Nuclear feature extraction for breast tumor diagnosis

W. Nick Street, William H. Wolberg and O. L. Mangasarian
Departments of Computer Sciences, Surgery, and Human Oncology
University of Wisconsin, Madison, WI 53706

ABSTRACT

Interactive image processing techniques, along with a linear-programming-based inductive classifier, have been used to create a highly accurate system for diagnosis of breast tumors. A small fraction of a fine needle aspirate slide is selected and digitized. With an interactive interface, the user initializes active contour models, known as snakes, near the boundaries of a set of cell nuclei. The customized snakes are deformed to the exact shape of the nuclei. This allows for precise, automated analysis of nuclear size, shape and texture. Ten such features are computed for each nucleus, and the mean value, largest (or "worst") value and standard error of each feature are found over the range of isolated cells.

After 569 images were analyzed in this fashion, different combinations of features were tested to find those which best separate benign from malignant samples. Ten-fold cross-validation accuracy of 97% was achieved using a single separating plane on three of the thirty features: mean texture, worst area and worst smoothness. This represents an improvement over the best diagnostic results in the medical literature. The system is currently in use at the University of Wisconsin Hospitals. The same feature set has also been utilized in the much more difficult task of predicting distant recurrence of malignancy in patients, resulting in an accuracy of 86%.

1. INTRODUCTION

The diagnosis of breast tumors has traditionally been performed by a full biopsy, an invasive surgical procedure. Fine needle aspirations (FNAs) provide a way to examine a small amount of tissue from the tumor; however, diagnosis with this procedure has met with mixed success.[4,5] By carefully examining both the characteristics of individual cells and important contextual features such as the size of cell clumps, physicians at some specialized institutions have been able to diagnose successfully using FNAs. However, many different features are thought to be correlated with malignancy, and the process remains highly subjective, depending upon the skill and experience of the physician. In order to increase the speed, correctness, and objectivity of the diagnosis process, we have used image processing and machine learning techniques.

2. CELL NUCLEUS LOCATION

2.1. Image preparation

The diagnosis procedure begins by obtaining a small drop of fluid from a breast tumor using a fine needle. The aspirated material is then expressed onto a glass slide and stained. The image for digital analysis is generated

0-8194-1138-8/93/$6.00

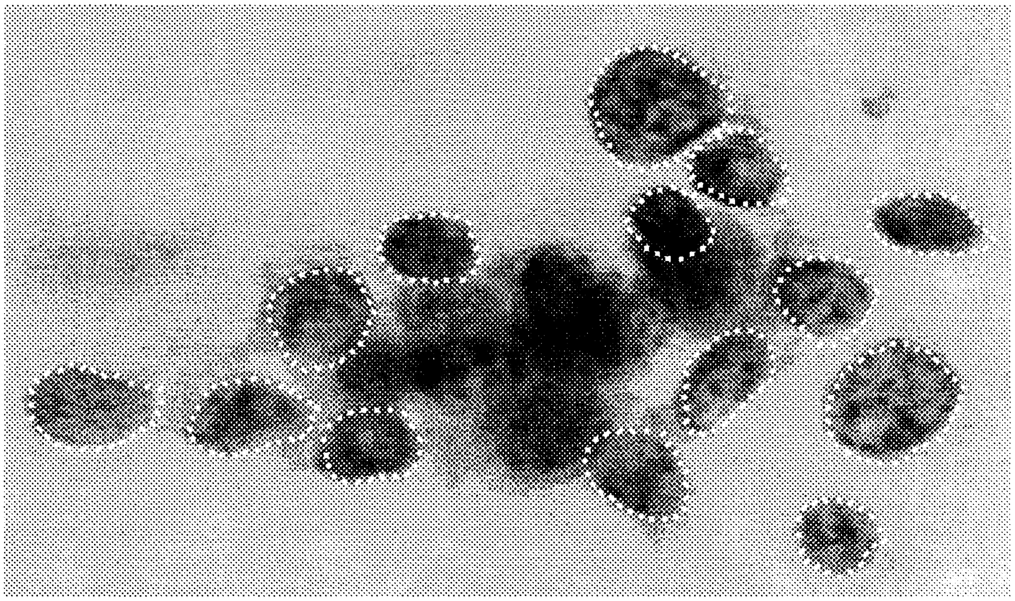

Figure 1: Initial Approximate Boundaries of Cell Nuclei
The user first draws a rough initial outline of some cell nucleus boundaries. Each outline serves as the initial position for a deformable spline which converges to an accurate boundary of the nucleus.

by a JVC TK-1070U color video camera mounted atop an Olympus microscope and the image is projected into the camera with a 63× objective and a 2.5× ocular. The image is captured by a ComputerEyes/RT color frame grabber board (Digital Vision, Inc., Dedham MA 02026) as a 512×480, 8-bit-per-pixel Targa file.

