Journal of Biomedical Informatics **34,** 67–73 (2001)
doi:10.1006/jbin.2001.1009, available online at http://www.idealibrary.com on IDEAL®

Faguo Yang*,[1] and Tianzi Jiang*,†,[2]

National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences,
Beijing 100080, People's Republic of China; and †School of Computer Science,
Queen's University of Belfast, Belfast BT7 1NN, United Kingdom

Received November 15, 2000; published online June 6, 2001

In this paper, we propose a novel approach to cell image segmentation under severe noise conditions by combining kernel-based dynamic clustering and a genetic algorithm. Our method incorporates a priori knowledge about cell shape. That is, an elliptical cell contour model is introduced to describe the boundary of the cell. Our method consists of the following components: (1) obtain the gradient image; (2) use the gradient image to obtain points which possibly belong to cell boundaries; (3) adjust the parameters of the elliptical cell boundary model to match the cell contour using a genetic algorithm. The method is tested on images of noisy human thyroid and small intestine cells. © 2001 Academic Press

Key Words: cell images; image segmentation; quantitative pathology; genetic algorithms.

1. INTRODUCTION

Pathologists often make diagnostic decisions by observing specimen cells, in particular, the geometric parameters of the cell such as the area, radius, and the circumference [1]. Thus, in an automated system it is very useful to accurately measure the geometric parameters of the cell.

As a precursor to accurate segmentation, shape modeling of objects is required. In past years, many methods for the segmentation of cell images have been presented [1–7]. These methods include region-based methods, threshold-based methods, and so on. Region-based methods employ region growing and region splitting and merging to segment the image. Threshold-based segmentation is a simple method based on single pixel classification; a feature value such as gray level is associated with each pixel; this value is compared to the threshold to classify a pixel as an object or background. The critical part is determination of the threshold; a simple method is to select threshold based on major value with histogram. More sophisticated version of this are given in Refs. [8, 9]. The problem with these is that they employ only local (single pixel) information; no a priori knowledge of object or region shape is brought to bear; such a priori knowledge can vastly improve segmentation in the presence of noise and other image corruption, resulting in poor conclusion between object and background. Cell images share the following characteristics:

- Poor contrast, i.e., object (cell) gray levels may be close to that of background;

[1] E-mail: fgyang@nlpr.ia.ac.cn.

[2] To whom correspondence should be addressed at National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100080, People's Republic of China. E-mail: jiangz@nlpr.ia.ac.cn.

1532-0464/01 $35.00

- Many cluttered objects in a single scene. A high number of occluding objects make image segmentation difficult;
- Low quality. Traditional staining techniques introduce a lot of inhomogeneity into the images, where not all of the parts of the same tissue are equally stained.

In the presence of noise, clutter, and occlusion, the segmentation of the cell image is an ill-posed problem [10]. However, as is usual with ill-posed problems, constraints may be imposed through a priori knowledge. That is, the segmentation performance can be greatly improved by incorporating a priori knowledge about shapes of cells. In fact, one of the most challenging issues in medical image segmentation is to extend traditional approaches of segmentation and object classification in order to include shape information rather than merely image intensity. In this paper, we not only use the edge information but also the shape information of the cell images to accurately segment the cell images. We also apply a genetic algorithm to segment the cell images. Our edge-based method consists of the following three parts: (1) detecting the possible edges of the cell; (2) locating the position of the cells approximately and finding out the image points, which most likely belong to each cell boundary; (3) constructing a cell contour model characterized by five parameters to accurately detect the cell contour and to eliminate the influence of the image noise. We can accurately describe the cell contour by adjusting the parameters of the cell boundary model. Thus, the image segmentation problem is finally transformed to an optimization problem.

The outline of this paper is as follows. Section 2 introduces the mathematical model used to describe the boundary of a cell. Section 3 describes a method by which we can obtain the edge information and the possible image points belonging to each cell boundary. In Section 4, we introduce the algorithm to optimize the parameters of the mathematical model to match the cell contour. Section 5 presents experimental results. Section 6 gives conclusions and a description of future work.

