GRIDDING THE SPOT CENTERS OF MICROARRAY IMAGES

Jinn Ho[†*], *Wen-Liang Hwang*[†], *Henry Horn-Shing Lu*[‡] , *and D.T. Lee*[†]

Institute of Information Science, Academia Sinica, Taiwan [†]
The Genomics Research Center, Academia Sinica, Taiwan [*]
Institute of Statistics, National Chiao Tung University, Taiwan [‡]

ABSTRACT

We use an optimization technique to accurately locate a distorted grid structure in a microarray image. By assuming that spot centers deviate smoothly from a checkerboard grid structure, we show that the process of gridding spot centers can be formulated as a constrained optimization problem. The constraint is equal to the variations of the transform parameter. We demonstrate the accuracy of our algorithm on two sets of microarray images. One set consists of some images from the Stanford Microarray Database; we compare our centers with those annotated in the Database. The other set consists of oligonucleotide images. We compare our results with those obtained by GenePix Pro 5.0. Our experiments were performed completely automatically.

1. INTRODUCTION

An important first step in gene expression analysis is detecting the position of a spot center, and labeling its corresponding coordinate in an micro-arrayer [2, 8]. This is called the spot gridding problem [15, 6]. Even though micro-arrayers arrange spots on a relatively regular checkerboard grid, spotting error irregularities that occur during the array manufacturing process makes accurate gridding of spot centers difficult. Deviations from microarray regularities are attributable to different causes, such as center-to-center spacing deviations of an arrayer, varied surface properties of the substrate, and imprecise movement of manufacturing devices [11]. Spots can also vary in size and position due to noise in the sample preparation and hybridization processes [13]. Dealing with spot center variations is the principal source of complexity in solving the gridding problem.

Some examples of image analysis software for spot gridding found in the Stanford Microarray Database (SMD) are ScanAlyze [10], GenePix [4], and Koadarray [5]. These require parameters and, at times, manual intervention to locate exact spot centers. We provide one constraint, which assumes that spot centers deviate smoothly. Deviations are modeled as a sequence of similarity transformations whose parameters vary smoothly. With this constraint, we can formulate the spot center gridding problem as a constrained optimization problem by combining a quantitative criterion that measures the correctness of the gridding result with a constraint that reduces local parameter variation. The problem can be solved numerically by an iterative algorithm.

We begin with block boundary detection to extract the block layout. A Bayesian approach, based on a multi-threshold Markov model [1], is combined with a model-based recognition method and a refinement algorithm to sequentially refine the spot centers from a sequence of thresholds. After obtaining the initial transform parameters, the Lagrange multiplier λ controls the balance between the correctness of the estimated transform parameters and the smoothness of the transforms in the final solution.

2. GRIDDING PROBLEM

Many approaches that solve the gridding problem require parameters as well as human intervention. The usual parameters are: the block layout structure, the width and height of each block, and the distances of spot centers of adjacent rows and columns. Even if these parameters are known, human intervention is still needed to adjust overly-deviated spot centers.

A reasonable way to do this is to assume that spot center deviations can be modeled as smoothly varying transformations, which can be characterized by a few local parameters. Let $\mathbf{x}$ and $\mathbf{y}$ be the coordinates of a pair of matched centers in the model and the image, respectively. We use $\mathscr{T}_{\mathbf{x},\mathbf{y}}$ to represent the distortion between $\mathbf{x}$ and $\mathbf{y}$, and assume that the distortion can be approximated by a similarity transform. Thus, $\mathbf{x}$ and $\mathbf{y}$ are related by the matrix form

$$\mathbf{y} = \mathscr{T}_{\mathbf{x},\mathbf{y}}(\mathbf{x}) \approx A(\mathbf{x},\mathbf{y})\mathbf{x} + \mathbf{b}(\mathbf{x},\mathbf{y}),$$

where

$$A(\mathbf{x},\mathbf{y}) = \begin{bmatrix} a(\mathbf{x},\mathbf{y}) & -b(\mathbf{x},\mathbf{y}) \\ b(\mathbf{x},\mathbf{y}) & a(\mathbf{x},\mathbf{y}) \end{bmatrix} \tag{1}$$