2.1.1. The user interface

The first step in successfully analyzing the digital image is to specify an accurate location of each cell nucleus boundary. A graphical user interface was developed that allows the user to input approximate initial boundaries of enough nuclei to provide a representative sample. The interface was developed using the X Window System and the Athena Widget Set on a DECstation 3100. A mouse button is used to trace a rough outline of some visible cell nuclei. These outlines are shown in Figure 1.

2.2. Snakes

Beginning with a user-defined approximate boundary as an initialization, the actual boundary of the cell nucleus is located by an active contour model known in the literature as a "snake".[7] A snake is a deformable spline which seeks to minimize an energy function defined over the arclength of a closed curve. The energy function is defined in such a way that the minimum value occurs when the curve accurately corresponds to the boundary of a cell nucleus. To achieve this, the energy function to be minimized is defined as the

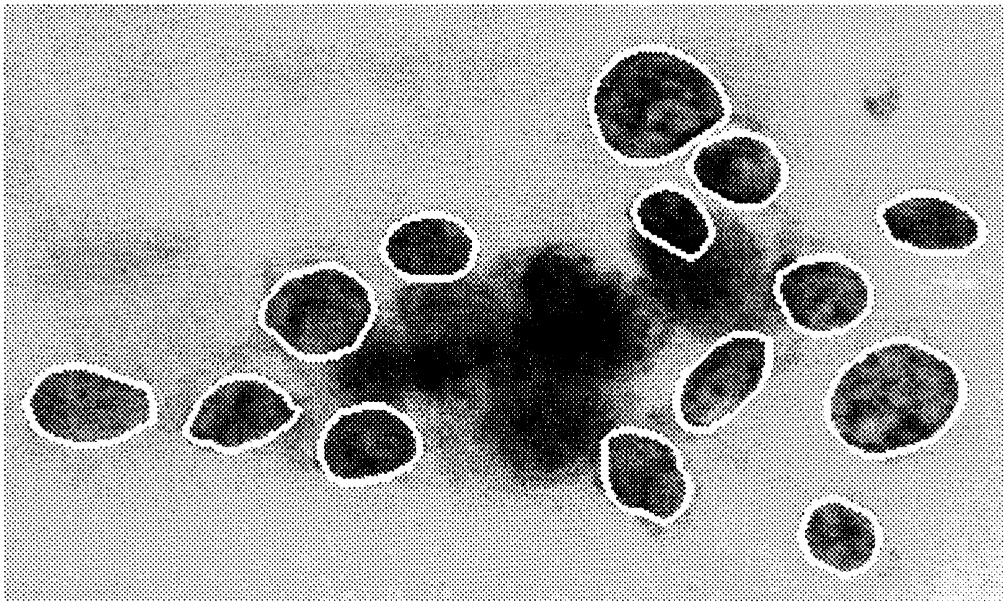

Figure 2: Snakes After Convergence to Cell Nucleus Boundaries
These contours are the final representation of the cell nuclei boundaries after the user is satisfied with the convergence of the snakes. This interactive process takes about two to five minutes.

following function of arclength s:

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

Here E represents the total energy integrated along the arclength s of the snake. The energy computation is a weighted sum of energy terms E_{cont}, E_{curv} and E_{image} with respective weights α, β and γ. To simplify the necessary processing, the energy function is computed at a number of discrete points along the curve, and the sum of these values is minimized. The component energy terms measure the following quantities:

- Continuity E_{cont}

  This term is constructed to penalize discontinuities in the curve. In the discrete case, this term measures how evenly spaced the snake points are. Note that this is a geometric property of the snake itself, and does not depend on the nucleus boundary that is being determined. 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 geometric term measures discontinuities in the curvature of the snake. Cell nuclei are more or less ellipsoidal; hence, points with abnormally high or low curvature, 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 radial line) is then compared to the average of such distances in a neighborhood of the point. The magnitude of the difference is this energy term E_{curv}.