2. CONTOUR MODEL OF CELL IMAGE

Most of the cells in the human body usually have elliptically shaped boundaries as shown in Figs. 1a and 2a. Here, we can see that the gray levels of the cells are lower than those of the background, but in general, the contrast is poor. Moreover, there is a lot of noise, clutter, and occlusion in the image. As has been indicated previously, the ill-posed problem of this form may be solved by imposed parameter constraints in the form of a priori information, In a model-based approach, cell image segmentation can be cast as a parameter optimization problem. If the model parameters of a cell boundary are determined, we can reconstruct the segmented image, which is used to extract meaningful geometric parameters for pathologists.

In order to impose such a priori knowledge, we use an ellipse equation to describe the boundary of a cell as follows:

$$\frac{[(x-x_0)\cos\theta+(y-y_0)\sin\theta]^2}{a^2} + \frac{[(x-x_0)\sin\theta+(y-y_0)\cos\theta]^2}{b^2} = 1 \qquad (1)$$

Equation (1) denotes an ellipse in the (x, y) domain, (x_0, y_0) determines the center of the ellipse, θ indicates the orientation of the ellipse, and a, b denote the major and minor of the ellipse. Thus, we have five parameters: x_0, y_0, a, b, and θ. Therefore, given five parameters to be determined, from boundary points are sufficient to determine them.

3. A NOVEL APPROACH TO DETECTION OF IMAGE POINTS IN CELL

3.1. Localization of Cells

Traditionally, edge-based segmentation has been divided into two independent stages: edge detection and edge linking. Under Marr's paradigm [11], boundary extraction is conventionally treated as a set of independent problems, where each problem has input information, a method to process them, and output information. This one-way flow of the information may yield wrong results because of error propagation. Consequently, we adopt a different scheme in this paper. In our scheme, we first find those image points that have a high probability belonging to a cell; these are used to determine an approximate location of the cell. This process can be viewed as getting high-level information from low-level information. Based on the approximate location and on an approximate model of a cell, we may then reevaluate decisions on whether pixels belong to the cell boundary. Thus, the method has a relaxation aspect. For edge detection, we use Canny's edge detector [12] to detect the image edges. Because of the influence of the noise, there are a lot of false directions in the gradient image (see Figs. 1b and 2b). Some of these edge points are connected. From the original cell images

