Published in final edited form as:
Med Biol Eng Comput. 2008 September ; 46(9): 943–947. doi:10.1007/s11517-008-0380-5.

IICBU 2008 - A Proposed Benchmark Suite for Biological Image Analysis

Lior Shamir[1,*], **Nikita Orlov**[1], **D. Mark Eckley**[1], **Tomasz J. Macura**[1,2], and **Ilya G. Goldberg**[1]

1 *Image Informatics and Computational Biology Unit, Laboratory of Genetics, NIA/NIH. 333 Cassell Dr., Baltimore, MD, 21224*

2 *Computer Laboratory, University of Cambridge, 15 Thomson Avenue, Cambridge, UK*

Abstract

New technology for automated biological image acquisition has introduced the need for effective biological image analysis methods. These algorithms are constantly being developed by pattern recognition and machine vision experts, who tailor general computer vision techniques to the specific needs of biological imaging. However, computer scientists do not always have access to biological image datasets that can be used for computer vision research, and biologist collaborators who can assist in defining the biological questions are not always available. Here we propose a publicly available benchmark suite of biological image datasets that can be used by machine vision experts for developing and evaluating biological image analysis methods. The suite represents a set of practical real-life imaging problems in biology, and offers examples of organelles, cells and tissues, imaged at different magnifications and different contrast techniques. All datasets are available for free download at http://ome.grc.nia.nih.gov/iicbu2008.

Keywords

Biological imaging; image analysis; image datasets; benchmarks

1. Introduction

In the past few years, automated image analysis tools have been becoming increasingly important in the field of biology, and are used in a wide range of applications in biological and medical research. Machinery for automated image acquisition has been providing vast pipelines of biological images of organelles, cells, tissues and full organisms, and the availability of wide bandwidths and large storage devices allow practical use of these images. However, while image acquisition systems can provide very many images, human analysis can be impractically slow. This limitation introduces the need for machine vision algorithms for automated analysis of biological images, which can complete a fully automated process for interpretation of biological experiments.

Computer vision algorithms are typically developed by pattern recognition and signal processing experts. In order to develop, test and evaluate the performance of novel algorithms, computer scientists and engineers must have access to datasets of biological images that represent actual biological questions. However, in many real-life cases machine vision specialists do not have convenient access to biological images, and do not work closely with biologists, who are better able to define the biological problems and direct the research towards

Phone: 410-558-8682, Fax: 410-558-8331, Email: shamirl@mail.nih.gov.

effective and practical solutions. Even when the data are available and well-defined, developers of biological image analysis algorithms often find it difficult to assess the performance of their methods due to the absence of performance figures for previously proposed algorithms using the same data.

While numerous researchers have made their data publicly available [5,6], the data are often not organized in a fashion that can be used conveniently by computer scientists, and performance figures and source codes of previously proposed algorithms are not always available for comparison.

Here we propose IICBU-2008 - a benchmark suite for biological image analysis that provides free access to several carefully chosen biological image datasets. The purpose of this benchmark suite is to make biological image datasets available to computer vision experts, and allow comparative analysis of different algorithms for assessing the performance of newly proposed methods. Each dataset represents an actual real-life biological problem defined by experimental biologists, and includes noise and imperfect images that are normally present in real-life biological datasets.

To facilitate use of our benchmark suite, all data are fully available for free download via the internet, with no registration or license agreement. Compilable source code of the image classification algorithm that was tested using these datasets is also publicly available, and allows scientists to compare their new methods with existing performance figures, as discussed in Section 3.

2. Image Datasets

The image datasets represent a broad range of biological imaging problems, and are based on actual biological experiments. Each dataset features organelles, cells, or tissues, so that the standard types of subjects are represented by at least two datasets. File format of all datasets is TIFF (8 or 16 bits), and the image sizes vary from 25×25 to 1388×1040 pixels. The selected datasets also represent different magnifications with different contrast techniques. The datasets included in IICBU-2008 benchmark suite are listed in Table I, and a more detailed description of each dataset is described below.

