SEGMENTING MICROARRAY IMAGE SPOTS USING AN ACTIVE CONTOUR APPROACH

Jinn Ho and Wen-Liang Hwang

Institute of Information Science, Academia Sinica, Taiwan

ABSTRACT

Inspired by Paragious and Deriche's work, which unifies boundary-based and region-based image partition approaches, we integrate the active contour(snake) model and the Fisher criterion to capture, respectively, the boundary and region information of microarray images. We then use the proposed algorithm to automatically segment the spots in the microarray images, and compare our results with those obtained by commercial software.

Index Terms— Microarray, Active contour, Snake, Fisher criterion

1. INTRODUCTION

As the DNA microarray can simultaneously measure thousands of gene expression levels on the genomic scale, it has enormous potential for biological, medical, and industrial applications [16, 7]. The fragments of genes, are spotted or printed on an array matrix as probes to detect gene expressions. Image processing techniques and statistical methods are applied to determine the expression levels of the spots in the microarrays in order to perform gene expression analysis.

According to Yang et al. [17], the processing of microarray images involves spot gridding, segmentation, and intensity extraction. The spot gridding task detects the positions of the spot centers and identifies their coordinates [2]. Existing commercial software provides semi-automatic algorithms to deal with the problem. An accurate and automatic algorithm for the case where the spot centers are smoothly distorted is provided in [11].

The goal of segmentation is to classify a pixel as either foreground inside a spot, or as background outside the spot. A number of segmentation techniques have been proposed [14, 13], some of which assume that the geometry of the spots is either a fixed circle or an adaptive circle [8]. However, the assumption is incorrect because a spot's morphology is not always a circle. Other techniques use hypothesis testing to segment the foreground and background [5], but this requires modeling the pixel intensity distributions, which is a difficult problem. Region growing based on the watershed algorithm proposed in *Spot* [17] can segment regions of irregular shape and does not need to model a region's probability; however, the segmentation results are not necessarily an optimization of some class separation criteria. The objective of the intensity extraction task is to calculate and normalize spot intensity in order to derive quality measurements [15]. The segmentation task is the focus of the present study.

For the spot segmentation task, we propose using the snake model to capture boundary information and the Fisher criterion to capture region information. The snake model [12] is very effective in segmenting objects whose boundaries can be approximately delineated by a set of large gradient points along a contour. The spot boundary is such an example. The Fisher criterion is based on discriminate analysis in statistics, which uses between-class and within-class statistics to form a criterion for class separation [9]. We adopt the Fisher criterion because it is simple and can be analyzed mathematically.

The operation of a snake model is only semi-automatic and the solution depends on the initial contour and the parameter values, both of which must be determined manually. The difficulty of solving the problems of selecting good initial contours and parameters for images with various signal-to-noise (SNR) levels prevents the model from operating automatically. Because of the enormous number of spots in microarray images, a semi-automatic process severely degrades the throughput of microarray analysis. To resolve these difficulties, we first modify the Markov chain Monte Carlo-based Climber algorithm to find a good initial contour [3], and then estimate the values of our parameters from that contour. Experiments on several synthesized and natural images show that it can find a good initial contour and estimate quality parameters. Using the proposed method, we segment the spots in microarray images with various SNR levels, and compare our results with those of $GenePix\ Pro$ 5.0 [8] and $Spot$ 2.0 [17].

The remainder of the paper is organized as follows. In the next section, we introduce our model. In Section 3, we present an automatic algorithm that finds a solution for our model. In Section 4, we validate our model by comparing it with other approaches. Finally, in Section 5 we present our conclusions.

2. DESCRIPTION OF THE MODEL

For simplicity, we assume that there are only two regions to be delaminated. However, the proposed model can be used to detect more than two regions simultaneously.