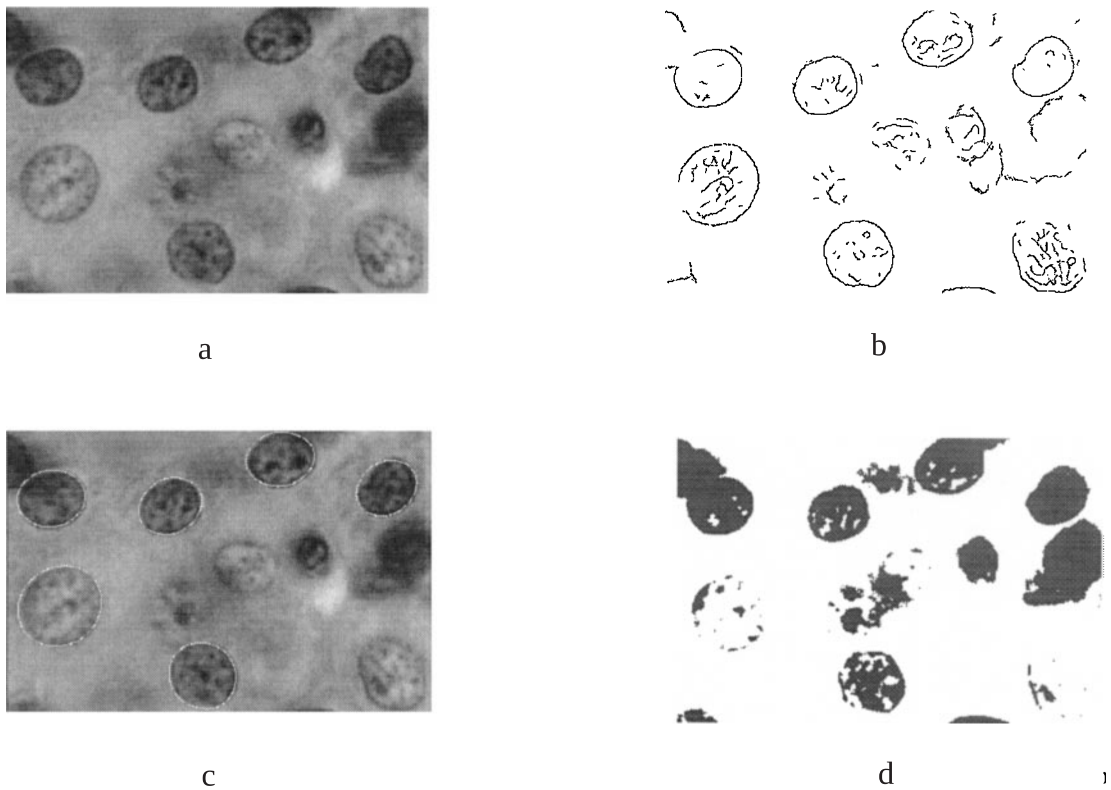

FIG. 1. Segmentation results of hypothyroid image. (a) Original image. (b) Edge image. (c) Our approach. (d) Histogram-based method.

(see Figs.1a and 2a), we can see that the variation of the gray levels near the cell boundary is large. Consequently, the edges formed due to image noise have small number of connected image points. Thus, we can use a threshold to decide whether an edge direction indicates a cell boundary. If the number of the connected image points is greater than the threshold, the edge becomes a potential boundary point. Otherwise, we reject the edge as due noise. The threshold can be selected by experience.

3.2. Kernel-Based Dynamic Clustering for Detection of Image Points in Cell

After detecting the approximate location of the cells, we use a kernel-based dynamic clustering method to find out the image points possibly belonging to each cell. In this method, a kernel k_j, which represents a cluster, is defined. A kernel can be a function, an image point sets or other models. In order to determine whether a sample point belongs to a cluster, a measurement $\Delta(y, k_j)$, which describes the similarity between sample point y and cluster k_j, is also defined. The steps involved in the method are as follows:

1. Determine an initial kernel k_j for each cluster.
2. For all sample points, follow the following rules to classify them: if, and $\Delta(y, k_j) \leq t$, then $y \in \gamma_j$, where y denotes a sample point, t is a threshold, and γ_j represents the jth cluster.
3. Update the kernel k_j. If all of the kernels do not change again, stop the iteration; otherwise return to the step 2.

Since the cell has an ellipse-like boundary, we use Gaussian function as the kernel, which can be formulated as

$$K_j(y) = \frac{1}{2\pi\|\Omega_j\|} \exp -\frac{1}{2}(y - m_j)\,\Omega_j^{-1}(y - m_j)^T \quad (2)$$

where m_j is the mean value of the samples and Ω_j is the covariance matrix.

The similarity between sample point y and cluster j is defined as follows:

