A Fast and Robust Approach for Automated Segmentation of Breast Cancer Nuclei

Kyoung-Mi Lee* and Nick Street†

Abstract

This paper presents an automatic segmentation method to locate cell nuclei in cytological images using two iterative generalized Hough Transforms (GHTs), one for obtaining knowledge on the size of the nuclei in the image, and the other for isolating nuclei themselves. In order to reduce the parameter space of the image, the original image is scaled down to half-sized and quarter-sized images. Using the information from the first iterative GHT on a quarter-sized image, the range of nuclear sizes is determined to limit the parameter space of the half-sized image. After the second iterative GHT on the half-sized image, nuclei are detected and segmented with edge information which helps determine the exact boundary. The results show that this method gives reduction in computation time and memory usage without loss of accuracy.

1 Introduction

Detecting objects in images is one of most interesting tasks of computer vision. An important example is cell detection in cytologic and histologic images for the detection of breast cancer. A typical algorithm for this detection is the Hough Transform (HT) [11], or its variant the generalized Hough Transform (GHT) [2], a robust, well known and very powerful template-matching algorithm to detect an arbitrary, but known, shape in a digital image. Primitives on the HT are represented by parametric curves with a number of free parameters. The principal concept of the HT is to define a mapping between an image space and a parameter space. Each edge point in an image is transformed by the mapping to determine cells in the parameter space whose associated parameters are such that the defined primitive passes through the data point. The chosen cells are accumulated and after all the points in an image have been considered, local maxima in the accumulator correspond to the parameters of the specified shape.

Because a curve with n parameters requires and n-dimensional parameter space [1], an ellipse with five parameters has a five-dimensional parameter space, two for the size of the image ($w \times h$), two for the size of templates (a and b) and one for the shape of templates (θ). Detecting ellipses

*Computer Science Department, University of Iowa, Iowa City, IA 52242, klee1@cs.uiowa.edu

†Management Sciences Department, University of Iowa, Iowa City, IA 52242, nick-street@uiowa.edu

in this way requires storage proportional to $wh|a||b||\theta|$, where $|i|$ represents the cardinality of the parameter. Hence, algorithms for ellipse detection are slow and memory intensive. To overcome excessive memory requirements, iterative GHT [6], a GHT extension, used a three-dimensional parameter space (a, b, and θ) for templates, and only two two-dimensional accumulators, one global and one local. GHT is iterated with elliptic templates defined by size parameters a and b and rotation parameter θ. At the end of each iteration, the local accumulator is compared to the global accumulator. If the value at location (x, y) of the local accumulator is higher than the value of the same location (x, y) of the global accumulator, that means the current template is a better match to a cell nucleus than any previous template. So the value at the local accumulator is stored in the global accumulator which always has the best match. After completion of the iterative process, each location of the global accumulator has the highest value of any local accumulator. Iterative GHT thus reduced the use of memory from $\Theta(wh|a||b||\theta|)$ to $\Theta(wh)$.

In previous work, Xcyt [5, 9], a graphical computer program for diagnosing breast cancer and predicting the course of the disease, used this iterative GHT algorithm. The system performs analysis of cytological features based on a digital scan of a breast fine-needle aspirate, diagnosis of the image as benign or malignant along with an estimated probability of malignancy, and prediction of when the cancer is likely to recur for cancerous samples. Iterative GHT is the first step toward automatically segmenting and analyzing nuclei in the Xcyt system. Even though the iterative GHT achieved memory efficiency, the Xcyt system suffered from excessive computational load. The presented algorithm solved this problem, reducing the range of four dimensional parameters (w, h, a, and b) by image scaling.

Automatic exact segmentation of medical images is very important, since applications need to extract precisely the interesting organic features in the human body. The Xcyt system uses an adaptive spline curve-fitting technique known as a snake [4] to determine the exact boundaries of the cell nuclei. A template detected by iterative GHT initializes the boundaries. To get more accurate boundaries, we added a local tuning algorithm.