and

$$\mathbf{b}(\mathbf{x},\mathbf{y}) = \begin{bmatrix} c(\mathbf{x},\mathbf{y}) \\ d(\mathbf{x},\mathbf{y}) \end{bmatrix} \tag{2}$$

is a translation matrix. We denote the collections of centers in the model and the image as $\{\mathbf{x}\}$ and $\{\mathbf{y}\}$ respectively. We use $[\mathbf{x},\mathbf{y}]$ to denote that $\mathbf{x}$ and $\mathbf{y}$ are a pair of matched centers. The mean squared error of a collection of matched centers is

$$e_e(\{[\mathbf{x},\mathbf{y}]\}) = \frac{1}{|\{[\mathbf{x},\mathbf{y}]\}|} \sum_{\{[\mathbf{x},\mathbf{y}]\}} ||A(\mathbf{x},\mathbf{y})\mathbf{x} + \mathbf{b}(\mathbf{x},\mathbf{y}) - \mathbf{y}||^2, \tag{3}$$

where $|\{[\mathbf{x},\mathbf{y}]\}|$ is the size of the matched pair. We can impose a smoothness constraint by minimizing the variations of $A(\mathbf{x},\mathbf{y})$ and $\mathbf{b}(\mathbf{x},\mathbf{y})$. If $\mathbf{x}_i$ and $\mathbf{x}_j$ are neighboring grids, then according to a smoothness constraint, the parameters of $A(\mathbf{x}_i,\mathbf{y}_i)$, $\mathbf{b}(\mathbf{x}_i,\mathbf{y}_i)$ and $A(\mathbf{x}_j,\mathbf{y}_j)$, $\mathbf{b}(\mathbf{x}_j,\mathbf{y}_j)$ should have similar values. A simple measurement of the smoothness of the parameters is

$$e_s(\{[\mathbf{x},\mathbf{y}]\}) = \frac{1}{2|\mathscr{A}\{\mathbf{x}\}|} \sum_{\mathbf{x}_i \in \mathscr{A}\{\mathbf{x}\}} \frac{1}{|\mathscr{N}\{\mathbf{x}_i\}|} \sum_{\{[\mathbf{x},\mathbf{y}]\}} V_A(i,j) + V_b(i,j), \tag{4}$$

where $V_A(i,j) = ||A(\mathbf{x}_i,\mathbf{y}_i) - A(\mathbf{x}_j,\mathbf{y}_j)||_F^2$, $V_b(i,j) = ||\mathbf{b}(\mathbf{x}_i,\mathbf{y}_i) - \mathbf{b}(\mathbf{x}_j,\mathbf{y}_j)||^2$ and $||.||_F$ is the Frobenius norm defined as $\sqrt{\sum_i \sum_j |b_{i,j}|^2}$. According to Equations 3 and 4, we need to find the set of matched pairs between $\{\mathbf{x}\}$ and $\{\mathbf{y}\}$ that minimizes $e_e + \lambda e_s$, where λ is a parameter that weights the error in the matched pair relative to the departure from smoothness of the transform parameters.

2.1 Numerical Solution

To find a numerical solution of $e_e + \lambda e_s$, we use a finite-element method, because it is easy to implement and achieves a satisfactory solution. Let $\mathbf{x}$ and $\mathbf{y}$ be paired, with $\mathbf{x}$ in the active set $\mathscr{A}\{\mathbf{x}\}$. If the coordinate of $\mathbf{x}$ is $[x_1(k,l) \ \ x_2(k,l)]^T$ and the coordinate of $\mathbf{y}$ is $[y_1(k,l) \ \ y_2(k,l)]^T$, then - to simplify the formulation of a numerical method - we denote $\mathbf{x}$ as $\mathbf{x}_{k,l}$, and $\mathbf{y}$ as $\mathbf{y}_{k,l}$; the parameters in $A(\mathbf{x}_{k,l}, \mathbf{y}_{k,l})$ as $a_{k,l}$ and $b_{k,l}$; and the parameters in $\mathbf{b}(\mathbf{x}_{k,l}, \mathbf{y}_{k,l})$ as $c_{k,l}$ and $d_{k,l}$. The mean squared error measurement in Equation 3 is, therefore:

$$e_e = \frac{1}{|\mathscr{A}\{\mathbf{x}\}|} \sum_{\mathbf{x}_{k,l} \in \mathscr{A}\{\mathbf{x}\}} \{E_e(k,l)\},$$

where $E_e(k,l) = (a_{k,l}x_1(k,l) - b_{k,l}x_2(k,l) + c_{k,l} - y_1(k,l))^2 + (b_{k,l}x_1(k,l) + a_{k,l}x_2(k,l) + d_{k,l} - y_2(k,l))^2$, and $|\mathscr{A}\{\mathbf{x}\}|$ is the size of the active set.