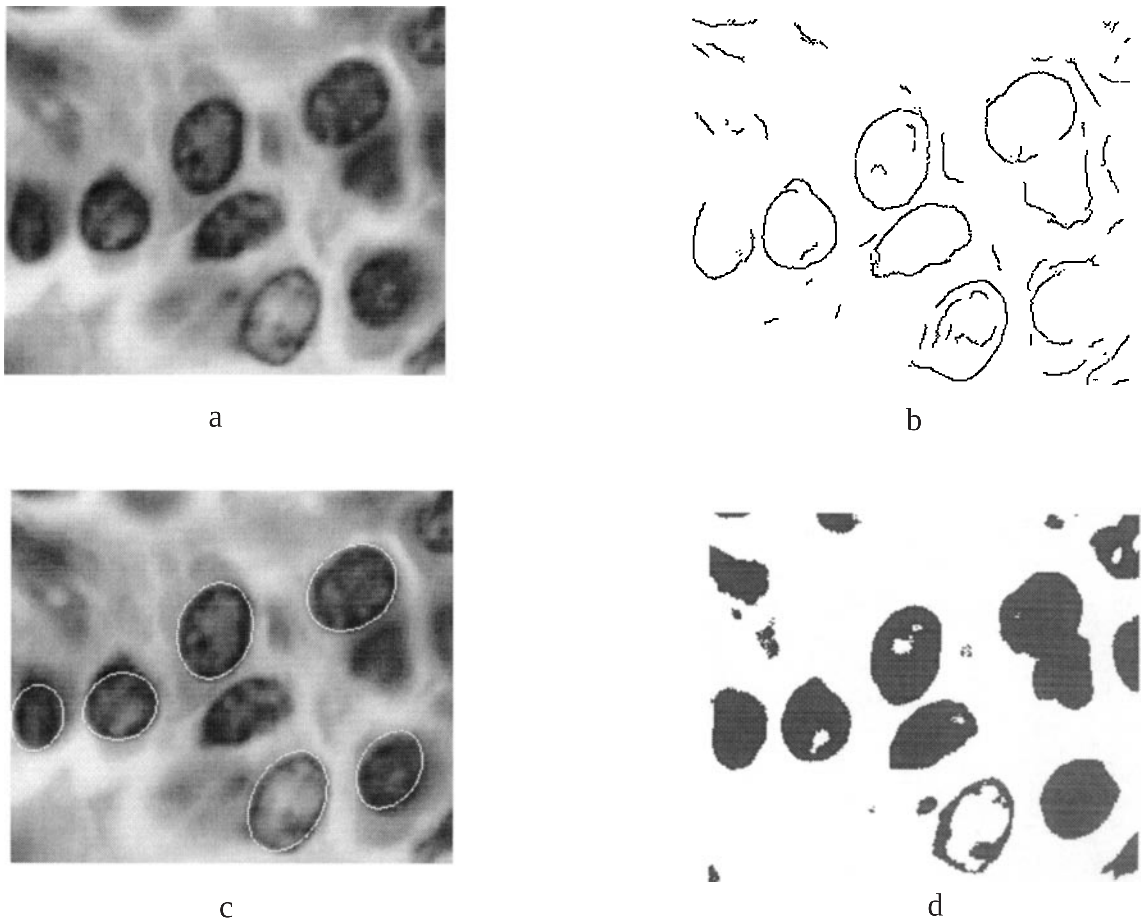

FIG. 2. Segmentation results of small intestine image. (a) Original image. (b) Edge image. (c) Our approach. (d) Histogram-based method.

$$\Delta(y, k_j) = \frac{1}{2}(y - m_j)\Omega_j^{-1}(y - m_j)^T + \frac{1}{2}\log\|\Omega_j\| \quad (3)$$

Using the above method, we can find out the image points that have a high probability belonging to each cell respectively. After the image points possibly belonging to a cell are detected, we can search an ellipse that best matching the cell boundary in a relatively small space. So, the time spent in finding the solution is decreased greatly.

4. EXTRACTING CELL BOUNDARY USING A GENETIC ALGORITHM

4.1. Genetic Algorithms

Genetic algorithms have proved to be effective for many optimization problems [13, 14]. They use heuristics to find global optima, and avoiding local optima. Here we briefly review some concepts of genetic algorithms and outline the basic steps of genetic algorithms for solving optimization problems. In order to solve an optimization problem using a genetic algorithm, we must do the following three basic things:

1. Define a representation. We should use a representation that is minimal yet completely expressive. Our representation should be able to represent any solution to our problem, but if at all possible we should design it so that it cannot represent infeasible solutions to our problem. If infeasible solutions are possible, then the objective function must be designed to penalize them approximately.

2. Define the genetic operators. The basic genetic operators include initialization, mutation, and crossover. The initialization operator defines how to create many initial solutions to the specific problem. The mutation operator tells

how to change some parts of a solution to produce a new solution to the problem. The crossover operator uses two existing solutions determine a new solution.

3. Define the objective function. Genetic algorithms are often more attractive than gradient search methods because they do not require complicated differential equations or a smooth search space. Genetic algorithms need only a single measure of how good a single individual is compared to other individuals.

A genetic algorithm consists of the following.

I. Data

(1) Chromosome: A chromosome is a solution to a specific optimization problem.

(2) Population: Population is a group of chromosomes to an optimization problem.

II. Function

(1) Fitness: Fitness is used to express how good a chromosome (an individual) is compared to others.

(2) Crossover: Crossover is a genetic operator that determines how to obtain one or two new chromosomes according to two existing solutions.

(3) Mutation: Mutation is a genetic operator that tells how to change an existing chromosome to form a new one.

III. Parameters

(1) Crossover probability: The probability for two existing chromosomes to crossover to produce two new chromosomes.