A description of the proposed algorithm is given in Section 2. In Section 3, the performance of the system is examined. Conclusions are presented in Section 4.

2 Description of the Method

Our method operates in two parts. It first applies a set of iterative GHTs to an image for detection and then uses a local tuning algorithm for segmentation. The iterative GHTs examine each location in the image at two scales, looking for locations that might contain a nucleus. The local tuning algorithm and snake are then applied to segment the nucleus.

2.1 Detection of ellipses

In this part, we use two iterative GHTs, one for obtaining knowledge on templates in the image and the other for detecting nuclei. Most of the computational time of the iterative GHT is spent performing the GHT for each template that might correspond to a nucleus in the image. But, the range of sizes of nuclei in any one image is typically much smaller than the number of templates that we generate. So, the time spent searching for templates which do not exist in the image is wasted. To reduce the search spaces of the templates, we need to know which sizes of nuclei exist in the image. This is achieved by an iterative GHT on a reduced image. The role of the first iterative GHT is to find the range of a and b that includes all nuclei in the image. The shapes and sizes of the nuclei are found in the second iterative GHT. So the first iterative GHT doesn't need to use all the templates.

Image scaling is used to reduce the parameter space of the image, w and h, and the iterative GHT is applied at the reduced size. This iterative GHT must have some invariance to position and scale. The amount of invariance determines the number of scales and positions at which it must be applied [7]. For the work presented here, we apply the iterative GHT at every pixel position in the image and scale the image down twice by a factor of 2.

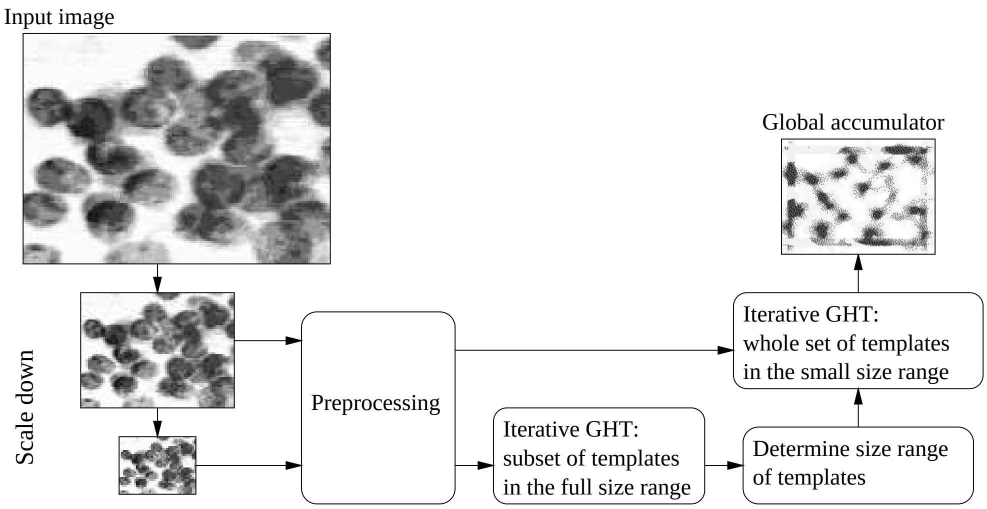
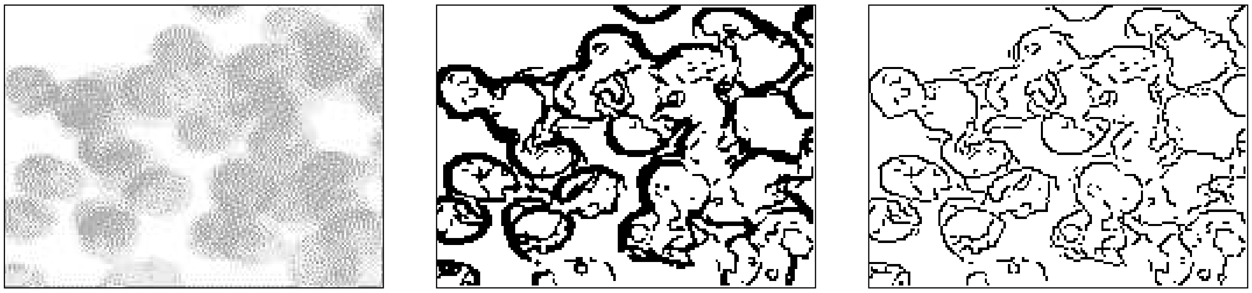