- Image E_{image}

  This is the only term that ties the snake's performance to the underlying image. In our case E_{image} measures the gray-level discontinuity along the snake. To quantify this discontinuity we convolve the area of the image corresponding to the snake point with a Sobel[1] edge detector and observe the resulting edge magnitude. This term is customized by taking advantage of the fact that cell nuclei are generally darker than the surrounding material. Hence, the edge detection template is rotated so that the expected edge is perpendicular to the radial line of the nucleus at that point. For instance, for a snake point directly above the center of the nucleus, the edge template

1	2	1
0	0	0
-1	-2	-1

  would be applied. In this way, gray scale discontinuities which are perpendicular to the radial line produce the highest edge score. E_{image} is defined so a sharp discontinuity minimizes the energy value.

The weights α, β and γ are empirically derived constants. For best performance on these images, γ is set somewhat higher than the others to ensure that the snake converges to any visible boundary. The curvature term determines the snake's shape in cases of low contrast or partial occlusion. The continuity term does not determine shape, but does prevent snake points from bunching together near areas of sharpest gray scale contrast.

In order to control computation time, the optimal local value of the energy function is approximated using a greedy algorithm due to Williams and Shah.[13] 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. The results of a typical image are shown in Figure 2.

3. NUCLEAR FEATURES

The computer vision diagnostic system extracts ten different features from the snake-generated cell nuclei boundaries. All of the features are numerically modeled such that larger values will typically indicate a higher likelihood of malignancy. The extracted features are as follows.

1. *Radius*

   The radius of an individual nucleus is measured by averaging the length of the radial line segments defined by the centroid of the snake and the individual snake points.

2. *Perimeter*

   The total distance between the snake points constitutes the nuclear perimeter.

3. *Area*

   Nuclear area is measured simply by counting the number of pixels on the interior of the snake and adding one-half of the pixels in the perimeter.

4. *Compactness*

   Perimeter and area are combined[1] to give a measure of the compactness of the cell nuclei using the formula $perimeter^2/area$. This dimensionless number is minimized by a circular disk and increases with the irregularity of the boundary. However, this measure of shape also increases for elongated cell nuclei, which do not necessarily indicate an increased likelihood of malignancy. The feature is also biased upward for small cells because of the decreased accuracy imposed by digitization of the sample. We compensate for the fact that no single shape measurement seems to capture the idea of "irregular" by employing several different shape features.

5. *Smoothness*

   The smoothness of a nuclear contour is quantified by measuring the difference between the length of a radial line and the mean length of the lines surrounding it. This is similar to the curvature energy computation in the snakes. See Figure 3.

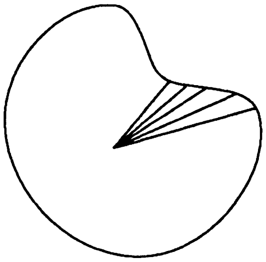

Figure 3: Radial Lines Used for Smoothness Computation

6. *Concavity*

   In a further attempt to capture shape information we measure the number and severity of concavities or indentations in a cell nucleus. We draw chords between non-adjacent snake points and measure the extent to which the actual boundary of the nucleus lies on the inside of each chord (see Figure 4). This parameter is greatly affected by the length of these chords, as smaller chords better capture small concavities. We have chosen to emphasize small indentations, as larger shape irregularities are captured by other features.

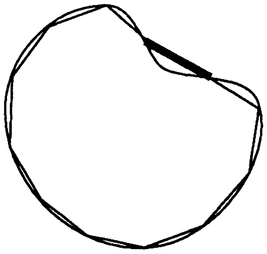

Figure 4: Chords Used to Compute Concavity

7. *Concave Points*

   This feature is similar to Concavity but measures only the number, rather than the magnitude, of contour concavities.

8. *Symmetry*

In order to measure symmetry, the major axis, or longest chord through the center, is found. We then measure the length difference between lines perpendicular to the major axis to the cell boundary in both directions. See Figure 5. Special care is taken to account for cases where the major axis cuts the cell boundary because of a concavity.