2.1. 2D Hela

2D HeLa [1] is a dataset of fluorescence microscopy images of HeLa cells, stained with various organelle-specific fluorescent dyes. The dataset includes 10 probes for intracellular organelles and structures. HeLa cells were stained with dyes (DAPI, MitoTracker, and DiOC6), or antibodies (Giantin, GPP130, Lamp2, Nucleolin, TfR, Actin, and Tubulin).

Automated identification of sub-cellular organelles is important when characterizing newly discovered genes or genes with unknown functions. It is possible to fluorescently tag the protein (s) produced by any given gene, and the ability to identify the organelle where the protein resides provides an important clue to its possible function. It is important to note that human experts have trouble distinguishing Endosomes and Lysosomes, and they also find overlapping Golgi compartments exceedingly difficult to differentiate.

2.2. CHO

CHO [2] is a dataset of fluorescence microscope images of CHO (Chinese Hamster Ovary) cells. The images were taken using five different labels, which are anti-giantin (Golgi), Hoechst 33258 (DNA), anti-lamp2 (Lysosomes), anti-nop4 (nucleoli), and anti-tubulin. This dataset represents the same type of problem as the dataset 2D HeLa (Section 2.1).

2.3. Pollen

The purpose of the dataset *Pollen* [3] is to train a computer program to automatically identify seven classes of pollen grains. This dataset contains small images (25×25 pixels) of seven different types of grains. *Pollen* is a lightweight dataset that represents a relatively simple image classification problem, and can be processed by CPU-consuming computer vision algorithms in a relatively short time.

2.4. RNAi

RNAi is a set of fluorescence microscopy images of fly cells (D. melanogaster) subjected to a set of gene-knockdowns using RNAi. The cells are stained with DAPI to visualize their nuclei. Each class contains 1024×1024 images of the phenotypes resulting from knockdown of a particular gene. Ten genes were selected, and their gene IDs are used as class names. The genes are CG1258, CG3733, CG3938, CG7922, CG8114, CG8222, CG 9484, CG10873, CG12284, CG17161.

The images were acquired automatically using a DeltaVision light microscope with a 60× objective. Each image is produced by deconvolution, followed by maximum intensity projection (MIP) of a stack of 11 images at different focal planes.

Automated analysis of this dataset can focus not only on the classification of the different gene classes, but also on assessing the similarities between the different phenotypes. Measuring and quantifying phenotype similarities can be a matter of considerable importance since different genes may be part of the same cellular mechanism, and therefore may produce phenotypes that are more similar to each other than phenotypes resulting from knocking down other genes. This type of analysis can be used for finding similarities between genes based on the phenotypes that they produce, in contrast to finding similarities using sequence analysis.

2.5. Binucleate

The purpose of the *Binucleate* dataset is to distinguish binucleate cellular phenotypes from normal mononucleate cells. The binucleate phenotype signals a failure in cell division, and is a common phenotypic target when screening for genes or compounds that affect cytokinesis, such as CG1258. Notably, a large proportion of chemotheraputic agents used for cancer treatment interfere with cell division as their primary mode of action.

The images feature cells from D. melanogaster, and were acquired at 60× using a fluorescence microscope and a fluorescent dye that binds DNA (DAPI). The images were acquired robotically as part of a high-throughput screen, so there is no human-based quality control. The *binucleate* dataset represents the same type of imaging problem as RNAi, which is automatic phenotype classification, but the data is many times simpler, and therefore classification accuracy can be higher.

2.6. Lymphoma

Malignant lymphoma is a cancer affecting lymph nodes. Three types of malignant lymphoma are represented in the set: CLL (chronic lymphocytic leukemia), FL (follicular lymphoma), and MCL (mantle cell lymphoma). The ability to distinguish classes of lymphoma from biopsies sectioned and stained with Hematoxylin/Eosin (H&E) can allow more consistent and less labor-consuming diagnosis of this disease. Since only the most knowledgeable and experienced expert pathologists specializing in these types of lymphomas are able to consistently and accurately classify these three lymphoma types using H&E-stained biopsies, automating this classification may also be considered as an important practical application of medical imaging.

The dataset is a collection of samples prepared by different histologists at a number of hospitals, so there is a large degree of staining variation that one would normally expect in real-life data of this kind. This dataset demonstrates a difficult classification problem of a real-life medical application (automated diagnosis of lymphoma).