Figure 1: Algorithm for detecting ellipses

The algorithm for the first part is shown in Figure 1. A preprocessing step is first applied to the reduced images. The preprocessing first smoothes the intensity values across the image. We use a median filtering algorithm to diminish spurious effects that can be present in a digital image as a result of a poor sampling. The Sobel edge detector is then used to detect sharp gray-scale gradients and determine their direction. Then, edge thinning is performed, which finds the local maxima of the gradient magnitudes. This reduces the number of edge components so that further analysis is facilitated. The preprocessing steps are shown in Figure 2.



(a) Median Filtering (b) Sobel edge detecting (c) Edge thinning

Figure 2: The steps in preprocessing

The preprocessed images are then passed through iterative GHTs. After iterating the GHT with an elliptic template and the local accumulator, the local accumulator is compared with the global accumulator. Since we are interested in finding only the best match among the templates, only the highest value is important and thus is stored in the global accumulator. Finally, the global accumulator includes the highest values of each image location. In our system, a four-layer global accumulator was used to preserve the following information: the accumulator values, two parameters for the size of the templates, and the rotation direction of the template. In this first stage, only a few templates are applied for each possible size.

The best template matches from the first stage are used to determine the search range for the second stage. For instance, if the well-matched nuclei have radius of around 12 to 14 pixels, we would exhaustively search the size range 11 to 15 in the second stage. We performed this step an a half-sized image as a compromise between the precision of full image and the computation speed of smaller images.

2.2 Segmentation of nuclei

After the second iterative GHT, the peak finding step is performed to find the location of each nucleus and the shape of the template that closely matches the nucleus. The area of the global accumulator that was included in the template is then eliminated, since neighbors of the high peak point are also likely to have high values. The Xcyt system first finds the ten best matches on all templates. More matches can be found by using the global accumulator which is already calculated.

The algorithm for the second part is shown in Figure 3. Since the iterative GHT is applied to the half-sized image, we scale up of the position of the best matches. A snake uses scaled positions along a template as initial points to precisely locate the boundary of the nucleus. This resolves a shortcoming of the snake which requires an initial position to be specified near the desired boundary,

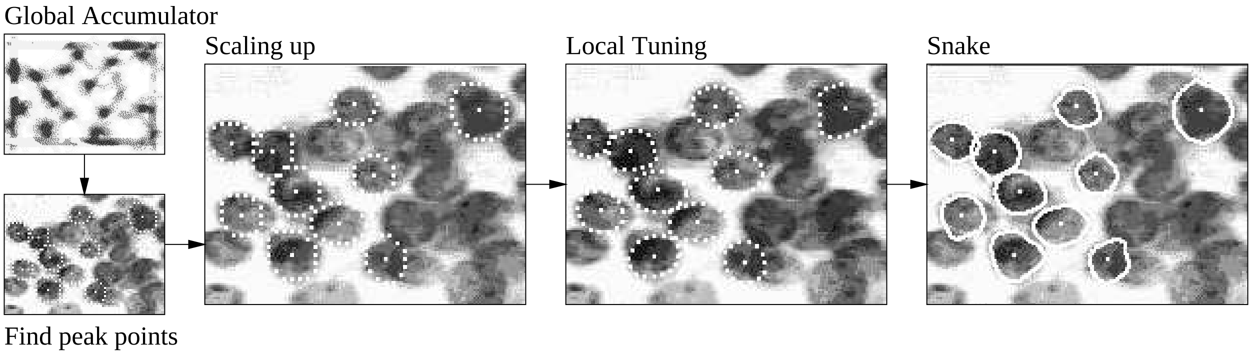


Figure 3: Algorithm for segmenting nuclei

and is sensitive to the specification of this contour [3]. Most nuclei are segmented well, while false boundaries by a wrong initialization may remain as in Figure 4.