We also use $\mathscr{N}(\mathbf{x}_{k,l})$ to denote the set of neighbors of $\mathbf{x}_{k,l}$ in the active set.
Equation 4 then becomes

$$e_s = \frac{1}{2|\mathscr{A}\{\mathbf{x}\}|} \sum_{\mathbf{x}_{k,l} \in \mathscr{A}\{\mathbf{x}\}} \frac{1}{|\mathscr{N}\{\mathbf{x}_{k,l}\}|} \sum_{\mathbf{x}_{k+i,l+j} \in \mathscr{N}\{\mathbf{x}_{k,l}\}} \{E_s(k,l)\}, \tag{5}$$

where $E_s(k,l) = [2(a_{k,l} - a_{k+i,l+j})^2 + 2(b_{k,l} - b_{k+i,l+j})^2 + (c_{k,l} - c_{k+i,l+j})^2 + (d_{k,l} - d_{k+i,l+j})^2]$ and $i,j \in \{-1,1\}$. We need $\{a_{k,l}\}$, $\{b_{k,l}\}$, $\{c_{k,l}\}$, and $\{d_{k,l}\}$ to minimize

$$e = e_e + \lambda e_s. \tag{6}$$

To solve this, we differentiate e with respect to $a_{k,l}$, $b_{k,l}$, $c_{k,l}$, and $d_{k,l}$ and set the derivatives to zero. The resultant equations are formed as a matrix representation and can be solved by the Jacobi iterative scheme.

The final quality of the Jacobi iterative solution depends on the quality of the initial solution. In the following section, we propose a method that finds a robust initial solution.

3. FINDING A ROBUST INITIAL SOLUTION

We use a sequence of robust image processing methods for a more automated process of finding effective initial transform parameters.

3.1 Boundary Detection and Block Extraction

In a microarray, spots are grouped into blocks, and we must delineate each block in order to identify the spot centers within the blocks. In [3], the blocks of a microarray are delineated from the vertical and horizontal projection profiles of the image. However, the method only works well provided that the microarray image has no rotation deviation.

Boundary Detection

To extract the boundaries of a slightly rotated image, we use the line equation $y = x \cdot A_H + B_H$ to represent the top or bottom boundary.

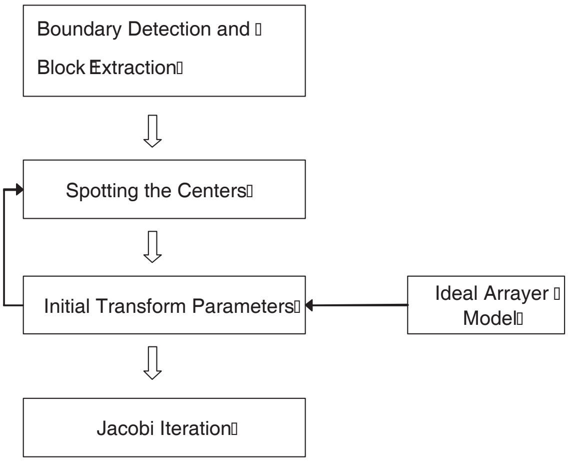

Figure 1: Our model-based approach algorithm.