A. Energy Form

We define $R = \{I(x,y)\}$ as an image of gray value pixels. A simple closed curve $\Gamma = \Gamma(s)$ on R divides the image into $\{R_1, R_2\}$, where $R = R_1 \cup R_2$ and $\Gamma = \partial R_1 \cap \partial R_2$. We denote M_1 and M_2 as the expected values of pixels in R_1 and R_2, respectively. The total energy induced by contour Γ is defined as the sum of the snake's energy and the region's energy. The former measures the properties along the contour, while the latter measures the statistical differences between the regions separated by the contour. The total energy E_{total} is written as

$$E_{total} = E_{snake} + \tilde{\gamma} E_{region}. \quad (1)$$

in which

$$E_{snake}(\Gamma) = \int_{\Gamma} (\frac{\alpha}{2}|\Gamma_s|^2 + \frac{\beta}{2}|\Gamma_{ss}|^2 - \| \bigtriangledown I\|^2) ds \quad (2)$$

We use the two-class Fisher discriminate criterion to represent E_{region} as $E_{region} = E_{within}/E_{between}$. where

$$\begin{aligned} E_{within} &= \iint_{R_1} (I - M_1)^2 + \iint_{R_2} (I - M_2)^2 \\ E_{between} &= (M_1 - M_2)^2 \end{aligned} \quad (3)$$

The within-class distance E_{within} measures the scatter of samples in R_1 and R_2 around their expected values and the between-class distance $E_{between}$ is the difference between the expected gray levels of R_1 and R_2.

We propose an iterative algorithm to find the solution contour of (1). The algorithm begins with an initial contour; then, at each iteration, a new contour is obtained by alternating the subsequent stages. In the first stage, by fixing the values of the between-class distance $E_{between}(\Gamma)$ and the model's parameters α, β, and $\tilde{\gamma}$, the algorithm finds the curve $\hat{\Gamma}$ that minimizes E_{total}. In the second stage, we calculate the between-class distance with respect to $\hat{\Gamma}$. The parameter values are then estimated by minimizing the mean square error (MSE) of the Euler equation in (1) with respect to $\hat{\Gamma}$, as shown in Fig. 1.

B. Euler Equation

By applying Green's theorem to E_{within} in (3) and setting $\gamma = \tilde{\gamma}/E_{between}$, we obtain

$$\begin{aligned} E(\Gamma) &= \int_{\Gamma} (\frac{\alpha}{2}|\Gamma_s|^2 + \frac{\beta}{2}|\Gamma_{ss}|^2 - \| \bigtriangledown I\|^2) ds \\ &\quad + \gamma \int_{\Gamma} L(s; v_s, v_{ss}) ds \\ &= \int_{\Gamma} F(s; v, v_s, v_{ss}) ds, \end{aligned} \quad (4)$$

where

$$\begin{aligned} F(s; v, v_s, v_{ss}) &= (\frac{\alpha}{2}|\Gamma_s|^2 + \frac{\beta}{2}|\Gamma_{ss}|^2 - \| \bigtriangledown I(v)\|^2) \\ &\quad + \gamma L(s; v, v_s), \end{aligned} \quad (5)$$

in which $v : [0,1] \rightarrow \mathbb{R}^2$; $v(s) = (x(s), y(s)) = \Gamma(s)$; and $x, y \in C^2([0,1])$. Using functional calculus the Euler equation for $\beta = 0$ becomes

$$-\frac{\partial \| \bigtriangledown I\|^2}{\partial x} - \alpha x_{ss} + \gamma[(I - M_1)^2 - (I - M_2)^2] y_s = 0, \quad (6)$$

$$-\frac{\partial \| \bigtriangledown I\|^2}{\partial y} - \alpha y_{ss} - \gamma[(I - M_1)^2 - (I - M_2)^2] x_s = 0. \quad (7)$$

The solution of the above can be obtained by an iterative procedure similar to that in [12].Because $\Gamma_{ss} = [x_{ss} \quad y_{ss}] = \kappa \vec{n}$, where κ is the curvature, and $\vec{n}$ is parallel to $[y_s \quad -x_s]$, (6) and (7) can be written as one equation:

$$-\nabla \| \bigtriangledown I\|^2 - \alpha \kappa \vec{n} - \gamma'[\frac{(I - M_1)^2 - (I - M_2)^2}{(M_1 - M_2)^2}]\vec{n} = 0. \quad (8)$$

This equation requires that a point on the optimal contour must satisfy (9) in the tangent direction ($\vec{t}$), and (10) in the normal direction ($\vec{n}$):

$$\nabla \| \bigtriangledown I\|^2 \cdot \vec{t} = 0, \quad (9)$$

$$-\nabla \| \bigtriangledown I\|^2 \cdot \vec{n} = \alpha \kappa + \gamma'[\frac{(I - M_1)^2 - (I - M_2)^2}{(M_1 - M_2)^2}]. \quad (10)$$

Equation (10) indicates that the optimal contour balances three terms: the first term is provided by the normal component of the gradients of the image, the second term is proportional to the curvature, while the last term measures the class separation.