2.7. Liver Gender (Caloric Restriction)

The Liver Gender dataset features microscopy images of tissues, acquired from 30 six month old male and female mice on a caloric restriction diet. Livers were extracted from sacrificed mice, sectioned, stained with hematoxylin and eosin (H&E), and imaged by a bright-field microscope. Fifty color images per liver were manually acquired using a Carl Zeiss Axiovert 200 microscope and 40× objective. Data was originally collected from 48 livers, but 18 livers were not suitable for imaging and were therefore excluded from the dataset. All images are 1388×1040 RGB TIFFs, with 12 bits of quantization per color channel.

In this benchmark, the data are used to define a binary image classification problem, which is the classification of the mouse gender based on the liver images. The purpose of this dataset is to test automatic analysis of tissue images.

2.8. Liver Gender (Ad-libitum)

The structure and data acquisition method of this dataset is similar to the gender classification of livers on caloric restriction diet. However, the ad-libitum diet of the mice resulted in less uniform liver images of the two genders, and provided a harder image classification problem than the liver gender under caloric restriction.

2.9. Liver Age

The liver age dataset contains images of female mice of four ages – 1, 6, 16, and 24 months, on ad-libitum diet. All slides were H&E stained by the same person in order to minimize the staining variability. The purpose of this benchmark is to test whether a computer program can associate images of mouse livers with one of four age classes. The continuous nature of the aging process also allows measuring the similarities between the different age classes, because similarity between the classes is expected to decrease as the difference between the ages gets larger.

3. Classification Accuracy Using WND-CHARM

Benchmark classification accuracies were computed in order to provide an estimation of the difficulty of each problem, and can be used to compare the performance newly developed algorithms with existing performance figures. The classification of all datasets was performed using the WND-CHARM [7,8] multi-purpose image classification method, which is designed to analyze a wide range of biological image datasets. The classification accuracy achieved for each dataset is given by Figure I.

Several of the image datasets were classified with very high accuracy - *Pollen*, *Binucleate*, and *Liver Gender (CR)*. Other datasets such as *HeLa*, *Lymphoma* and *RNAi* were classified with an accuracy range of 80-85%, reflecting a higher degree of complexity. The lowest performance was recorded on the *Liver Age* dataset, which was classified with accuracy of 51%.

Since there is no “typical” biological experiment, there is also no “typical” biological image dataset, so that a proposed image analysis method that can effectively serve biologists should be able to handle different types of biological data. Therefore, we suggest that the efficacy of

novel biological image analysis methods should be evaluated using several different biological datasets.

C and Matlab source code of WND-CHARM can be downloaded freely as part of Open Microscopy Environment [4] via CVS at http://www.openmicroscopy.org (recommended), or as a “tarball” at http://www.phy.mtu.edu/~lshamir/downloads/ImageClassifier

4. Discussion

Here we described a publicly available biological imaging benchmark suite that contains datasets of several real-life experiments. Since the datasets represent actual biological data, they include also imperfect images and background noise, which is practically unavoidable in many cases of high content screening applications.

The benchmark suite currently does not cover the entire range of all standard biological datasets, and we hope that other biologists will provide additional datasets for upload to this extensible resource. By making this work publicly available we hope to initiate an on-going effort, which will evolve incrementally and adjust itself to the current and future needs of biological image analysis and high content screening. The datasets are available for free download via the world wide web at http://ome.grc.nia.nih.gov/iicbu2008, and no registration or license agreement are required. We encourage investigators who own datasets of biological images to contribute their data, and make them available to the scientific community through this benchmark suite.

Acknowledgements

This research was supported by the Intramural Research Program of the NIH, National Institute on Aging. The datasets *Hela* and *CHO* are from Murphy lab, CMU. The dataset *Pollen* was contributed by Andrew Duller, Henry Lamb and Ian France, and the dataset *Binucleate* was provided by Aaron Straight, Stanford U. We would also like to thank Cathy Wolkow and Wendy Iser for their assistance with the acquisition and definition of the *Terminal Bulb Aging and C. elegans Muscle Aging* datasets, and Elaine Jaffe for providing the data for the Lymphoma dataset.

