Zach Pincus
CS223b Progress Report

0. Abstract

Modern techniques in microscopy, combined with new high-throughput screening methods from genetics and molecular biology, enable the generation of a tremendous number of individual data sets with very little in common between data sets. This poses a challenge for automated classification of images in such datasets because despite amount of data, there are relatively few instances of any given class in any particular set. The use of transductive learning techniques, which are designed to make efficient use of information in a limited data set, may allow automated or semi-automated classification to be successful in such situations.

1. Introduction and Background

Historically, the practice of microscopy can be broken into two major domains based on the amount of data generated and the method of analysis. The first could be called "low throughput" methods, characterized by many small research experiments analyzed manually by one to three investigators. Though computers may be used in the image generation and analysis, these techniques are in general not highly automated and are user-driven. The other major domain of microscopy in biomedicine is in industrial "high throughput" microscopy, where automated methods are used to score the results of thousands of identical experiments a day. Classical high throughput applications include screening cell cultures for response to potential pharmaceuticals, and evaluating the results of medical tests such as blood cell counts or pap smears.

Decreases in the cost of microscope automation have now brought the ability to generate large numbers of micrographs (in the hundreds and thousands) easily and rapidly from the exclusive purview of industrial scientists to the average academic research lab. Though these advanced tools for the automation data capture are now available to research scientists, software tools for the automation of data analysis are not. This is in large part because the task is much different: this "medium throughput" work is often characterized by a great variety in the experimental assays performed and in the cellular substrates imaged, which stands in stark contrast to the industrial case of one assay on one cell type, repeated continuously. Current software tools reflect this bias: they focus on defining simple, rapid "scripts" which can mechanically perform the desired analyses with a relatively simple toolbox of segmentation and statistical methods. (See Clemex Vision software [ref. CLEMEX] for an example of industry-standard analysis tools.

Though most defined analysis tasks can be automated with such scripts, this is often not worthwhile for individual researchers who will not be repeating a particular analysis task from day to day, but instead moving on to new experiments in the course of their research. Moreover, much research is inherently exploratory, and exploration is not a task easily scripted. Thus, there is need for intelligent systems for research biology that can, with minimal supervision, perform a wide variety of analyses or search for surprising outliers.

Many academic image analysis questions fall into the realm of pattern classification: finding relevant subpopulations of cells and analyzing their statistical makeup as compared to control populations. Classification of every cell in every image as "normal" or "abnormal" or binning them into different phases of the cell life cycle is a common initial step in many assays. Therefore, tools which can learn such classifiers would be extremely useful to research biologists.

Though pattern classification, especially with regard to image analysis, is heavily studied in computer science and engineering, the application of those techniques to biological imaging is relatively unstudied. In fact, only one researcher has published actively in this area [BOLAND2001, BOLAND1999]. The methods outlined in this work rely primarily on neural networks trained on features hand-tuned and hand-selected for particular tasks.

My approach to this task is to pursue general-purpose classification in a manner that reflects the realities of biological data generation. In the high-throughput case, a well-trained classifier with low generalization error is important, because the classifier would have to classify new data every day. However, in the academically relevant medium-throughput case, there is rarely new data to classify with a given classifier beyond the "test set." Thus, any clues to the nature of the data embedded in the structure of the given data set, or any features defined specifically over that data set, can be used to increase classification accuracy in this setting.

This learning paradigm, in which the structure of the data to be classified is used to train a classifier, is known as *transductive* learning. Much recent scholarship in machine learning has discussed different methodologies for transduction, but there have been few, if any, applications of this technique in bioinformatics in general, and biological image analysis in specific.

Unfortunately, I as of yet have no results relevant to my own application of these techniques to biological image analysis. I have many results regarding the appropriate way to compile and link OpenCV on Mac OS X, and about patches to the National Library of Medicine's Image Segmentation Toolkit that I have submitted, and about learning to use heavily templated C++ code. More to the point, I have preliminary results on appropriate techniques for separating foreground from background in micrographs.

