COLLECTIONS OF CLASSIFIERS TUNED FOR CELL FINDING WITH AN APPLICATION TO BUILDING DIGITAL CELL ATLASES OF DROSOPHILA EMBRYOS

William Beaver[1], Adam Pare[2], David Kosman[2], Ethan Bier[2], William McGinnis[2], and Yoav Freund[1]

[1]Department of Computer Science and Engineering, UCSD, La Jolla, CA 92093
[2]Cell and Developmental Biology Department, UCSD, La Jolla, CA 92093

The different combinations of genes active in different cells control the development and diversity of multicellular organisms. Yet the codes that control this process, written in both cis–regulatory and protein–coding DNA sequence, are poorly understood. This is despite the availability of imaging techniques that allow detection of gene activity at the resolution of single transcription units on chromosomes in individual interphase diploid nuclei[1]. While it has been shown that digital cell atlases are indeed possible[2], producing query-able models from imaging for Systems Biology[3] is far from commonplace for even the simplest of organisms.

While much algorithmic work has been done on transforming pixels to meaningful objects (i.e. segmenting cell nuclei)[4], this necessary first step is far from a solved problem. Algorithm tuning remains a critical component and requires expertise in computer vision and image processing. This expertise is called upon repeatedly with new experimental data: when changes in the model organism, hybridization method, or the microscope result in changes to the generated images. We have successfully developed *adaptive* segmentation methods[5] that, while tuned for particular types of images, are tuned using machine learning methods. The advantage machine learning provides is rather than have the experimentalist tune the computer vision parameter(s) directly—this, again, requires an understanding of computer vision—the experimentalist instead tunes the system by providing direct feedback. The feedback mechanism is designed to be a of a natural form: given a collection of segmented pixels, the experimentalist labels each collection as "a partial nucleus", "a combination of two nuclei", "a complete nucleus", etc. From these these labeled collections we build classifiers that mimic the experimentalist's recognition of various segments.

Using the human labeled examples as input, classifiers are built using JBoost[6], an open source, Java implementation of the Adaboost machine learning algorithm[7]. Adaboost generates a mapping from segments to scores, where high scores correspond to high confidence that the segment is correct (Figure 1a–g). Adaboost is a general-purpose learning algorithm which is particularly powerful when combining many weakly predictive features. The predictive features in this case are the values of the morphological features of the segments (Figure 1f). JBoost generates classifiers in a form of decision tree called Alternating Decision Trees (ADT)[8]. The ADT classifier (Figure 1g) acts a as proxy for the experimentalist, and we use these classifiers to guide the search for proper segmentations. As the experimental conditions change for certain collections of pixels, or as the objects we are looking for change, we generate new classifiers using newly labeled data and can often use existing classifiers to bootstrap the labeling process. We have developed tools to decrease the time needed to generate a new classifier by an order of magnitude (from weeks or days, to days or hours).