(2) Mutation probability: The probability for an existing chromosome to alter itself to produce a new chromosome.

With the above notations, we can describe the basic steps of a genetic algorithm as follows:

Step 1. Use an initializing genetic operator to initialize the population.

Step 2. Select some chromosomes to crossover to produce a group of new chromosomes.

Step 3. According to the mutation probability, mutate the offspring produced by the crossover operator.

Step 4. Insert the offspring into the population. Determine whether the stopping criteria are satisfied. If satisfied, stop the iteration; otherwise go to the Step 2.

4.2. Extraction of Cell Boundary with a Genetic Algorithm

In this section, we present a genetic algorithm for extracting the cell boundaries. The input information is N possible edge points of the cell. Each of these image points has an associated index, which is a number from 1 to N. We can select five of them to represent an ellipse. Let $I = (I_1, I_2, \ldots, I_5)$ denote the index of the five image points selected. For the sake of convenience, we assume that $I_i < I_j$ for $i < j$, $i, j = 1, \ldots, 5$. The objective function that we define is as follows [15]:

$$f(I) = \sum s(r_{I_i}^2) \tag{4}$$

where s is the step function: $s = 1$ if r_j is greater than or equal to the template width, and $s = 0$ otherwise. The objective function counts the number of points within a fixed distance of the ellipse. Let P denote the population with M chromosomes $I_1, I_2, \ldots, I_M$. Let L denote the size of the population. Let NUM represent the number of generations. Let N_r denote the number of chromosomes replaced in each generation. For convenience, let N_r be an even number. The genetic algorithm-based method for cell boundary extracting can be described as follows:

Step 1. Initialization: To initialize each chromosome, randomly generate five integers between 1 and N (N is the number of the possible edge image points) as the index of five image points. We repeat the process for L times to initialize the whole population. Let $k = 1$, where k is the iteration step indices.

Step 2. Selection and crossover: We use the fitness proportional model to select chromosomes to reproduce offspring. To obtain N_r children, we should select $N_\mathrm{r}/2$ mothers and fathers, respectively. Each pair of parents has two children. The crossover operator is described as follows: an integer m between 1 and 5 is randomly generated; exchange the genes that have the index m of the two chromosomes selected as parents. So, N_r new chromosomes are reproduced.

Step 3. Mutation: For all of the newly generated chromosomes, according to a mutation probability to mutate them. The mutation method is described as follows: an integer m between 1 and 5 and another integer n between 1 and N are randomly generated. The gene having the index m of the chromosome is replaced by the integer n.

Step 4. Replacement scheme: We use the newly generated N_r chromosomes to replace those having a low fitness. If the stop criterion (see Remark) is satisfied, stop the iteration; otherwise, go to the Step 2.

Remark: *First, we compute the fitness of each chromosome and the fitness can be viewed as a distribution function. Then we compute the entropy of the distribution function. When the entropy does not change again, we decide that the genetic algorithm has converged, and we stop. If the number of generations N is reached, we also stop.*

5. EXPERIMENTAL RESULTS

In this section, we present our experimental results on segmenting the cell images of hypothyroid and small intestine and make a comparison with the histogram-based methods. The histogram-based method can be described as follows: (1) calculate the histogram of the image to be segmented. (2) According to the histogram of the image, try to find a proper threshold to segment the image. (3) If the gray level of a pixel is smaller or equal to the threshold, the pixel will be classified to the object cluster; otherwise the pixel will be classified into the background. The threshold can be found out through a minimal error method or a maximum entropy method [8, 9]. In our experiment, we let the crossover probability $P_c = 0.6$, the mutation probability $P_m = 0.1$, the number of generation NUM = 250, and the population size $L = 100$. Figures 1c and 2c are our experimental results on segmenting the cell images of hypothyroid and small intestine respectively. The edge images are obtained using Canny’s edge detector. The parameters used in Canny’s edge detector are as follows: The higher threshold is 0.9 and the lower threshold is 0.8; the covariance is 1.6.