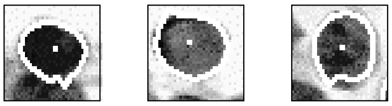


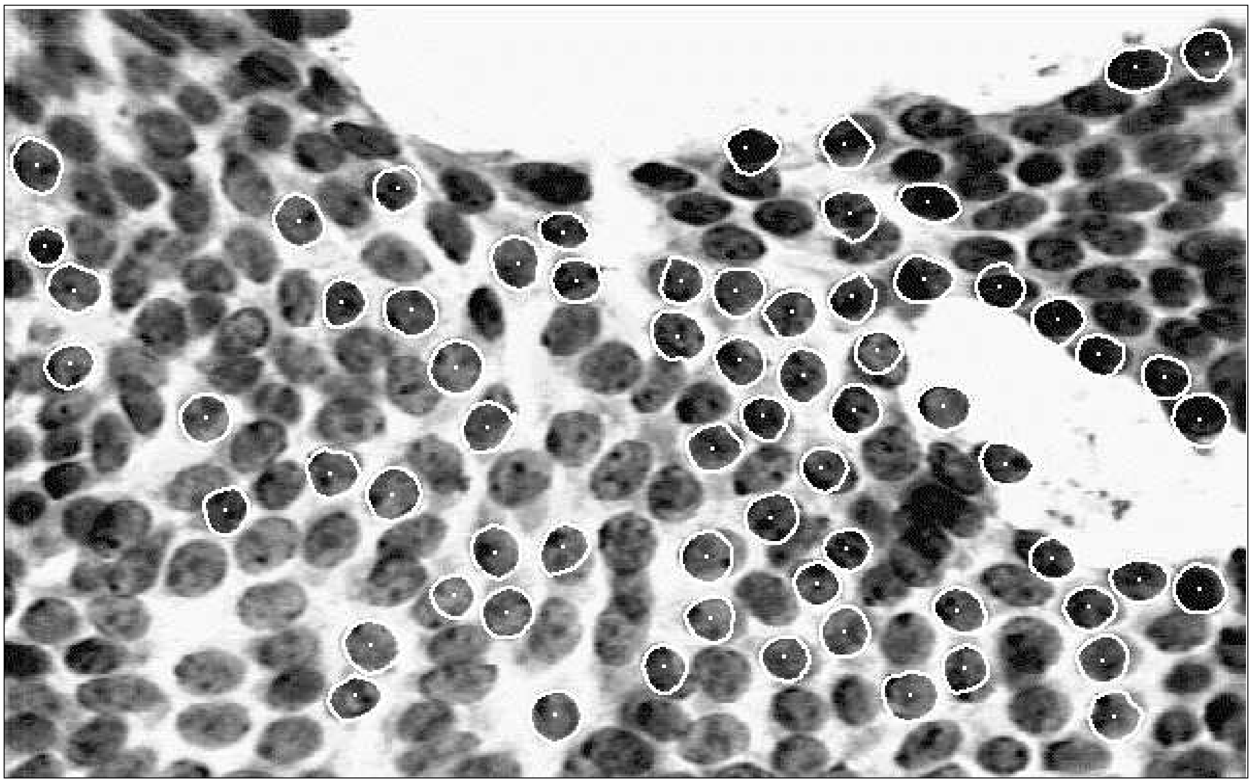
Figure 4: Case of false boundary

To find proper initial points, we use the edge information since the exact boundary should be placed on an edge. But using edge-based techniques may fail due to the similarity of objects within the scene or indistinct boundaries between image objects [10]. Combining the snake with a local edge-based tuning algorithm improves the resulting accuracy and robustness. In this work, we simply look for the nearest edge from each point along the template and then consider this a new initial point of snake.

