Toward Automated Cancer Diagnosis: An Interactive System for Cell Feature Extraction*

Nick Street
Computer Sciences Department
University of Wisconsin-Madison
street@cs.wisc.edu

Abstract

Oncologists at the University of Wisconsin-Madison have identified nine cell features which allow them to diagnose breast tumors from a fine-needle aspirate. Currently, the physician examines the sample under a microscope and assigns a number in the range 1-10 to each feature, where larger numbers indicate a higher chance of malignancy. These values, viewed as 9-dimensional vectors, are then used as the basis for diagnosis.

This paper presents an interactive system which uses computer vision techniques to extract a similar set of five features from a digitized image of the sample. The cell nuclei are isolated by use of deformable models known as snakes. Feature values are then found by computing some estimators for the desired quantities and are then scaled into the 1-10 range. Promising preliminary results and possible future directions for this research are presented.

1 Introduction

The goal of automating the diagnosis of breast tumors has led Dr. William H. Wolberg [3, 4, 6], oncologist at University of Wisconsin Hospitals, to isolate nine relevant cell features, such as uniformity of cell nucleus size and cell clump thickness. These features are isolated in tumor cells taken from a fine needle aspirate. Once the physician has assigned values to each of these features for a significant number of samples (both benign and malignant) the resulting nine-dimensional points can be used to train an automatic separation system. The pattern separator currently in use is described briefly below. Later samples can be diagnosed by the separator with a high rate of accuracy.

Current research is dedicated to further automating the diagnosis task by evaluating features from a single digitized image of the aspirate. Since the sample being examined is now reduced to a two dimensional image and some

*This research is directed by O. L. Mangasarian in collaboration with Dr. W. H. Wolberg and supported by AFOSR Grant AFOSR 89-0410 and NSF Grant CCR-8723091

detail is lost, certain parameters such as clump thickness can no longer be determined. We will show that with automated computation of feature values taken from this digital image we are able to correctly diagnose a high percentage of images. This is accomplished with minimal user intervention and very few training samples.

2 Current Work and Results

2.1 Image Preparation

After a cytology slide has been prepared, an area with a large number of representative cells is brought into focus under a microscope and photographed. A print made from this film is then scanned, using a camera attached to a PC with the appropriate digitizing hardware. The input to our analysis is then a 242×256-pixel gray-scale (8 bits / pixel) image. (Dr. Wolberg has recently obtained the necessary scanning equipment to eliminate the photograph step.)

2.2 The User Interface

Since locating a sufficient number of cells automatically proved to be infeasible due to variations in the different images, a graphical user interface was developed which allows the user to input the approximate location and size of enough cells to provide a significant statistical sample. The interface was developed using the X Window System and the Athena Widget Set on a DECstation 3100. In its current form, the interface requires the user to input the name of the image to be processed and the expected size of the cell nuclei in the image. (Measurements are currently in pixels but could easily be changed to units of the user's choice.) The user then uses a mouse button to pick a point near the center of each cell nucleus. If desired, the user may then pick another point near the edge of the nucleus. The two points define a radius of the circle which is drawn as the beginning approximation. Otherwise, a default size is used for the radius and the corresponding circle is drawn. See Figure 1.

2.3 Snakes

Once the position and approximate size have been determined, the actual boundary of the cell nucleus is located by use of an active contour model known in the literature as a "snake" [2]. A snake is a deformable spline which minimizes an energy function at a certain number of points. Using the circle as the initial set of points, the energy function is defined in such a way that the minimum value should occur when the points are evenly spaced around the boundary of a nucleus.

To achieve this, the energy function to be minimized is defined as follows:

