Automatic DNA Microarray Gridding based on Support Vector Machines

Dimitris Bariamis, Dimitris Maroulis, *Member, IEEE* and Dimitris K. Iakovidis, *Member, IEEE*

***Abstract*—This paper presents a novel method for DNA microarray gridding based on Support Vector Machine (SVM) classifiers. It employs a set of soft-margin SVMs to estimate the lines of the DNA microarray grid by maximizing the margin between the lines and the spots. This process comprises an efficient and effective approach of separating the spots into distinct rows and columns. The classifiers are trained using the spot locations as training vectors. The results obtained from the application of the proposed method on reference microarray images illustrate its robustness in the presence of artifacts, noise and weakly expressed spots. The comparative evaluation presented reveals its advantageous performance over a state of the art gridding approach. The gridding quality achieved exceeds 95% in terms of the total number of perfectly gridded spots.**

I. Introduction

DNA microarrays are devices that enable monitoring of the expression levels for thousands of genes in each experiment, rendering them a valuable tool of biotechnology. An experiment involves the isolation of two mRNA samples to be compared. The two samples are labeled with distinct fluorescent dyes, commonly Cy5 and Cy3, hybridized with the known genes that are printed on the microarray and scanned at the wavelength of each dye. The output of an experiment is a high resolution digital image for each wavelength. A microarray image consists of a matrix of blocks, each of which contains a number of rows and columns of spots. The intensity of each spot signifies the degree of hybridization of the sample to a known gene, thereby indicating the expression level of the particular gene.

The processing of microarray images is usually performed in three steps, namely gridding, segmentation and intensity extraction. Gridding involves assigning coordinates to each spot, whereas segmentation handles the separation of the spot pixels (foreground) from the background. In the last step, the intensity of the foreground and background is extracted from the respective pixels. Since gridding is the first step in the microarray image processing, its results significantly affect the accuracy of the following steps and the extracted spot and background intensities.

A gridding algorithm should be able to address several issues that arise during the processing of microarray images, such as rotation, irregular spot sizes and shapes, spots of very low or zero intensity, as well as noise and various artifacts that are introduced by the wet lab process. Furthermore, the algorithm should not require user intervention or parameter fine-tuning, in order to facilitate high-throughput processing of large amounts of data and avoid the dependence of the results on the user input.

Several methods have been proposed for microarray gridding, but most rely on some user input or adjustments, such as those implemented in ScanAlyze [1] and ImaGene [2]. Only a few state of the art methods address the problem of automatic gridding. Such methods are based on mathematical morphology [3], Markov random fields [4], Voronoi diagrams [5], Bayesian grid matching [6], genetic algorithms [7] or a combination of approaches [8]. However, there are still problems that have to be resolved before fully automatic gridding can take place. For example, the method proposed in [3] requires that grid rows and columns are strictly aligned with the x and y axes; the region segmentation approach proposed in [4] fails to detect many weak signal spots; in [8], the number of rows and columns of spots per grid is required; the method proposed in [6] is quite complex; and the genetic approach [7] is very time-consuming.

In this paper we propose the use of a soft-margin linear Support Vector Machine (SVM) classifier [9] for DNA microarray gridding that overcomes the aforementioned issues. A spot detection step selects spots that have specific properties, filtering out any irregularities and artifacts. The remaining spots are then automatically separated into rows and columns by estimating the distance between consecutive rows and columns of spots, as well as the image rotation angle. The SVM classifier automatically sets the separating lines between consecutive rows or columns so as to maximize the margin between the lines and the spots. The motivation for the using the linear SVM classifier in a gridding application was its well known geometric properties as a maximum-margin classifier, as well as its generalization ability and tolerance to outliers. These features provide robustness in the presence of weakly

Manuscript received July 5, 2008. This work was realized under the framework of the Reinforcement Program of Human Research Manpower ("PENED 2003" – 03ED324), co-funded 25% by the General Secretariat for Research and Technology, Greece, and 75% by the European Social Fund.

D. Bariamis is with the Dept. of Informatics and Telecommunications, University of Athens, Panepistimiopolis, 15784 Athens, Greece. (e-mail: d.bariamis@di.uoa.gr).