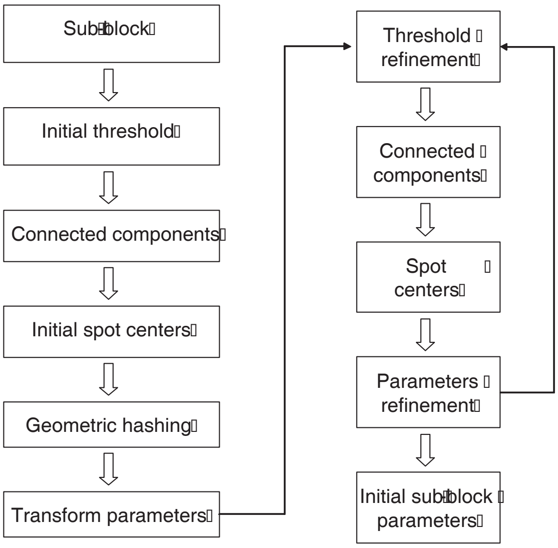

Figure 2: Sub-block transform parameters acquisition.

In order to find the boundary line L that is tangent to the edges of spot centers located at a boundary, we use the Gaussian-like weighting function

$$W(d_L) = W(i,j) \cdot exp(-{d_L}^2)$$

in which $W(i,j) = |\partial I(i,j)/\partial j|$ and $d_L = j - (i \cdot A_H + B_H)$.

The weighting function gives more weight to a pixel that is closer to the line L, or one that has a greater absolute intensity gradient. To find the line, we look for the A_H and B_H that minimize the weighted squared error $Err_H(A_H, B_H) = \sum_{i,j} W(d_L) \cdot ({d_L}^2)$.

By differentiating Err_H with respect to A_H and B_H and setting the results to zero, we have

$$\begin{aligned} \sum_{i,j} jW(d_L) \cdot d_L\{d_L^2 + 1\} &= 0, \text{ and} \\ \sum_{i,j} W(d_L) \cdot d_L\{d_L^2 + 1\} &= 0. \end{aligned}$$

If a point is close to line L, then we have $d_L \to 0$. This means that $(d_L^2 + 1)$ can be approximated as $\exp({d_L}^2)$. If we

denote $\sum_L$ to be the summation of all points near L, we have the approximations of the two equations above, which are $\sum_L jW(i,j) \cdot d_L = 0$, and $\sum_L W(i,j) \cdot d_L = 0$.
The solutions of the above equations are

$$\begin{aligned} A_H &= (1/d) \cdot \{\sum_L jW(i,j) \sum_L iW(i,j) \\ &- \sum_L W(i,j) \cdot \sum_L j \cdot i \cdot W(i,j)\}, \text{ and} \\ B_H &= (1/d) \cdot \{\sum_L jW(i,j) \sum_L j \cdot i \cdot W(i,j) \\ &- \sum_L j^2 W(i,j) \cdot \sum_L iW(i,j)\}, \end{aligned}$$

where $d = (\sum_L jW(i,j))^2 - \sum_L W(i,j) \cdot j^2 \sum_L W(i,j)$.

Block Extraction

After extracting four boundary lines, we slightly modify them such that they form a rectangular box. To extract blocks, we project along each boundary line and select the blocks from the projection profile as described in [3].

3.2 Initial Distortion Estimation

After the blocks are delineated, we estimate an initial distortion of the spot centers in each block. A good initial estimation of block distortion is obtained by dividing a block into sub-blocks, and assuming that the transform within a sub-block is the same anywhere in that sub-block. Thus, each sub-block has a transform. Dividing a block into several sub-blocks and assuming each has the same transform allows us to efficiently apply the geometric hashing algorithm and obtain transform parameters.

There is a trade-off between the solutions of geometric hashing and computation time cost, *i.e.*, a sub-block with more spots obtains a better result at a higher computation time cost.

Geometric Hashing to Find Matched Pairs in a Sub-block

We assume that local distortion within a sub-block can be approximated by a similarity transform. That is, model point $\mathbf{x}$ and sub-block point $\mathbf{y}$ are related by the matrix transformation

$$\mathbf{y} \approx \begin{bmatrix} a & -b \\ b & a \end{bmatrix} \mathbf{x} + \begin{bmatrix} c \\ d \end{bmatrix}. \quad (7)$$

Geometric hashing can find the parameters of this similarity transform between the model points and sub-block points, according to an invariant property. We define a frame from a pair of model points and assign the coordinate $[0\ \ 0]^t$ to one point and $[1\ \ 0]^t$ to the other. These two points are called a basis pair. The coordinates of all other points with respect to the same basis will be preserved after applying any similarity transform to the points.