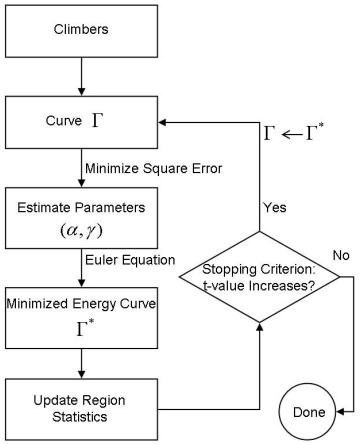

Fig. 1. Block diagram of our approach. First, we use the Climber algorithm to find the initial contour. The parameters are then generated, and the contour's energy is minimized. The process is repeated after the statistics of the regions have been modified in the minimizing energy step.

3. SOLUTION OF OUR MODEL

To solve the Euler equation, we need the initial contour and the model's parameters. First, we describe methods for obtaining the initial contour and estimating the parameters. We then present an iterative algorithm that obtains the solution of our model.

3.1. Initial Contour Detection

The snake-balloon approach tries to solve the initial contour problem by adding an external force to the snake model [6]. We adopt a different approach based on the Climber algorithm [3], derives a stochastic ridge estimation method that is easy to implement and remarkably robust against noise. The algorithm randomly places a large number of independent climbers in a time-frequency plane. Although each climber moves with equal probability in the time direction, it is prevented from moving in the frequency direction. However, it is encouraged to climb to reach the peaks of the local energy functions by a Hastings-Metropolis penalization and a temperature schedule similar to that in the simulated annealing algorithm. Thus, as the temperature approaches zero, the climber settles on a suitable ridge contour. The flowchart of the algorithm is given in Fig. 2 and illustrates the results of applying the Climber algorithm to a heart-shaped image.

3.2. Parameter Estimation

After obtaining the initial contour, we need to determine the values of the parameters. A contour is the solution of our model if we can find the values of the parameters such that the contour and the values satisfy the Euler equation. For the case where there are no suitable values, we estimate the parameters by minimizing the mean-square-error (MSE) of the Euler equation with respect to the contour.

To estimate (α, γ) of a closed curve, we first select the sample pixels Γ_1. Let $\Gamma_1 = \{(x(i), y(i))|\ i = 1, \cdots, s\}$ be the sample points of the given contour, and $K(i) = [(I(x(i), y(i)) - M_1)^2 - (I(x(i), y(i)) - M_2)^2]$. The MSE $e_1^2(\Gamma_1)$, $e_2^2(\Gamma_1)$ of (6) and (7) are, respectively,

$$e_1^2(\Gamma_1) = \frac{1}{s}\sum_i \left[\alpha x_{ss}(i) - \frac{\partial \| \bigtriangledown I\|^2}{\partial x(i)} - \gamma K(i) y_s(i)\right]^2,$$

$$e_2^2(\Gamma_1) = \frac{1}{s}\sum_i \left[\alpha y_{ss}(i) - \frac{\partial \| \bigtriangledown I\|^2}{\partial y(i)} - \gamma K(i) x_s(i)\right]^2.$$

The (α^*, γ^*) that minimizes the MSE satisfies $\frac{\partial e_1^2}{\partial \alpha} = 0$, $\frac{\partial e_1^2}{\partial \gamma} = 0$, $\frac{\partial e_2^2}{\partial \alpha} = 0$, and $\frac{\partial e_2^2}{\partial \gamma} = 0$.

3.3. Alternative Refinement Algorithm

After initial contour detection, the algorithm estimates the optimal parameters that minimize the MSE of the contour, and then solves the Euler equation using the contour and the parameters to obtain a new contour. The statistics of the region partitioned by the new contour are then updated, followed by updating of the parameters. Based on the obtained contour, and the updated statistics and parameters, at each iteration, the algorithm solves the Euler equation and generates a new contour. The process is repeated until a certain stopping criterion is reached. We use the t-test of the interior and exterior regions separated by the contour as the measurement for stopping. The algorithm stops if the t-test value of the current contour is smaller than that of the previous contour.