D. Maroulis is with the Dept. of Informatics and Telecommunications, University of Athens, Panepistimiopolis, 15784 Athens, Greece. (corresponding author, phone: +302107275317, e-mail: d.maroulis@di.uoa.gr).

D. K. Iakovidis is with the Dept. of Informatics and Telecommunications, University of Athens, Panepistimiopolis, 15784 Athens, Greece. (e-mail: dimitris.iakovidis@ieee.org).

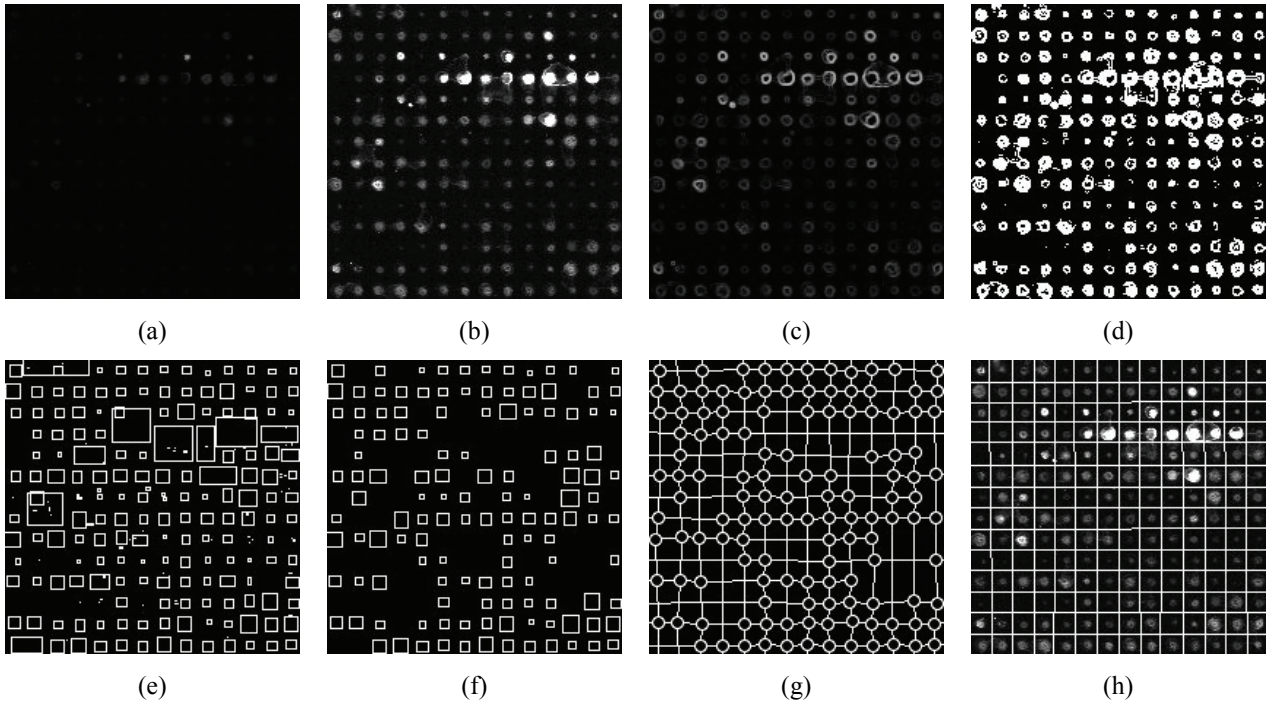

Fig. 1. Gridding algorithm steps: (a) original image, (b) normalization, (c) edge-detection, (d) thresholding, (e) spot detection, (f) valid spots, (g) rows and columns of spots, (h) SVM-based gridding

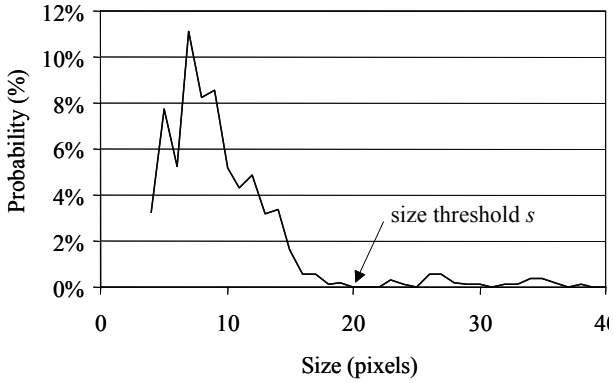

Fig 2. Histogram of rectangle sizes

expressed spots and in the presence of irregularities or artifacts.

