, and B.S Manjunath, "Retina layer segmentation and spatial alignment of antibody expression levels," in *ICIP*, Sep 2007, pp. II –421–II – 424.