$$E = \int (\alpha(s)E_{cont} + \beta(s)E_{curv} + \gamma(s)E_{image})ds$$

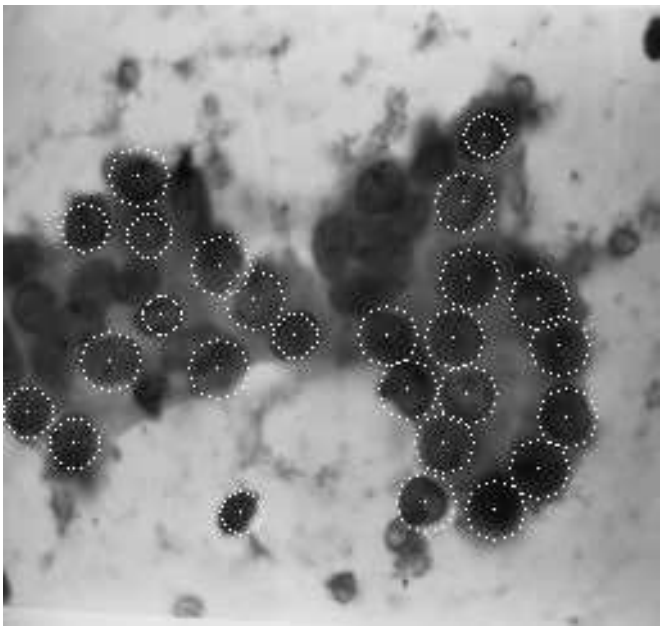

Figure 1: Cell Image with Initial Placement of Snakes

where E represents the total energy integrated across the arclength s of the snake. At a particular snake point, the energy computation is a weighted sum of energy terms E_{cont}, E_{curv} and E_{image} with respective weights $\alpha(s)$, $\beta(s)$ and $\gamma(s)$. The energy terms measure the following quantities:

- Continuity E_{cont}: This term measures how evenly spaced the snake points are. Note that this is a desirable property of the snake itself, and does not depend on the nucleus boundary it is attempting to isolate. To compute this, the distance from a snake point to one of its neighbors is found and compared to the average distance between adjacent points. The magnitude of this difference is then E_{cont}.

- Curvature E_{curv}: This term measures discontinuities in the curvature of the snake. Cell nuclei are more or less ellipsoidal, so points with abnormally high or low curvature, as compared to a circle, are penalized. Taking advantage of this knowledge about the nuclear shape, the following method was adopted. First, the 'center' of the snake (center of mass of the snake points) is located. The distance from a snake point to the center (i.e., length of a radius if the nucleus is circular) is then compared to the average of such distances in a neighborhood of the point. The magnitude of the difference is the energy term.

- Image E_{image}: This term measures the gray-level contrast between pixels inside the snake and those outside the snake. Here, we are using the

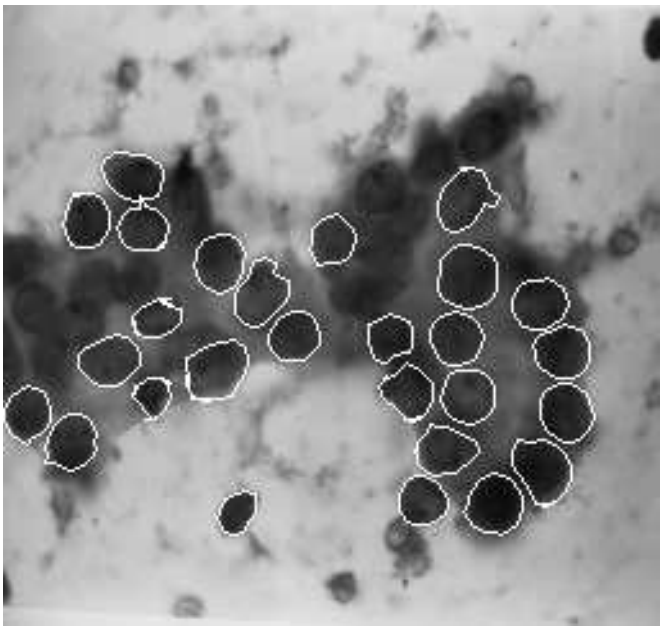

Figure 2: Snakes After Convergence to Cell Nucleus Boundaries

domain-specific knowledge that cell nuclei are generally darker than the surrounding material. A set of directional edge detectors is used to compute this term.

In the current system, the weights α, β and γ are empirically derived and do not actually depend on s. For best performance on a typical image, γ is set somewhat higher than the others to ensure that the snake converges to the visible boundary.

In order to control computation time, the optimal local value of the energy function is approximated using a greedy method algorithm due to Williams and Shah [5]. If the function value at a particular snake point can be lowered by moving the point to an adjacent pixel, then it is moved, thus possibly affecting the energy computation at other points. The process is repeated for each point until all points settle into a local minimum of the energy function.