It is worth noting that the approach proposed in [5] is equivalent to using an 1NN (nearest neighbor) classifier, which requires thorough filtering of outliers and the introduction of artificial spots in place of the spots that are very weakly expressed. In contrast to this approach, the use of SVM in the proposed method allows a significant tolerance to outliers and enables gridding without requiring interpolation of non-expressed spots.

The rest of the paper is organized in three sections. Section II describes the proposed gridding methodology. The results of the experiments conducted are presented in Section III, and the conclusions of this study are summarized in Section IV.

II. Methodology

A number of preprocessing steps are initially applied to discover the locations of the spots, as well as the distance between consecutive rows and columns of spots in a DNA microarray. Once extracted, that information is used to train a set of linear SVM classifiers, which produces the lines that form the microarray image grid. Each SVM classifier is trained with the spot locations as training vectors and produces a grid line. In short, the proposed methodology consists of four steps:

1. Image preprocessing
2. Spot detection
3. Distance estimation between consecutive rows and columns
4. SVM-based gridding

A. Image preprocessing

The first step involves normalization of the microarray image (Fig. 1a), in order to enhance its dynamic range (Fig. 1b). The edges of the spots are detected by the application of the Sobel operator on the image (Fig. 1c). A threshold t is used to isolate the sharpest edges, which correspond to prevalent spots (Fig. 1d). If the value of t is lower than optimal it might lead to several pixel groups per spot in cases where the spot edges are not sharp, otherwise a higher value might lead to merging of several spots into one pixel group in case the distance between them is too small.

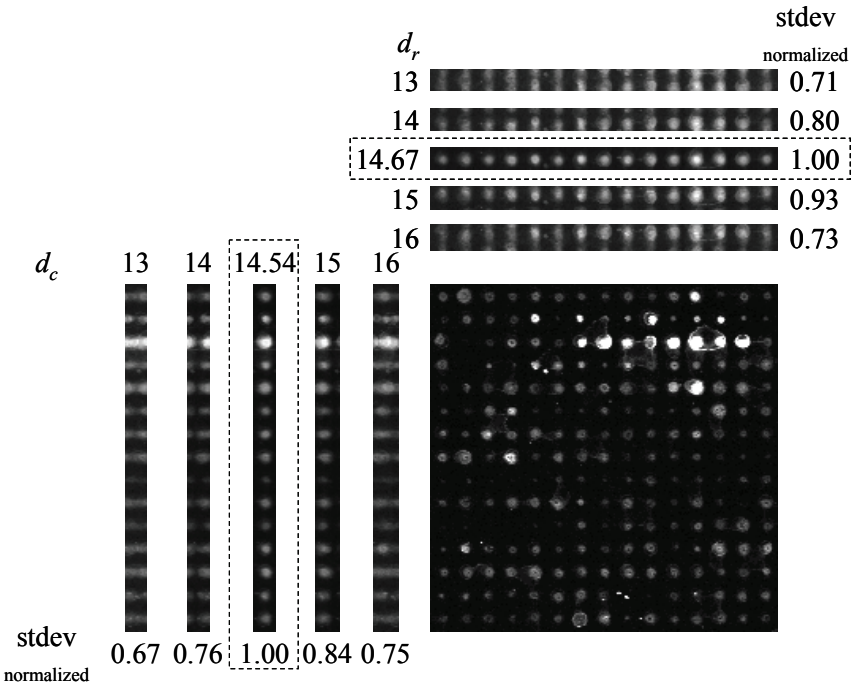

Fig 3. Calculation of the distance between rows d_r and columns d_c in a subgrid.

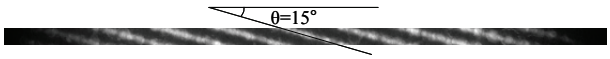

Fig. 4. Detection of subgrid rotation angle

B. Spot detection

The thresholded image is analyzed, in order to locate pixel groups, each of which contains pixels that reside on a single spot edge. Each group is represented as a rectangle that circumscribes the pixels of the group, as illustrated in Fig. 1e. Ideally, each rectangle should contain a single microarray spot, however in some occasions it might also include artifacts or multiple merged spots. Subsequently, only the rectangles that have specific size and shape characteristics are considered valid, as shown in Fig. 1f.