This is a substantially more challenging problem than I initially anticipated due to the nature of the light microscopy imaging modality. In general, several images are taken of the same sample under different forms of illumination. Typically, one of these images serves as a "key" image that defines the space occupied by each cell. Typically, the key image is taken with *phase contrast* microscopy. A phase image provides an indication of the relative optical density (refractive index) of each object in the sample. Bacteria (Figure 1a) are phase-dense, and are quite easy to extract from images. However, mammalian cells (Figure 1b) are nearly the same phase density as water, and are largely discernable by the *phase ring* artifacts around the edges of the cells. In particular, the texture within cells can be extremely similar to that without, so texture-based differentiation of cell from extracellular environment is extremely challenging.

In other experiments, the key image can be a fluorescent protein or other soluble tag that serves as a "volume marker" which illuminates the entire cell volume when interrogated with light of a certain wavelength (Figure 1c). A key problem in segmenting volume-marked cells, or closely packed cells imaged with phase contrast microscopy is that simple thresholding is not sufficient to differentiate individual cells: some sort of shape-based morphology operation is necessary.

A common technique for this problem is to assume that cells are largely convex objects, and that a pair of matching concavities on opposite sides of an object signifies that that object should be split in two [BELIEN2002]. In general, there are few published techniques for extracting cells from background. In practice, simple threshold-based segmentation with hand-tuning is used.

2. Methods

Subproblem: Segmentation

One commonality among all of the image modalities that an automated algorithm might be required to segment is the presence of strong edges. In the case of phase contrast imagery, the edges are contrast-enhanced by the bright phase ring artifacts. In the case of fluorescent imagery, there is in general a discernable boundary between the black background and fluorescent cell volume. Parts of the cell volume may not fluoresce due to physical constraints (e.g. a cell nucleus excluding fluorescent dye, leading to cells with dark spots in the centers). Therefore, edges are the only certainty in this modality as well.

As such, an edge-finding filter seemed like a logical first step to segmenting these challenging images. I chose to use a modified Canny filter to detect these edges. My modifications were first to convert the algorithm to take percentile inputs instead of raw "edge-value" thresholds for the hysteresis thresholding. This allows images from very different modalities to be evaluated with the same parameters, and it makes much more intuitive sense: "take the top 10% of the edges and follow them" is more reasonable than "take the edges above value 5437…" My second modification is essentially to use a zero lower threshold, so that edges are followed to their farthest logical extent. Given that cells are by definition closed objects, where a single exterior edge defines a particular cell, and given that the sharpest edges in both phase and fluorescent imagery usually correspond to cell boundaries, it makes sense to try to follow a bright edge as far as it will go. Coupled with a stringent upper threshold, this succeeds in finding the edges of cells in all of the image types I have tried.

Once edges have been traced, they must be extracted and converted into some object representation. Given that cells have internal structure, there may also be edges defining intracellular boundaries. Therefore, my approach is to extract each connected edge and treat it as a closed polygon, provided it meets some "roundness" constraint (informed by the fact that most cells are mostly round; this approach will of course fail on cells like neurons. Spatially overlapping polygons will then be assembled into a single outline with Boolean operations.

If further segmentation is necessary, advanced techniques like level-set segmentation methods available in the ITK toolkit will be used to refine the polygons. These methods are not appropriate for initial segmentations, because they require fairly detailed initial guesses as to the location of edges.

Once objects are segmented, they will be extracted and aligned. The alignment will be to ensure that cells with distinct major and minor axes are appropriately registered. If the polygons are non-round, alignment can proceed trivially based on the two-dimensional second moments of the binary shape. This implicitly fits a Gaussian to the shape, and this parametric representation can be used to rotate each shape into a given alignment. If the shapes are somewhat round, they may still have major and minor axes of intensity within the binary outline. To determine this, eigenvalues and vectors of the intensity image in the region will be calculated, a la Harris corner detection. If there is a substantial disparity in eigenvalues, the eigenvectors will be treated as the natural basis for the cell coordinates, and rotation will be applied to bring the major axis of brightness into alignment with other such axes.

Subproblem: Transductive Classification
I plan to use off-the-shelf support vector machine classifiers to provide a baseline non-transductive classification. To implement transductive classification, I will follow the work of [JOACHIMS1999], which essentially defines a wrapper around simple SVM training. This wrapper attempts to classify unlabeled data and then train on those classifications in an expectation-maximization framework. Additionally, I will use simple SVMs in a transductive approach by defining features that implicitly encode the structure of the data set.

