BENCHMARK FOR EVALUATING BIOLOGICAL IMAGE ANALYSIS TOOLS

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

Biological images are critical components for a detailed understanding of the structure and functioning of cells and proteins. Image processing and analysis tools increasingly play a significant role in better harvesting this vast amount of data, most of which is currently analyzed manually and qualitatively. A number of image analysis tools have been proposed to automatically extract the image information. As the studies relying on image analysis tools have become widespread, the validation of these methods, in particular, segmentation methods, has become more critical. There have been very few efforts at creating benchmark datasets in the context of cell and tissue imaging, while, there have been successful benchmarks in other fields, such as the Berkeley segmentation dataset [1], the handwritten digit recognition dataset MNIST [2] and face recognition dataset [3, 4]. In the field of biomedical image processing, most of standardized benchmark data sets concentrates on macrobiological images such as mammograms and magnet resonance imaging (MRI) images [5], however, there is still a lack of a standardized dataset for microbiological structures (e.g. cells and tissues) and it is well known in biomedical imaging [5].

We propose a benchmark for biological images to: 1) provide image collections with well defined *ground truth*; 2) provide image analysis tools and evaluation methods to compare and validate analysis tools. We include a representative dataset of microbiological structures whose scales range from a subcellular level (nm) to a tissue level (μm), inheriting intrinsic challenges in the domain of biomedical image analysis (Fig. 1). The dataset is acquired through two of the main microscopic imaging techniques: transmitted light microscopy and confocal laser scanning microscopy. The analysis tools[1] in the benchmark are designed to obtain different quantitative measures from the dataset including microtubule tracing, cell segmentation, and retinal layer segmentation.

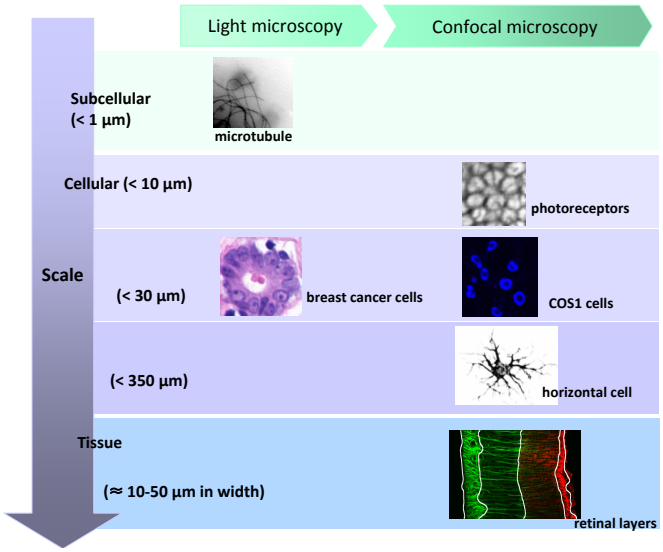

Fig. 1. Example dataset provided in the benchmark.

This research is supported by NSF ITR-0331697.

[1] All analysis tools mentioned in this work can be found at http://www.bioimage.ucsb.edu/publications/.

Additionally, in the proposed benchmark, ground truth is manually created from part of each dataset. Evaluation methods are provided to evaluate the performance of the analysis tools using the ground truth. The benchmark includes standard evaluation measures and *ad hoc* methods designed for specific applications. In the following, we briefly explain the evaluation measures used at various scale level.

Subcellular level. Tracing curvilinear structures is one of the fundamental problems for extracting information about structures such as blood vessels, microtubules, and similar entities. In order to understand microtubule dynamics, biologists study how microtubules grow and shorten by analyzing the stacks of images acquired through transmitted light microscope. Microtubule stacks and corresponding ground truth are part of the benchmark. An automatic method for extracting curvilinear structures from live cell fluorescence images is also integrated in the benchmark. To assess the performance of automated tracing algorithms, four evaluation measures are proposed to compare the tracing result to grount truth: 1) tip distance, 2) trace distance, 3) length difference, and 4) combination of 1), 2) and 3). Tip distance error is the Euclidean distance between the tip position defined by ground truth and that detected by the algorithm. Trace distance error is the average distance between the points on the ground truth and the points on the traced microtubule. Length difference is simply the difference between the length of the ground truth and the traced microtubule. When these errors are satisfied the following conditions which are set by biologists, the tracing algorithm are cosidered as an acceptable method. The conditions are: 1) tip distance smaller than 0.792 μm, 2) length difference is smaller than 0.792 μm, and 3) trace distance (mean) is smaller than 0.396 μm.

Cell level. Cell/nuclei segmentation is the first step of any further analysis of images at cellular level since the resulting counts of cells or nuclei provide a quantitative information crucial to cell viability. A nucleus detection method that count cells, nuclei, or other objects in sectioned materials, is supported in the benchmark framework and used to detect nuclei in the outer nuclear layer (ONL) within retinal images. Retinal images acquired through confocal microscopy and manually counted cells within the ONL are part of the benchmark. A simple evaluation method is integrated to evaluate the nucleus detection method. The error in cell counting is computed by the percentage error between manual counts obtained from three different experts and the result by the nucleus detector.

Tissue level. The retina consists of multiple layer of nerve cells and synapses. Since each layer has a different structure which consists of the group of cell bodies or synaptic terminals, the intact architecture of layers is crucial to retinal function. Layer segmentation simplifies image analysis task for understanding the function of retina before and after injury. Two retinal segmentation methods, confocal retinal images, and ground truth of the layers are supported in the benchmark. Several evaluation measures are integrated to test the performance of two automatic segmentation methods: 1) distance$_{layer}$: averaged distance between ground truth and segmented boundaries of a layer obtained by Fast Marching; 2) Precision: the ratio between the number of true positive and detected pixels; 3) Recall (sensitivity): the ratio between the number of true positive pixels and ground truth, 4) 1-specificity: the fraction of false positive pixels; 5) F measure: harmonic mean of precision and recall for each layer; 6) weighted F measure: a weighted sum of F-measures for each layer. The weight of each layer is determined by its area in proportion to the total area of all layers.

In summary, the benchmark will help researchers to validate, test, and improve their algorithms, and provide biologists a guidance of algorithms' limitations and capabilities. The benchmark can be easily extended to other dataset and analysis tools. The proposed benchmark provides a unique and publicly available datasets as well as image analysis tools and evaluation methods. The benchmark is integrated in Bisquik (http://flour.ece.ucsb.edu:8080/bisquik/) at UCSB.

1. REFERENCES

[1] 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 Int'l Conf. Computer Vision*, July 2001, vol. 2, pp. 416–423.

[2] Y. LeCun, "The mnist database of handwritten digits," *http://yann.lecun.com/exdb/mnist/*.

[3] P. J. Rauss P. J. Phillips, H. Moon and S. Rizvi, "The feret evaluation methodology for face recognition algorithms," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 10, October 2000.

[4] A. Georghiades, P. Belhumeur, and D. Kriegman, "From few to many: Illumination cone models for face recognition under variable lighting and pose," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 6, pp. 643660, 2001.

[5] T. W. Nattkemper, "Automatic segmentation of digital micrographs: A survey," in *Proc. of 11th World Congress on Medical Informatics (MEDINFO)*, San Francisco, USA, 2004, AMIA/IMIA.