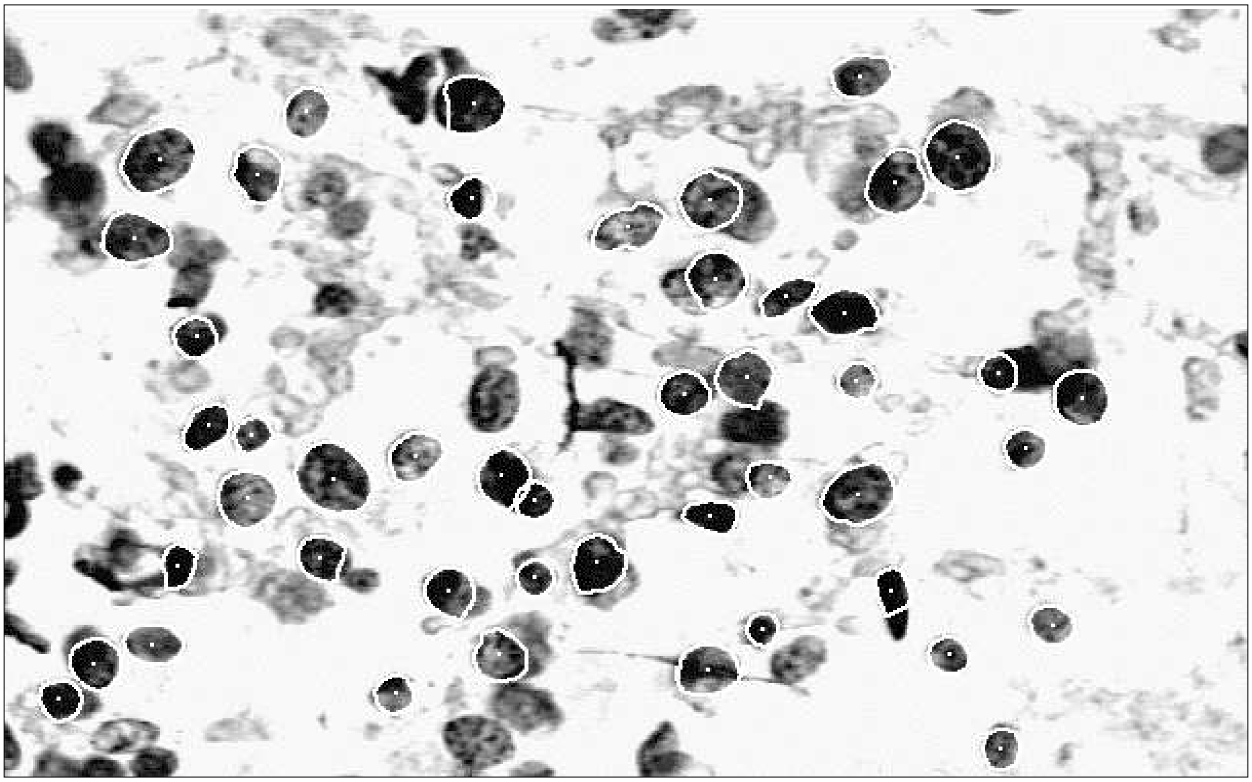
3 Experimental Results

This algorithm was tested on a set of cytologic images, each of which is gray-scaled with 640×400 pixel spatial resolution. Using built-in methods in Java images were reduced to 320×200 as a half-sized image and 160×100 as a quarter-sized image. The images contain various numbers of nuclei with radius ranging from 8 up to 32 pixels. Thus we used templates with size 4 to 16 for a half-sized image and 2 to 8 for a quarter-sized image. Figure 5 shows the results of two images that were tested.