In order to assess the validity of each pixel group, the histogram of the sizes of the circumscribed rectangles is created, as illustrated in Fig. 2. A size threshold s is derived from the histogram at its leftmost zero bin, thus rejecting any pixel groups that are larger. The threshold for the case shown in Fig. 2 is s=20 pixels. Furthermore, the rectangles should be quasi-square in order to contain only one microarray spot, therefore the ratio r of the smaller to the larger side of each rectangle must be close to unity.

C. Distance estimation between consecutive rows and columns

Given a microarray image of $x \times y$ dimensions and an estimate of the distance d_r between the rows of its spots, the image is segmented into subimages of size $x \times d_r$ pixels. The subimages are then accumulated into a single $x \times d_r$ image. Such images for several values of d_r are illustrated in Fig. 3. The distance d_r is represented by a floating point value, resulting in a tiling that only partially includes the pixels that reside at the edges of each subimage. The gray level values of those pixels are then linearly interpolated. The value of d_r for which the standard deviation of the gray levels in the resulting image is maximized, is the estimated distance between the rows. Figure 3 presents the normalized standard deviations that correspond to each value of d_r. This process is repeated to calculate the distance d_c between columns.

By analyzing the resulting $x \times d_r$ image, it is possible to calculate the angle of rotation of the original microarray image, in order to automatically get aligned. Figure 4 depicts a $x \times d_r$ image produced from a microarray image that has been rotated by 15 degrees. It is evident that the angle of rotation can be easily extracted.

Having the distances d_r and d_c, the spots detected in the previous step can be divided into rows and columns as shown in Fig. 1g.

D. SVM-based gridding

In this step, the spots that belong to each pair of consecutive rows k and k+1 of the grid are isolated. The coordinates of the center pixel of each spot form a vector. If the spot belongs to row k, it is assigned to class +1, else it is assigned to class -1. Therefore, for each pair of rows, a set D (Eq. 1) of vectors x_i and their respective classes c_i is created and it is provided as a training set to a linear SVM classifier.

$$D = \left\{ (\overline{x_i}, c_i) \mid \overline{x_i} \in \Re^2, c_i \in \{-1, +1\} \right\} \quad (1)$$

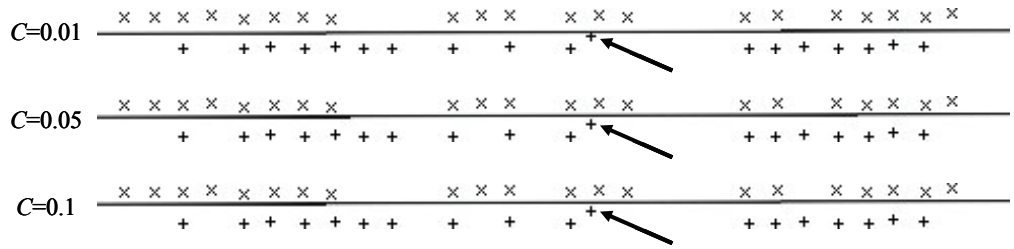

Fig 5. Comparison of separating lines as a function of the SVM cost parameter *C*. An outlier vector is denoted by the arrow.

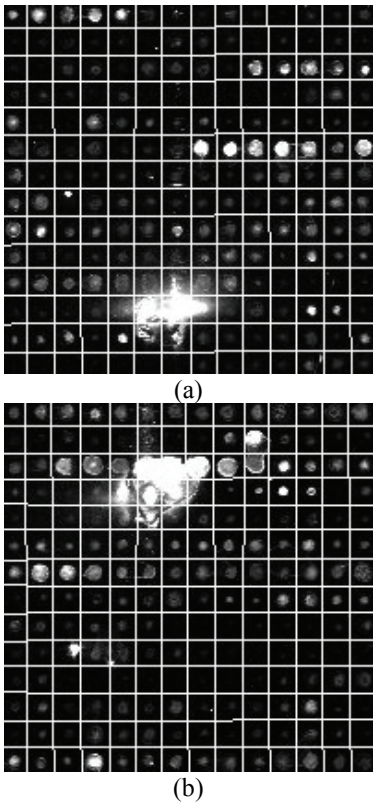

Fig. 6. Gridding results in images with artifacts

TABLE I
GRIDDING PERFORMANCE COMPARISON

	Perfect	Marginal	Incorrect
Proposed method	95.1%	4.5%	0.4%
Zacharia et al. [7]	94.6%	4.8%	0.6%