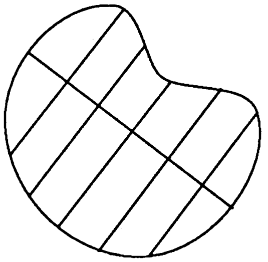

Figure 5: Segments Used in Symmetry Computation

9. *Fractal Dimension*

   The fractal dimension of a cell is approximated using the "coastline approximation" described by Mandelbrot.[9] The perimeter of the nucleus is measured using increasingly larger 'rulers'. As the ruler size increases, decreasing the precision of the measurement, the observed perimeter decreases. See Figure 6. Plotting these to values on a log scale and measuring the downward slope gives (the negative of) an approximation to the fractal dimension. As with all the shape features, a higher value corresponds to a less regular contour and thus to a higher probability of malignancy.

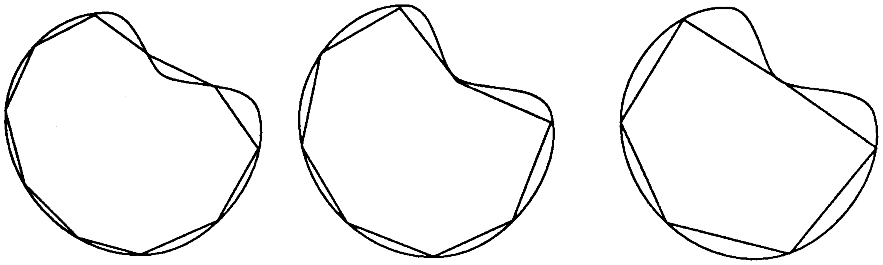

Figure 6: Sequence of Measurements for Computing Fractal Dimension

10. *Texture*

    The texture of the cell nucleus is measured by finding the variance of the gray scale intensities in the component pixels.

All of the shape features were verified using idealized phantom cells. They were shown to increase as the boundaries became less regular, and to be largely uncorrelated with the size of the contour.

The mean value, extreme (largest) value and standard error of each feature are computed for each image. The extreme values are the most intuitively useful for the problem at hand, since only a few malignant cells may occur in a given sample.

4. DIAGNOSTIC RESULTS

A set of 569 images has been processed in the manner described above, yielding a database of 30-dimensional points. The problem then becomes one of pattern separation, that is, determining how these points can best be separated into benign and malignant sets. The classification procedure used is a variant on the Multisurface Method (MSM)[10,11] known as MSM-Tree (MSM-T).[2,3] This method uses a linear programming model to iteratively place a series of separating planes in the feature space of the examples. If the two sets of points are linearly separable, the first plane will be placed between them. If the sets are not linearly separable, MSM-T will construct a plane which minimizes the average distance of misclassified points to the plane, thus nearly minimizing the number of misclassified points. The procedure is recursively repeated on the two newly created regions. Although the algorithm includes a pruning procedure to reduce the size of the resulting decision tree, our results were obtained by manually restricting the number of separating planes, and thus the number of decision regions.

In order to generate a classifier which generalizes well to unseen cases, we sought to minimize not only the number of separating planes but also the number of features used. The resulting single-plane classifier separates the points based on three feature values: mean texture and extreme values of area and smoothness. The plane, shown in Figure 7, separates 97.3% of the cases successfully. In order to estimate the performance on unseen cases, a ten-fold cross-validation[12] was performed. This train-and-test procedure divides the dataset into ten randomly selected, equally sized parts and uses each as a test set on a classifier created from the others. It thus provides a prediction of how well the classifier would perform on the universe of unseen cases. This estimate is unbiased and also very accurate in cases such as ours which have a fairly large number of training samples. In the case of this classifier, the predicted correctness was 97.0%.

To aid the physician and give the most complete picture of the classifier's effectiveness, we have implemented a method of varying the position of this single separating plane. As with many medical tests, the reliability of our diagnostic procedure is graded by the following numbers: $Sensitivity = \frac{correct\ positive}{total\ positive}$, $Specificity = \frac{correct\ negative}{total\ negative}$. By moving the plane parallel to itself we can vary the specificity and sensitivity of the test. In practice, the plane is moved to include the point being examined, and the specificity and sensitivity are computed for the resulting classifier. The process is shown in two dimensions in Figure 8.

The digital features have also been used to predict prognosis, that is, whether or not the cancer will recur at some future time in patients. with malignant tumors. Selecting an endpoint of two years, 124 samples were used. Table 1 shows the leave-one-out testing[8] results and the features used in the classifier with the best test set separation for various combinations of features and planes. Note that this more ambiguous data required more planes (but still only a few features) to get satisfactory separation. Also, the best feature sets seem to follow a distinct pattern: one size feature plus one shape feature, with additional features adding only marginal correctness.