The experimental results with the threshold methods are also given out in Figs. 1d and 2d. From Figs. 1 and 2, we can see the advantages of our method compared to the histogram-based methods. Our approach is very tolerant of noise. As long as the gray levels near the cell boundary differ by small amount, our method can actually detect the cell boundary and complete the segmentation task well. When two cells are located very closely, the simple histogram-based methods cannot distinguish them, but our method can extract them successfully. So our approach has some ability to cope with occlusions of the cell boundaries. The time required to find out the image points belonging to each cell is about 30 s, based on a 550-MHz processor. The time spent in adjusting the parameters of the cell contour model to match the cell boundary well is also about 30 s. So, total time needed to process an image like Fig. 1a is about 1 min. Because we first use kernel-based dynamic clustering method to find out which image points possibly belonging to each cell, the search space is reduced greatly and the time spent in finding the optimum ellipse to match the cell boundary is decreased accordingly.

6. CONCLUSION AND FUTURE WORK

In this paper, we have proposed a novel approach to cell image segmentation under severe noise conditions by combining kernel-based dynamic clustering and a genetic algorithm. In our algorithm, we make use of not only the edge information but also a model that states that the cell boundary has an ellipse shape. Due to the a priori (model) knowledge of the cell boundary being incorporated in our approach, our method has a high ability to resist noise. Image points that possibly belonging to a cell are first determined using a kernel-based dynamic clustering method, so the search space is greatly reduced and the time used to optimize the objective function is decreased accordingly. The results obtained indicate a promising direction for further research into automatic initialization, which is especially important for designing automatic algorithm in biomedical applications. The experimental results have shown that our approach is effective and efficient.

We have shown that our method is capable handling partial occlusions. However, more work is needed to solve severe occlusions.

ACKNOWLEDGMENTS

The authors are highly grateful to two anonymous referees for their significant and constructive comments and suggestions, which improve the paper very much. The second author also thanks his colleague, Dr. Jonathan G. Campbell, at the Queen’s University of Belfast for carefully reading this paper and valuable suggestions. This work was partially supported by the Natural Science Foundation of China, Grants 69705002 and 697908001, and the State Commission of Science and Technology of China, Grant G1998030503.

REFERENCES

1. Wu HS, Barba J, Gil J. A parametric fitting algorithm for segmentation of cell images. IEEE Trans Biomed Eng 1998; 45:400–7.

2. Garbay C. Image structure representation and processing discussion of some segmentation methods in cytology. IEEE Trans Pattern Anal Mach Intell 1986; 8:140–7.
3. Garrido A, Perez N. Applying deformable templates for cell image segmentation. Pattern Recognit 2000; 33:821–32.
4. Mouroutis T, Roberts SJ, Bharath AA. Robust cell nuclei segmentation using statistical modeling. BioImaging 1998; 6:79–91.
5. Simon I, Pound CR, Partin AW, Clemens JQ, Christensbarry WA. Automated image analysis system for detecting boundaries of live prostate cancer cells. Cytometry 1998; 31:287–94.
6. Wu HS, Gil J. An iterative algorithm for cell segmentation using short-time Fourier transform. J Microsc 1996; 184:127–32.
7. Wu HS, Barba J, Gil J. Iterative thresholding for segmentation of cells from noisy images. J Microsc 2000; 197:296–304.
8. Kapur JN, Sahoo PK, Wong AKC. A new method for gray-level picture thresholding using the entropy of the histogram. Comput Vis Graph Image Process 1985; 29:273–85.
9. Kittler J, Illingworth J. Minimum error thresholding. Pattern Recognit 1986; 19:41–7.
10. Poggio T, Torre V. Ill-posed problems and regularization analysis in early vision. Proc AARPA Image Understanding Workshop 1984; 257–63.
11. Marr D. Vision: a computational investigation into the human representation and processing of visual information. San Francisco, CA: Freeman, 1982.
12. Canny J. A computational approach to edge detection. IEEE Trans Pattern Anal Mach Intell 1986; 8:679–98.
13. Goldberg DE. Genetic algorithm in search, optimization and machine learning. Reading, MA: Addison Wesley, 1989.
14. Rudolph G. Convergence analysis of canonical genetic algorithm. IEEE Trans Neural Network 1994; 5:96–101.
15. Ke Q, Jiang T, Ma S. A tabu search method for geometric primitive extraction. Pattern Recognit Lett 1997; 18:1443–51.