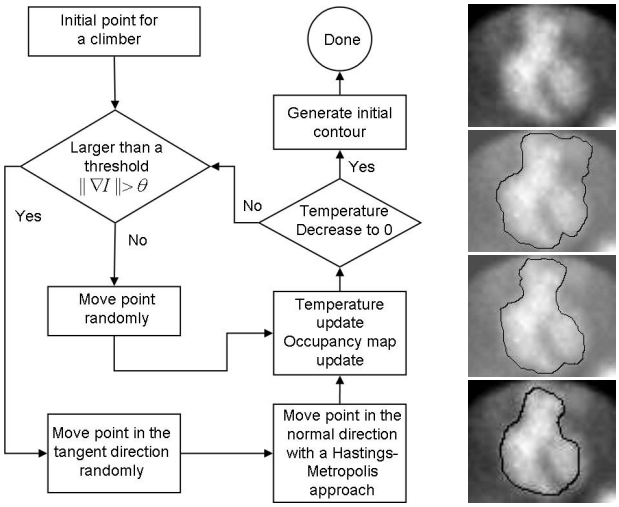

Fig. 2. Left: block diagram of the Climber algorithm. Right: the steps in the evolution of a climber's contour from an input image to the final result.

4. PERFORMANCE EVALUATION

We conduct experiments and evaluate the performance of our algorithm on the microarray images of different manufactured techniques. The experiment parameter for θ is 15% and the threshold for obtaining $\hat{C}$ is the top 10% of the occupation measure in C.

We evaluate and compare our spot segmentation results with those obtained by other algorithms for two sets of microarray images. One set contains some poor quality images from the Stanford Microarray Database (SMD) [10], while the other contains Agilent 60-mer oligonucleotide microarrays whose specifications are given on the related web pages [1]. In the experiments, we separate each spot region from adjacent regions manually, and process the segmentation algorithm inside each region. To evaluate the performance of the proposed algorithm, we compare it with the representative image analysis methods and software in $GenePix\ Pro\ 5.0$, which detects spots by circular boundary adjustment, and $Spot\ 2.0$, which detects spot regions by seed region growing. For the different segmentation results, we calculate the two-sample t-test value between the gray level pixels in the foreground and background, and use it to assess the performance of a segmentation algorithm. As shown by the figures, the distributions of the t-values of our method are statistically larger that those of the other methods, which indicates that the contours derived by our method generally yield better seg-

mentation results.

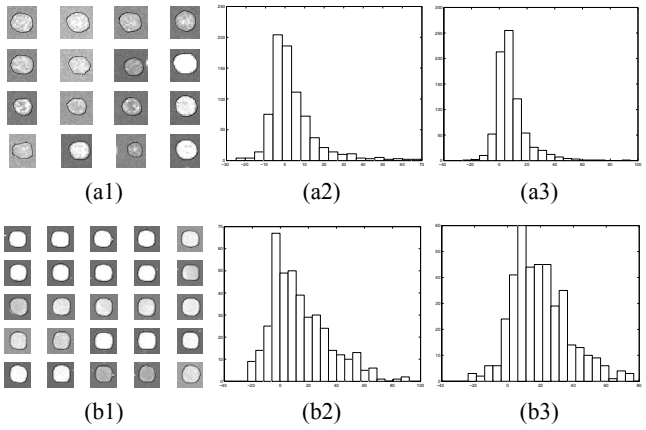

Fig. 3. Comparison of the t-test values of different methods. The data set comprises the spots of subblock $(2, 1)$ of the $LC23N085$ microarray image, which contains $784 = 28 \times 28$ spots, where (a1)shows some results of our algorithm applying on different spots. (a2)shows the histogram of the t-value difference between our algorithm and that of $Spot$. (a3)shows the histogram of the t-value difference between our algorithm and that of $GenePix\ Pro\ 5.0$; (b1) (b2) (b3) are the corresponding results on a subblock of an oligonucleotide microarray image which contains $400 = 20 \times 20$ spots.