Figure 6 (a) shows the computation time (in seconds) for the segmentation process with iterative GHT and our algorithm with image scaling. Because the presented algorithm performs the second iterative GHT using the number of templates detected in the quarter-sized image, the computation time varies depending on the image. So the computation time of the image in Figure 5(a) which

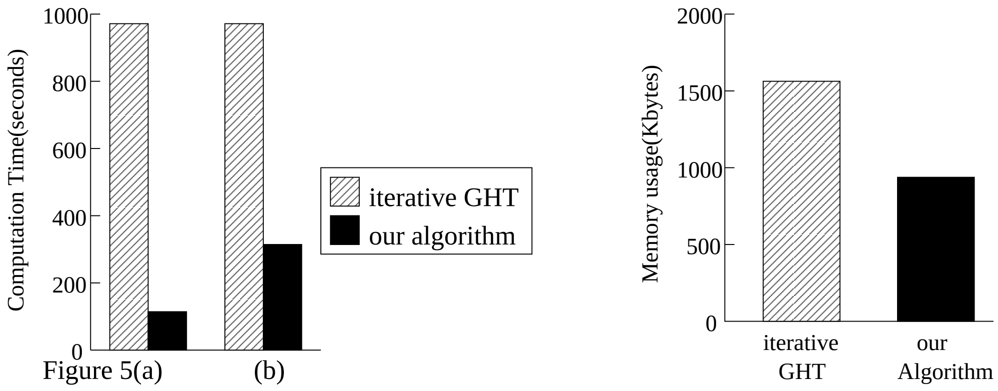


(a)



(b)

Figure 5: Result images : (a) image which includes similar-sized nuclei, radius 11 to 13, (b) image which includes various-sized nuclei, radius 8 to 15



(a) Time comparison

(b) Memory comparison

Figure 6: Results

includes similar-sized nuclei, radius 11 to 13 is smaller than that of the image in Figure 5(b) which includes various-sized nuclei, radius 8 to 15. Estimated memory usage is plotted in Figure 6 (b). Since the global accumulator with image scaling is smaller than without image scaling, memory is saved even after the additional accumulators. We note that memory usage for a traditional five-dimensional GHT covering the same space of ellipses would require 1242 Mbytes, more than 1250 times the memory used by our system.

We used two measures to evaluate the performance of the algorithm: sensitivity and accuracy. Sensitivity is the likelihood that a nucleus will be detected in case it has marked by a expert. It is defined as follows:

$$Sensitivity = \frac{TP}{TP+FN}$$

where TP (true positive) are those nuclei which have been segmented by the users and also detected by the proposed algorithm and FN (false negative) are those nuclei which have been marked but are not detected by the system. Accuracy is defined as the likelihood that the segmentation of a nucleus is actually associated with a nucleus segmented by a user.

$$Accuracy = \frac{TP}{TP+FP}$$

where FP (false positive) are those nuclei which are segmented by the algorithm but have not been segmented by the user [8]. As shown in Figure 7, the two algorithms produce very similar results.

4 Conclusion

In this work we described a cytologic image segmentation algorithm to reduce the excessive time with image scaling and to determine the exact nuclear boundaries with snakes and local tuning. The proposed algorithm efficiently exploits both region and edge information. The results show that it