If a snake fails to converge to the intended cell nucleus boundary after a preset number of iterations, it is deleted and a new one tried in its place. However, most of the snakes perform reasonably well, even in this relatively information-poor environment. For example, even in regions where gray scale information is lacking (say, where cells overlap, or where the background matter is as dark as the nucleus) the curvature property gives the snake a shape which is approximately the same as a human would draw, given the same task. See Figure 2.

Feature	Estimator
size	mean nuclear radius
shape	variance of nuclear radii
texture	variance of intensity levels in interior of nucleus
uniformity of size	variance of size across sample
uniformity of shape	variance of shape across sample

Table 1: Cell Features and their Estimators

2.4 Computation of Feature Estimators

Once all or most of the cell nuclei in an image have been isolated, statistics about the sample are estimated. The features we use are based on some of the nine features originally proposed and manually measured by Dr. Wolberg. For example, texture is a quantity not specifically measured in [6], but rather it attempts to combine two more specific features. The desired features and the estimators used in the current implementation are shown in Table 1.

The numbers are scaled into the range 1-10 so that they can easily be compared both to one another and to the numbers provided by Dr. Wolberg. This scaling can be done in several different ways and will be discussed later.

2.5 Results

In order to discuss results we must first describe the system used to separate the points into benign and malignant groups. Previous research [3, 4, 6] has developed a piecewise-linear separator based on linear programming techniques. This separator uses the multisurface method (MSM) which generates a series of hyperplanes in the appropriate dimensional space to isolate groups of only benign or only malignant data points. Future diagnosis can then be performed by the location of the point in question relative to the hyperplanes. Alternately, the results of MSM training can be used to generate weights for a neural network, which is then used for diagnosis [1]. Using the nine features that are evaluated by Dr. Wolberg and the other oncologists, the MSM system has successfully diagnosed about 98% of the 256 cases given to it over the past three years.

The numbers obtained from the current vision-based research were first tested with MSM using only three parameters; size, shape, and texture. Thirty-six photographic samples provided by Dr. Wolberg — 21 benign and 15 malignant — were analyzed, and the resulting 3-dimensional points are shown in Figure 3. As a heuristic aid to training, the two benign points which "dominate" some of the malignant points (i.e., they have higher values for each parameter) were removed. The remaining points were tested using an "all but one" methodology: the separator was trained with 33 points and tested on the remaining point. Out of the 34 cases thus tested, the remaining point was correctly diagnosed in 30 cases, giving an 88% success ratio. Using a randomly selected sample of two-thirds of the points for training, the success rate fell to about 81%, with a few of the borderline cases being repeatedly misdiagnosed. When