5. CONCLUSIONS AND FUTURE WORK

We have described a system that uses image processing and machine learning techniques to diagnose breast tumors by non-invasive fine needle aspiration. The system utilizes an interactive interface that allows fast, accurate and objective diagnosis, even by untrained observers. The system is now in use at the University of Wisconsin Hospitals and is one tool used by doctors there for diagnosis of breast cancer, the second most deadly form of cancer in the U.S. since 1970.[6]

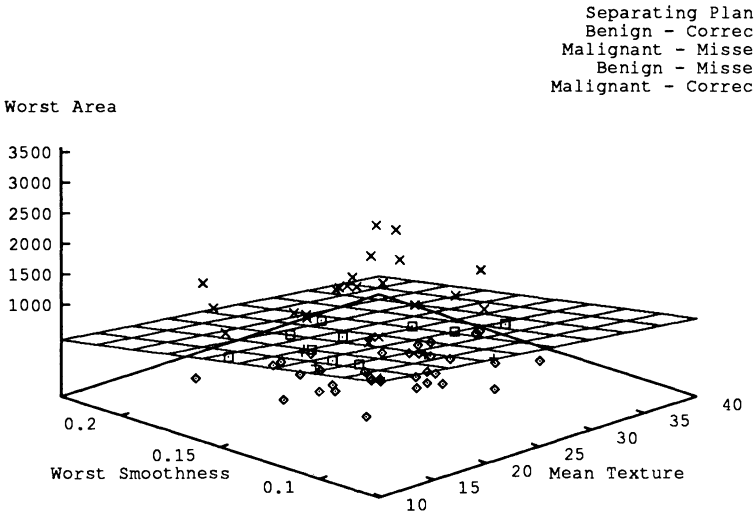

Figure 7: Separating Plane in Three Dimensions
In order to clarify the plot, only 10% of the correctly classified benign and malignant points are shown here. All of the misidentified points are shown.

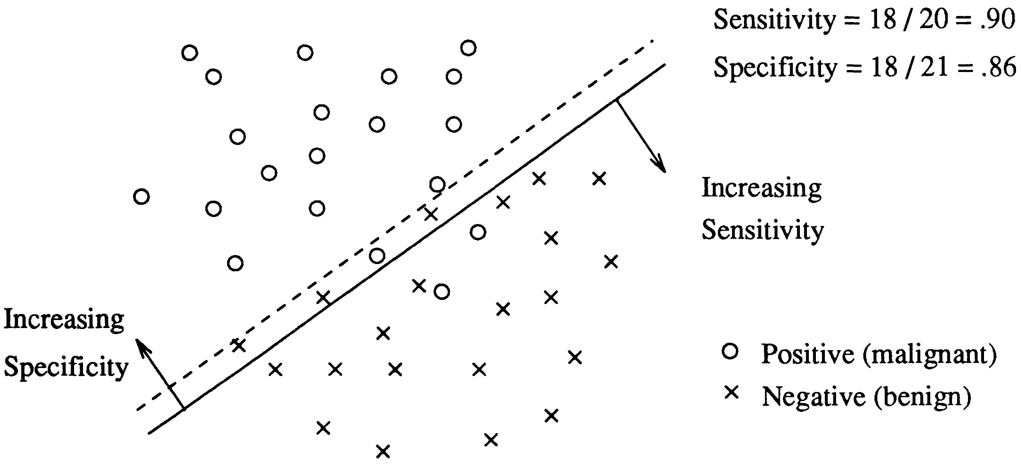

Figure 8: Adjusting Sensitivity and Specificity
A possible separating plane (here, simply a line) for a two-dimensional data set, and the resulting sensitivity and specificity of the classifier, is shown. By moving the line either direction parallel to itself (i.e., along its normal vector) these numbers can be adjusted. For instance, the dotted line represents a separator with 100% specificity at the cost of lower sensitivity.