One such a feature set can be generated by first decomposing the data set into a preliminary set of relevant features. Then, in this feature space, calculate the pairwise distance between each point. Now, represent each data instance as the vector of distances to each other instance. In this manner, the structure of the data set is encoded in the feature representation, and simple SVM training is in some sense transductive.

If the pairwise distance matrix is too large, feature reduction a la J Shi's talk on surprising event detection can be applied. In this framework, "important" points in the initial feature space are defined as the K-means cluster centers of the data set in that feature space. Then, one can represent each data instance as a vector of distances to only the "important" points in that space.

Subproblem: Feature Decomposition
For transductive training of any sort to be applicable, an initial decomposition of image pixels into reasonable features is necessary. First, I will attempt to use "eigenfeature" methods (that is, principle component analysis of the dataset) to generate relevant features. I will also attempt to implement texture detection of some sort to provide a different feature decomposition to test against.

3. Results

I have very limited current results, given the tremendous amount of time and effort I have put into this work so far. Unfortunately, most of this time was spent in utterly failed false starts, or in attempting to get various image processing toolkits functioning properly and to learn their usage. (I ruled out using Matlab for a variety of reasons. I hope for this work to become a component of a larger system that I am building for my dissertation, and for better or worse, Matlab is not appropriate for such a component.)

I do have my modified Canny segmentation worked out. Using a 3x3 Sobel operator for differentiating the data, a 99^{th} percentile upper threshold, and a 0^{th} percentile lower threshold, I achieve excellent results on images from all major modalities that I will have to deal with. See Figures 3-5 for typical results.

4. Next Steps and Conclusion

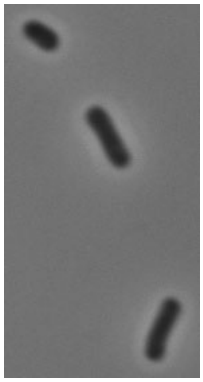
My planned next steps are to follow my proposed methodology laid out above.

In summary, I have utterly failed to achieve anything of significance so far. That said, I believe that I have finished the hardest part of the task: gathering an appropriate set of tools and learning their use. Given that I have more substantial expertise in machine learning algorithms and the tools available, I certainly believe that that portion of my project will proceed much more smoothly than has the initial data extraction portion.

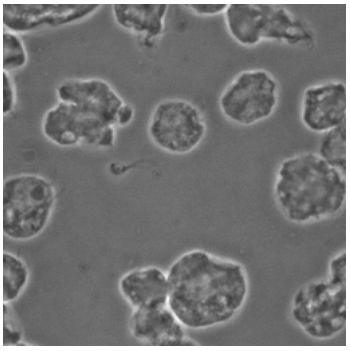
5. References

[CLEMEX] http://www.clemex.com/Products/ImageAnalysis/Software/VisionPE.aspx
[BOLAND2001] Boland MV, Murphy RF. Bioinformatics. 2001 Dec;17(12):1213-23. A neural network classifier capable of recognizing the patterns of all major subcellular structures in fluorescence microscope images of HeLa cells.
[BOLAND1999] Boland MV, Murphy RF. Trends Cell Biol. 1999 May;9(5):201-2. Automated analysis of patterns in fluorescence-microscope images.
[BELIEN2002] Belien JA, van Ginkel HA, Tekola P, Ploeger LS, Poulin NM, Baak JP, van Diest PJ. Cytometry. 2002 Sep 1;49(1):12-21. Confocal DNA cytometry: a contour-based segmentation algorithm for automated three-dimensional image segmentation.
[JOACHIMS1999] Joachims T. Proceedings of ICML-99, 16th International Conference on Machine Learning. Transductive Inference for Text Classification using Support Vector Machines.

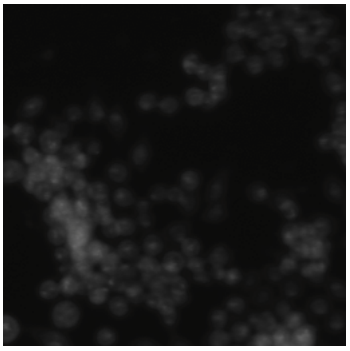
Figure 1



a. Bacteria in phase-contrast.



b. Mammalian cells in phase-contrast.



c. Mammalian cells with fluorescent volume marker.

Figure 2: Edge images (30% lower threshold, 99% upper, 7x7 Sobel operator.)