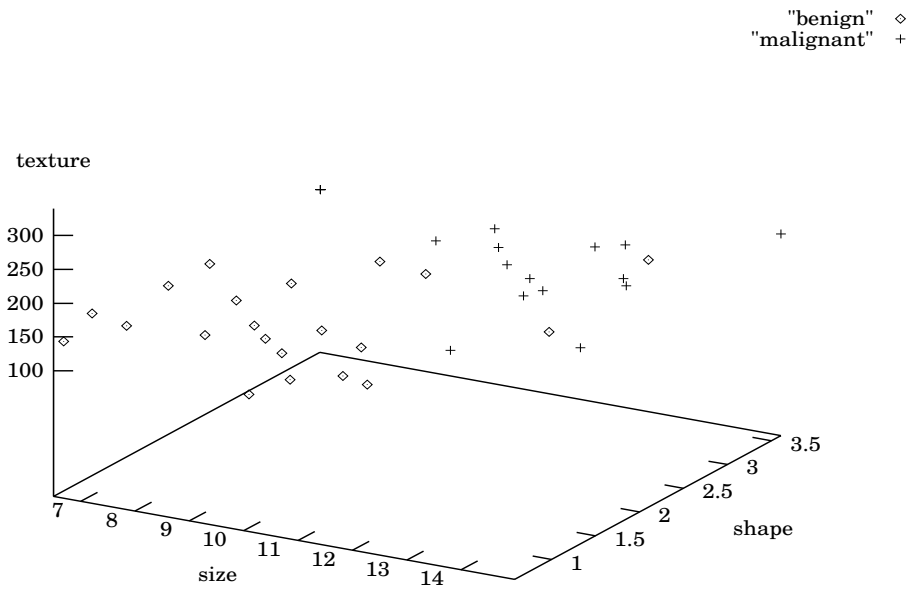

Figure 3: Resulting Points in 3-Dimensional Space

two additional features, uniformity of size and uniformity of shape, were also extracted from the images, the success rate using all-but-one testing fell to 79%. The causes of this drop are not clear and require further study. One possible explanation is that there were no dominating benign points that could be removed to enhance separation in the five-dimensional space. Nonetheless, the results to date are very encouraging, especially in light of the small number of samples that were available. Moreover, using the Dr. Wolberg's original 9-dimensional points, the MSM classifier correctly diagnosed only 81% of these cases using the all-but-one testing method.

3 Future Directions

As mentioned, Dr. Wolberg has obtained hardware to streamline our research. With a new microscope-mounted camera and the appropriate graphics hardware, he is now able to capture the sample in digital format and transfer the image electronically. We currently have nearly 600 samples of benign and malignant cases. The new system gives us somewhat better image resolution, which may increase the precision of the extracted feature values.

The energy function defining the snakes has many degrees of freedom; its performance should be optimized if we ever hope to use this as the basis of a "turnkey" diagnostic system for use by medical technicians. The three components of the energy function can be defined in different ways. For instance, the directional gray-scale derivative estimator could consider a larger or smaller

area, in effect smoothing the underlying image to a greater or lesser degree. Also, the weights associated with each of the three components could vary according to the specific sample rather than remaining constant.

As this work progresses, other significant cell features may be found to be automatically computable, especially with the increased resolution of the system being assembled. Also, different estimators for the current features may prove to be more accurate; for instance, the variance-based estimators appear to have a certain bias which will be addressed in later versions. Further, some subset of the current set of five features may provide the best separation.

Another question which has not been adequately considered is that of scaling the parameter values into the same range, in this case, integers from 1 to 10. Is direct linear scaling (in effect, using the raw numbers), which has been used in all testing to date, the best we can do? A clustering algorithm, which repeatedly combines the "closest" pair of adjacent clusters of sorted values, has been implemented and produced comparable results. Considering how a human expert might assign these values ("This looks like about a 6.") indicates that this could be an area where significant advances could be made.

Most importantly, the approach we have taken could be adapted and applied to many different cytology tasks. By working with doctors in different fields, we could expand the types of cells we are able to analyze as well as the types of features we are able to extract. Taken together with the MSM separator or some other automated pattern recognition system, the visual interface could be tailored to allow medical technicians in different fields to do cell analysis with a high degree of confidence in their results.

References

[1] K. P. Bennett & O. L. Mangasarian. Neural network training via linear programming. University of Wisconsin Computer Sciences Tech. Rpt. 948, July 1990, to appear in P. M. Pardalos (Editor), *Advances in Optimization and Parallel Computing*, North Holland, Amsterdam 1992.

[2] M. Kass, A. Witkin, and D. Terzopoulos. Snakes: Active contour models. In *Proc. of First International Conf. on Computer Vision*, 1987, pages 259-269.

[3] O. L. Mangasarian, R. Setiono & W. H. Wolberg. Pattern recognition via linear programming: Theory and application to medical diagnosis. In *Large-scale numerical optimization*, Thomas F. Coleman and Yuying Li, editors, SIAM, Philadelphia 1990, pages 22-30.

[4] O. L. Mangasarian and W. H. Wolberg. Cancer diagnosis via linear programming. In *SIAM News 23(5)*, September 1990, pages 1-18.

[5] D. J. Williams and M. Shah. A fast algorithm for active contours. In *Proc. 3rd Int. Conf. on Computer Vision*, Osaka, Japan, Dec 4-7, 1990, pages 592-595.

[6] W.H. Wolberg and O. L. Mangasarian. Multisurface method of pattern separation for medical diagnosis applied to breast cytology. In *Proceedings of the National Academy of Sciences U.S.A. 87*, 1990, pages 9193-9196.