5. CONCLUSION

We have integrated the snake model and the Fisher criterion to segment spots in microarray images. The initial contour is obtained by the robust Climber algorithm. The segmentation problem is then solved with an iterative algorithm, where the parameters of our model and the contour are modified alternately until the t-value of the regions cannot be improved further. The proposed algorithm's performance is superior because it is automatic and can segment the spots of microarray images without human intervention. The experiment results on microarray data manufactured by different techniques also demonstrate that our algorithm outperforms other methods.

6. ACKNOWLEDGMENTS

We would like to express our gratitude to Professor Wen-Hsiung Li of the University of Chicago for insightful suggestions.

7. REFERENCES

[1] Agilent Technologies, http://www.chem.agilent.com/scripts /generic.asp?lpage= 10692&prodcol=Y.

[2] J. Buhler, T. Ideker, D. Haynor "Dapple: Improved Techniques for Finding Spots on DNA Microarray", *UW CSE Technical Report UWTR 2000-08005*, 2000. http://www.cs.wustl.edu/ jbuhler/research/dapple/.

[3] R. Carmona, W.L. Hwang, B. Torr'esani, "Multiridge Detection and Time-Frequency Reconstruction", *IEEE Trans. on Signal Processing*, vol. 47, no. 2, pp. 480-492, February 1999.

[4] D. Cremers, M. Rousson "Review of Statistical Approaches to Level Set Segmentation: Integrating Color, Texture, Motion, and Shape", *International Journal of Computer Vision*, 2006, To appear.

[5] Y. Chen, E. R. Dougherty, M. L. Bittner, "Ratio-Based Decisions and the Quantitative Analysis of cDNA Microarray Images", *Journal of Biomedical Optics*, 2, 364-374, October 1997.

[6] L. D. Cohen, "On Active Contour Models and Balloons", *CVGIP: Image Understanding*, vol. 53, no. 2, pp. 211-218, March 1991.

[7] M. B. Eisen, P. O. Brown, "DNA Arrays for Analysis of Gene Expression", *Methods Enzymol 303,* 179-205 (1999).

[8] GenePix Pro, http://www.axon.com/gn_GenePixSoftware.html.

[9] K. Fukunaga, "Introduction to statistical pattern recognition", *Academic Press, New York*, 1972.

[10] J. Gollub, C. A. Ball, G. Binkley, K. Demeter, D. B. Finkelstein, J. M. Hebert, T. Hernandez-Boussard, H. Jin, M. Kaplper, JC Matese, M. Schroeder, PO Brown, D. Botstein, and G. Sherlock, "The Stanford Microarray Database: data access and quality assessment tools", *Nucleic Acids Res.*, 31(1):94-96, January 1, 2003.

[11] J. Ho, W. L. Hwang, H. H. S. Lu, D. T. Lee, "Gridding Spot Centers of Smoothly Distored Microarray Images", *IEEE Trans. on Image Processing*, vol. 15, no. 2, pp. 342-353, February 2006.

[12] M. Kass, A. Witkin, D. Terzopoulos, "Snakes: Active Countour Models", *International Journal of Computer Vision*, 1988, pp.321-331.

[13] M. Katzer, F. Kummert, G. Sagerer, "Methods for Automatic Microarray Image Segmentation", *IEEE Transactions on Nano-Bioscience*, 2(4):202-214, 2003.

[14] R. Nagarajan, "Intensity Based Segmentation of Microarray Images", *IEEE Trans. Med. Imaging*, 22(7), pp. 882-889, 2003.

[15] G.K. Smyth, Y.H. Yang, T. Speed, "Statistical Issues in cDNA Microarray Data Analysis", *Methods Mol Biol.*, 2003;224:111-36.

[16] M. Schena, D. Shalon, R. W. Davis, P. O. Brown, "Quantitative Motoring of Gene Expression Patterns with a Complementary DNA Microarray", *Science,* Oct 20;270(5235):467-70, 1995.

[17] Y. H. Yang, M. J. Buckley, S. Dudoit, T. P. Speed, "Comparison of Methods for Image Analysis on cDNA Microarray Data", *Journal of Comuptational and Grahical Statistics,* 11:108-136, 2002.