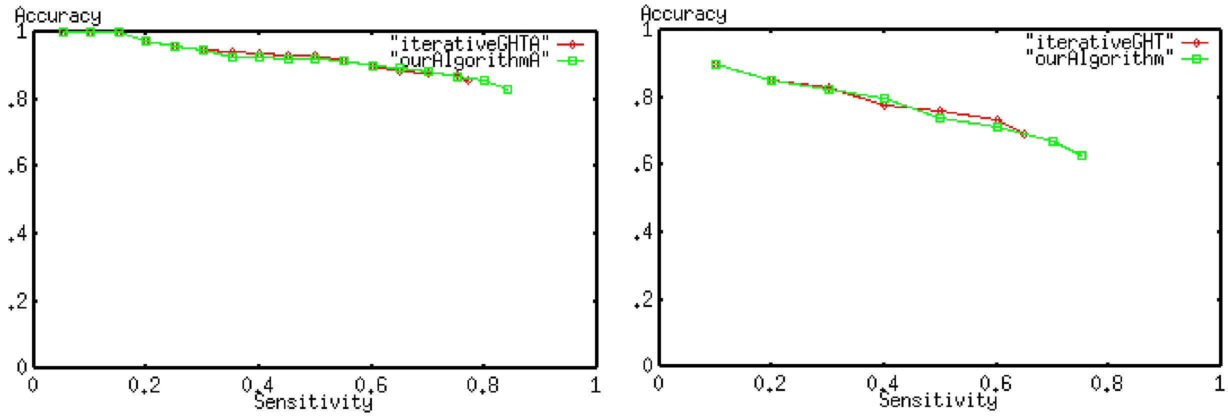


Figure 7: Sensitivity and accuracy: (a) image shown Figure 5(a), (b) image shown in Figure 5(b)

is a reliable method for segmenting nuclei in cytological images and for extracting components of interest, which is a key step for diagnosing breast cancer and predicting the course of the disease.

5 Acknowledgements

The authors wish to thank to the following people for their contributions to the Xcyt project: W.H. Wolberg, O.L. Mangasarian, H. Oh, and S.R.K.R. Mallina. This work was supported in part of NIH grant CA64399-04 and NSF grant IIS-99-96044.

References

[1] Aguado, A.S. and Nixon, M.S., "A new Hough transform mapping for ellipse detection", *Technical report*, University of Southampton, Department of Electronics and Computer Science, 1995. 1995/6 Research Journal.

[2] Ballard, D.H., "Generalizing the Hough transform to detect arbitrary shapes", *Pattern Recognition*, vol. 13, no. 2, pp. 111-122, 1981.

[3] Chalana, V., Costa, W. and Kim, Y., "Integrating region and edge information using regularization", *Proceedings of the SPIE Conference on Medical Imaging*, SPIE, 1995.

[4] Kass, M., Witkin, A. and Terzopoulos, D., "Snakes: Active contour models", *International Journal of Computer Vision*, pp. 321-331, 1988.

[5] Mangasarian, O.L., Street, W.N. and Wolberg, W.H., "Breast cancer diagnosis and prognosis via linear programming", *Operations Research*, vol. 43, no. 4, pp. 570-576, July-August 1995.

[6] Oh, H. and Street, W.N., "A memory-efficient generalized Hough transform for segmenting cytological images", 1998, under review.

[7] Rowley, H.A. and Kanade. T., "Neural network-based face detection", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 1, pp. 23-38, Jan. 1998.

[8] Tsapatsoulis, N., Schnorrenberg, F., Pattichis, C., Kollias, S. and Schizas, C., "A neural network assisted algorithm for automated detection of breast cancer nuclei", *Proceeding of 13th International Conference Digital Signal Processing*, Santorini, 1997.

[9] Wolberg, W.H., Street, W.N. and Mangasarian, O.L., "Machine learning techniques to diagnose breast cancer from image-processed nuclear features of fine needle aspirates", *Cancer Letters*, vol. 77, pp. 163-171, 1994.

[10] Xuan, J., Adali, T. and Wang, Y., "Segmentation of magnetic resonance brain image: Integrating region growing and detection", *International Conference of Image Processing*, pp. 544-547, 1995.

[11] Yuen, H.K., Illingworth, J. and Kittler J., "Detecting partially occluded ellipses using the Hough transform", *Image and Vision Computing*, vol. 7, no. 1, pp. 31-37, Feb. 1989.