In this way, if model points and sub-block points are related by a similarity transform, we can derive the parameters of the similarity transform from the matched basis pair in the model and sub-block. A detailed discussion of geometric hashing can be found in [7, 9, 14].

Gridding Centers Using a Multi-threshold Markov Model

Because there are various signal intensities and noise levels in a microarray, using a threshold to distinguish signals from noise may yield either a spot pattern with insufficient signal information, or one with too much noise information. Thus, we use a Bayesian approach on a Markov random field model to locate spots. We begin with a coarse threshold to binarize a sub-block image. We then compute connected components of the resultant image and find the component centers. The transform between the model and sub-block centers is obtained by geometric hashing. We refine the threshold and repeat the above procedure, replacing geometric hashing with a Bayesian approach that uses a Markov model to refine parameters.

Tree-based Outlier Correction

Because local distortion varies smoothly, the transform parameters of neighboring sub-blocks should have similar values. An error in a previous parameter estimation can thus be adjusted, based on the estimated parameters of neighboring sub-blocks. If the transform of a sub-block is inconsistent with its neighboring sub-blocks, we say that the sub-block is an outlier. A simple way to define an outlier is to let the rotation θ and scale s be the similarity transform of the sub-block within a given threshold.

The quadtree structure allows us to correct any number of outlier sub-blocks. If the children of a node p are not all outliers, parameters of the inlier children nodes are used to calculate the parameters of the parent node p . Resultant parameters are passed to all outlier children of p and become the new parameters of the outlier nodes. If all children of p are outliers, then p is an outlier. We can use the parameters obtained from the inlier sibling of p as the new parameters of p.

4. PERFORMANCE EVALUATION

We evaluate our spot gridding algorithm by comparing our results with those obtained by applying other algorithms to two sets of microarray images. One set contains some poor quality images from SMD, while the other contains Agilent 60-mer oligonucleotide microarrays whose specifications are on the related web pages [12].

The Agilent microarrays are some of the best quality oligonucleotide chips currently available commercially. We use these sets of images to demonstrate that our method can accurately grid the spot centers of images of varying quality produced by different technologies. We implement our algorithm using the Windows XP platform and all images are processed in the Matlab environment. The gray level image in SMD takes eight bits, while that of Agilent's image takes sixteen bits. Throughout our experiments, we use gray-scale images and set the control parameter λ to $\frac{1}{16}$. Our experiments show that this setting is robust for different microarray images.

Figures 4 shows the super-position of our detected spot centers over the block provided by SMD. All spot centers detected by our method are located within their corresponding boxes. The mean and standard deviation of the Distance of our centers and the SMD centers (centroids of boxes) are respectively 1.7 and 1.1 pixels(Figures 5). In addition, our method can extract accurate spot centers from a small rotated microarray image.

5. CONCLUSIONS

A large proportion of grid distortion can be approximated by a locally smooth distortion. In this manuscript, we pro-

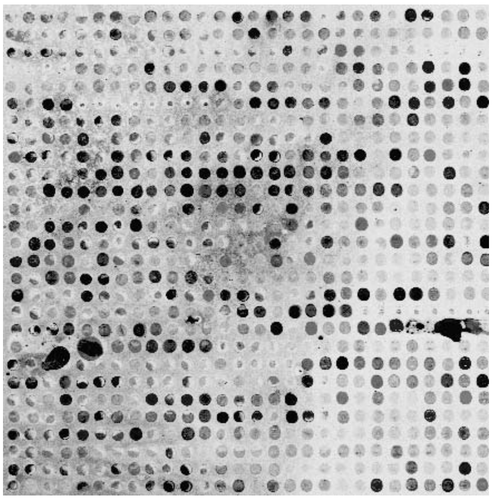

Figure 3: Block (1,1) of $lc30n010$ provided by a SMD microarray image.