In this work we describe uses for these collections of classifiers: initially as part of toolset aimed at building digital cell atlases that encode the spatial and temporal activity of a variety of developmental regulatory genes in embryos of *Drosophila melangaster* (Figure 1h–l), and later as part of a classifier distribution system: a repository of classifiers available for use by the wider biological community. This repository is envisioned to allow sharing of classifiers designed for specific experimental setups among labs who contribute classifiers and data back to the larger system. We believe that the successful completion of this work will mark an important step in the use of adaptive systems in imaging and in the future use of imaging within Systems Biology.
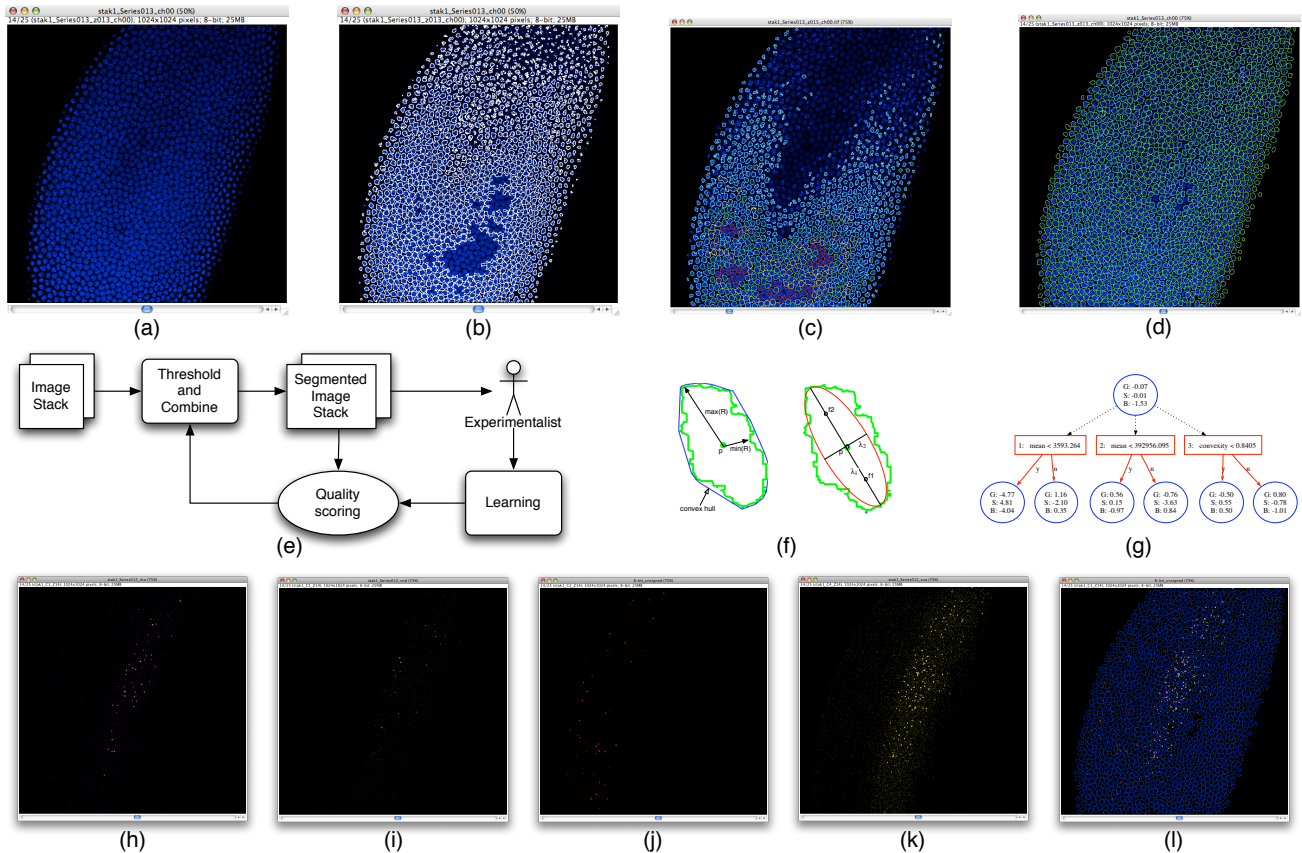
Fig. 1: (a) Example of slice from confocal microscopy labeled with DAPI and (b) the results when a general, global threshold used to find nuclei. (c) A variety of segmentations are labeled by experimenters based on "correctness". (d) Results of local thresholding algorithm that uses (e) classifier approach based on (f) morphological features extracted from each example nuclues. (g) An alternating decision tree showing the first three decision rules of our classifier for DAPI. (h–k) Application of this technique to mRNA channels which leads to (l) spatial model of transcription within nuclei. Higher resolution images are available at `http://eye.ucsd.edu`

1. REFERENCES

[1] Dave Kosman, Claudia M. Mizutani, Derek Lemons, W. Gregory Cox, William McGinnis, and Ethan Bier. Multiplex detection of RNA expression in Drosophila embryos. *Science*, 305(5685):846–, 2004.

[2] Fuhui Long, Hanchuan Peng, and Eugene Myers. Automatic segmentation of nuclei in 3d microscopy images of C. elegans. In *4th IEEE International Symposium on Biomedical Imaging: From Nano to Macro*. ISBI, 2007.

[3] Sean G Megason and Scott E Fraser. Imaging in Systems Biology. *Cell*, 130(5):784–795, 2007.

[4] Gang Lin, Monica K. Chawla, Kathy Olson, John F. Guzowski, Carol A. Barnes, and Badrinath Roysam. Hierarchical, model-based merging of multiple fragments for improved three-dimensional segmentation of nuclei. *Cytometry Part A*, 63A(1):20–33, 2005.

[5] W. Beaver, D. Kosman, G. Tedeschi, E. Bier, W. Mcginnis, and Y. Freund. Segmentation of nuclei in confocal image stacks using performance based thresholding. In *4th IEEE International Symposium on Biomedical Imaging*, pages 53–56. ISBI, 2007.

[6] JBoost. `http://www.cs.ucsd.edu/~aarvey/jboost/`.

[7] Yoav Freund and Robert E. Schapire. A Short introduction to Boosting. *Journal of Japanese Society for Artificial Intelligence*, 14(5):771–780, September 1999. Appearing in Japanese, translation by Naoki Abe, Available from `http://www.cse.ucsd.edu/~yfreund/papers/`.

[8] Yoav Freund and Llew Mason. The Alternating Decision Tree Learning Algorithm. In *Machine Learning: Proceedings of the Sixteenth International Conference*, pages 124–133, 1999.

[9] Robert E. Schapire, Yoav Freund, Peter Bartlett, and Wee Sun Lee. Boosting the margin: A new explanation for the effectiveness of voting methods. *The Annals of Statistics*, 26(5):1651–1686, October 1998.