		Number of Planes		
		1	2	3
Number of Features	1	SE Perimeter 71.8		
	2	SE Perimeter SE Smoothness 74.2	W Radius W Fractal Dim 79.8	
	3	SE Area SE Compactness SE Fractal Dim 75.0	M Area W Concave Pts W Fractal Dim 81.5	M Smoothness M Compactness M Fractal Dim 83.9
	4	M Radius M Area SE Concave Pts SE Fractal Dim 76.6	M Texture W Area W Concavity W Fractal Dim 86.3	M Texture M Compactness W Area W Fractal Dim 81.4

Table 1: Features and Testing Correctness for Prognosis Data
All subsets of k features were tested for training set separation. The subset which demonstrated the best separation was then tested using the leave-one-out approach, and the percent correctness is shown. (M = mean, SE = standard error, W = worst)

Future directions for this research will be driven both by the need to improve the existing diagnostic and prognostic systems and the possibility of generalizing this approach to other forms of cancer as well as other cellular diseases. There are three distinct paths along which this research will move: sample preparation, image processing and pattern separation.

Most of the issues involved in the preparation of the sample lie in the medical realm, with one exception. A certain selection bias is introduced in the process when the physician decides what part of the sample will be digitized. While the bias is very difficult to quantify, it is possible that if the physician suspects the sample to be malignant, then the selected cells will reflect that suspicion. This bias could be reduced by selecting a number of different areas for digitization, or possibly eliminated altogether by automating the selection process.

In the area of image processing, certainly our feature set is not comprehensive, and new features may be better suited to the analysis of other diseases. However, the richest area for future work would seem to be the snake model. While this kind of deformable contour model is a very powerful tool, it has the disadvantage of requiring a fairly precise initialization in order to converge to the desired contour. One interesting possibility would be to apply machine learning to the process in such a way that the model becomes tailored to the particular type of object being detected. In our case, recall that the weights assigned to the various components of the energy function were empirically derived. These could instead be learned, using both the snake's performance (speed to convergence and resulting energy) and a subjective user grade as feedback. Further, domain-specific heuristics, such as the directional edge detectors and expected elliptical shape, might also be learned.

Current work in pattern separation is concentrating on the problem of feature selection. The exhaustive search through the space of possible feature subsets is clearly unacceptable for larger subsets. Various heuristic search techniques, as well as machine learning approaches, are being considered for selecting both the feature subset and the number of separating planes which are necessary. Different pattern separation approaches are being considered that take greater advantage of the speed and flexibility of MSM-T.

6. ACKNOWLEDGEMENTS

This study was supported in part by Air Force Office of Scientific Research grant AFOSR 89-0410 and National Science Foundation grant CCR-9101801. The authors also wish to thank Kurt de Venecia of Helava and Associates, San Diego, CA for suggesting the use of the fractal dimension measurement.

REFERENCES

[1] D. Ballard and C. Brown. *Computer Vision.* Prentice–Hall, Inc, Englewood Cliffs, New Jersey, 1982.

[2] K.P. Bennett. Decision tree construction via linear programming. In *Proc. of the 4th Midwest Artificial Intelligence and Cognitive Science Society Conference*, pages 97–101, 1992.

[3] K.P. Bennett and O.L. Mangasarian. Robust linear programming discrimination of two linearly inseparable sets. *Optimization Methods and Software*, 1:23–34, 1992.

[4] W.J. Frable. *Thin-needle aspiration biopsy. Major problems in pathology 14.* WB Saunders Co., Philadelphia, 1983.

[5] R.W.M. Giard and J. Hermans. The value of aspiration cytologic examination of the breast. a statistical review of the medical literature. *Cancer*, 69(2104-2110), 1992.

[6] M.S. Hoffman. *The World Almanac and Book of Facts 1993.* World Almanac, New York, NY, 1992.

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

[8] P. Lachenbruch and P. Mickey. Estimation of error rates in discriminant analysis. *Technometrics*, 10:1–11, 1968.

[9] B.B. Mandelbrot. *The fractal geometry of nature.* W.H. Freeman and Company, New York, NY, 1977.

[10] O.L. Mangasarian. Multi-surface method of pattern separation. *IEEE Trans on Information Theory*, IT-14:801–807, 1968.

[11] O.L. Mangasarian. Mathematical programming in neural networks. Technical Report 1129, Computer Sciences Department, University of Wisconsin, Madison, Wisconsin 53706, December 1992.

[12] M. Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society*, 36:111–147, 1974.

[13] D. J. Williams and M. Shah. A fast algorithm for active contours. In *Proc. Third Int. Conf. on Computer Vision*, pages 592–595, Osaka, Japan, December 1990.