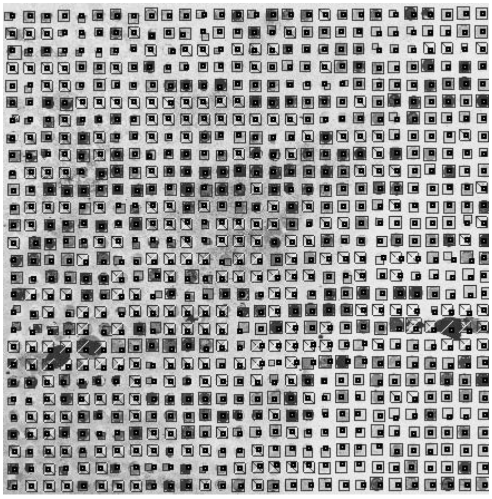

Figure 4: The super-position of our detected spot centers on the block in Figure 3.

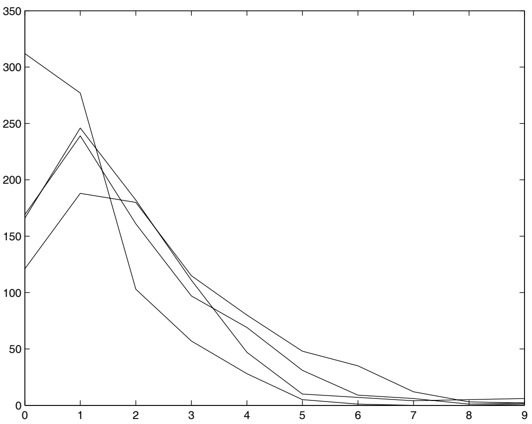

Figure 5: Distance histogram of our spot centers and block (1,1), (1,2), (2,1), (2,2) of $lc30n010$ provided by SMD.

pose an optimization approach to grid the exact spot centers of a microarray image, whose grids are slowly varying similarity transforms. A Bayesian approach and a multi-threshold Markov model are used to find robust initial parameters. The initial parameters are refined by Jacobi iterations, which solves our optimization problem. Experiments show that our method can robustly extract accurate spot centers from microarrays with local smooth grid distortions. In practice, however grid distortions can be discontinuous. Improving our method for images of discontinuous distortion grids is an issue worth further study.

REFERENCES

[1] C. A. Bouman and M. Shapiro, “A Multiscale Random Field Model for Bayesian Image Segmentation”, *IEEE Trans. on Image Processing,* vol. 3 , no. 2, pp. 162-177, March 1994.

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

[3] R. Fabbri, L. da F. Costa, J, Barrera, “Towards Non-Parametric Gridding of Microarray Images”, *14th International Conference on Digital Signal Processing,* vol.2, 1-3 July 2002, pp.623-626.

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

[5] Koadarray, http://www.koada.com/koadarray.

[6] C. Kooperberg, T. G. Fazzio, J. J. Delrow, and t. Tsukiyama, “Improved Background Correction for Spotted DNA Microarrays”, *Journal of Computational Biology,* vol.9, no. 1, pp. 55-66, 2002.

[7] Y. Lamdan, J. T. Schwartz, and H. J. Wolfson, “Affine Invariant Model-Based Object Recognition”, *IEEE Transactions on Robotics and Automation*, vol.6, No.5, pp.578-589, October 1990.

[8] D. J. Lockhart and E. A. Winzeler “Genomics, Gene Expression and DNA Arrays”, *Nature,* vol. 405, pp. 827-836, June 2000.

[9] I. Rigoutsos and R. Hummel, “A Bayesian Approach to Model Matching with Geometric Hashing”, *Computer Vision and Image Understanding*, vol.61, No.7, pp.11-26, July 1995.

[10] ScanAlyze, http://rana.lbl.gov/EisenSoftware.htm

[11] M. Schena, “Microarray Analysis”, *Wiley-Liss,* 2003.

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

[13] Y. Tu, G. Stolovitzky, and U. Klein, “Quantitative Noise Analysis for Gene Expression Microarray Experiments”, *PNAS,* vol.99, no. 22, Oct. 29, 2002, pp. 14031-14036.

[14] H. J. Wolfson, I. Rigoutsos, “Geometric Hashing: an Overview”, *IEEE Computational Science and Engineering*, vol.4, Issue.4, pp.10-21, Oct-Dec 1997.

[15] Y. H. Yang, M. J. Buckley, S. Dudoit, and T. P. Speed, “Comparison of Methods for Image Analysis on cDNA Microarray Data”, *Journal of Computational and Graphical Statistics,* vol.11, pp.108–136, 2002.