References

1. Boland MV, Murphy RF. A neural network classifier capable of recognizing the patterns of all major subcellular structures in fluorescence microscope images of HeLa cells. Bioinformatics 2001;17:1213–1223. [PubMed: 11751230]
2. Boland MV, Markey MK, Murphy RF. Automated recognition of patterns characteristic of subcellular structures in fluorescence microscopy images. Cytometry 1998;33:366–375. [PubMed: 9822349]
3. Duller AWG, Duller GAT, France I, Lamb HF. A pollen image database for evaluation of automated identification systems. Quaternary Newsletter 1997;89:4–9.
4. Goldberg IG, Allan C, Burel JM, Creager D, Falconi A, Hochheiser H, Johnston J, Mellen J, Sorger PK, Swedlow JR. The Open Microscopy Environment (OME) Data Model and XML file: open tools for informatics and quantitative analysis in biological imaging. Genome Biology 2005;6:R47. [PubMed: 15892875]
5. Martone ME, Zhang S, Gupta A, Qian X, He H, Price DL, Wong M, Santini S, Ellisman MH. The cell-centered database: a database for multiscale structural and protein localization data from light and electron microscopy. Neuroinformatics 2003;1:379–375. [PubMed: 15043222]
6. Martone ME, Sargis J, Tran J, Wong WW, Jiles H, Mangir C. Database resources for cellular electron microscopy. Methods Cell Biol 2007;79:799–822. [PubMed: 17327184]
7. Orlov N, Shamir L, Macura TJ, Johnston J, Goldberg IG. WND-CHARM: Multi purpose image classification using compound transforms. Pattern Recognition Letters 2008;29:1684–1693.
8. Shamir L, Orlov N, Eckley DE, Macura TJ, Johnston J, Goldberg IG. Wndchrm – an open source utility for biological image analysis. BMC Source Code for Biology and Medicine 2008;13:3.

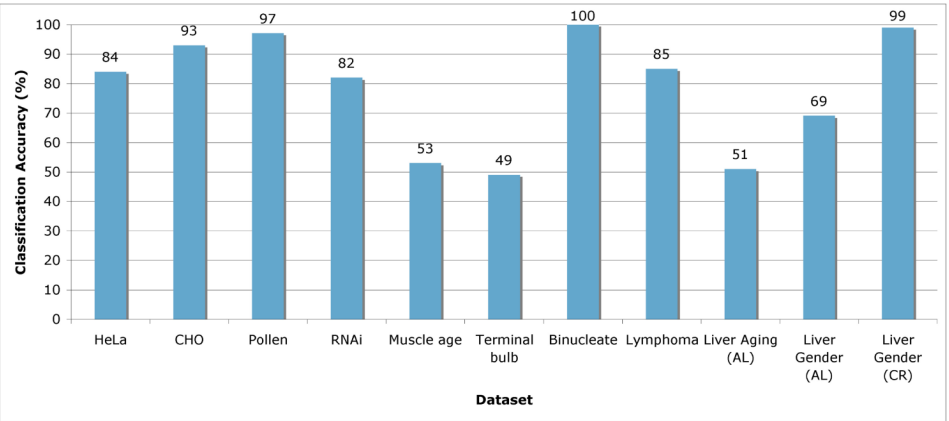

Fig. I.
Classification accuracy of the datasets using WND-CHARM image classification algorithm.

Table I

Number of classes, number of images, format, size and microscopy of the datasets of IICBU-2008.

Dataset	# of classes	# of images	Image format	Microscopy
Pollen	7	630	25×25 8 bit TIFF	Phase contrast
RNAi	10	200	1024×1024 16 bit TIFF	Fluorescence
Binucleate	2	40	1280×1024 16 bit TIFF	Fluorescence
Lymphoma	3	375	1388×1040 32 bit TIFF (color)	Brightfield
Liver gender (caloric restriction)	2	256	1388×1040 32 bit TIFF (color)	Brightfield
Liver gender (ad-libitum)	2	522	1388×1040 32 bit TIFF (color)	Brightfield
Liver aging	4	850	1388×1040 32 bit TIFF (color)	Brightfield
2D HeLa	10	860	382×382 16 bit TIFF	Fluorescence
CHO	5	340	512×382 16 bit TIFF	Fluorescence