The classifier produces the separating line

$$\overline{w} \cdot \overline{x} - b = 0 \tag{2}$$

that maximizes the margin between the vectors x_i which represent the spots belonging to the two rows. Considering the fact that the spots reside on distinct rows, the set of training vectors is linearly separable and can be successfully classified using a hard-margin SVM classifier. Nevertheless, we have chosen a soft-margin classifier to diminish the effects of misdetected spots that act as outliers in the training set. Thus, the margin is maximized when

$$\min \frac{1}{2}\|w\|^2 + C\sum_i \xi_i \tag{3}$$

is minimized under the constraints

$$c_i(\overline{w} \cdot \overline{x_i} - b) \geq 1 - \xi_i \tag{4}$$

The cost parameter *C* (Eq. 3) determines the effect that outliers or noise might have on the separating lines produced by the SVM. Large values of *C* result in separating lines that are mostly determined by any outliers or noise and are not optimal. On the other hand, if a smaller value of *C* is used, the separating lines follow the general trend of the training set given to the classifier, ignoring any outliers.

III. RESULTS

The dataset used for the evaluation of the proposed method consists of the 25 DNA microarray images, from the Stanford Microarray Database (SMD), used in [7]. The images have 1900×5500 pixels and 16-bit gray level depth. Each of these images includes 41472 spots, equally distributed in 48 blocks. They have been produced for the study of the gene expression profiles of 54 specimens of acute lymphoblastic leukemia. The specimens span 37 positive and 17 negative to BCR-ABL [10], which is a fusion gene product resulting from translocation between the 9th and the 22th chromosomes.

The gridding performance of the proposed method was evaluated using *C*=0.01 and *t*=16. The SVM cost parameter *C* determines the effect that outliers or noise might have on the separating lines that the SVM produces. Figure 5 illustrates the separating lines produced using *C*=0.01, *C*=0.05 and *C*=0.1 for a training set that includes an outlier, which is denoted by the arrow. It is evident that in the case of *C*=0.01, the outlier is virtually ignored, whereas in the case of *C*=0.1, it determines the positioning of the separating line, resulting in a line that is significantly closer to most of the vectors of the top row. Therefore, a small value of *C*=0.01 should be selected for successful gridding. The value of *t* was experimentally determined.

Each spot was evaluated as being perfectly gridded when all its pixels reside within its respective grid cell, marginally gridded when more than 80% of its pixels reside within its respective grid cell and incorrectly gridded when less than

80% of the spot pixels reside within its respective grid cell. The evaluation results are shown in Table I. Out of more than a million spots present in the data set, 95.1% spots were perfectly gridded, whereas 4.5% and 0.4% were marginally and incorrectly gridded respectively. These results show that the proposed method achieves higher quality gridding than the state of the art method presented in [7]. Additionally, it is significantly faster than the genetic algorithm approach.

Figure 6 illustrates an indicative gridding example with the proposed method. It can be noticed that the obtained gridding is accurate even in the presence of considerable artifacts in the microarray image.

IV. Conclusion

In this paper, we presented a novel method for automatic microarray gridding, which is based on the soft-margin linear Support Vector Machine classifier as a means of achieving high accuracy and robustness.

A spot detection step prior to the use of the classifier facilitates the artifact removal process. Subsequently, the distance between consecutive rows and columns of spots, as well as the image rotation angle, is estimated and the spots are organized into rows and columns. The SVM produces the separating lines of the grid so as to maximize the margin between the lines and the spots, and displays high tolerance to outliers that result from misdetected spots or artifacts and to weakly expressed spots.

Overall, the proposed method achieves successful gridding of DNA microarray images in the presence of the following conditions:

- Irregular and weakly expressed spots
- Noise and artifacts
- Rotation

The experimental results on reference DNA microarray images showed that the proposed method outperforms the state of the art method presented in [7], providing the potential of achieving perfect gridding for the vast majority of the spots.

References

[1] M. B. Eisen, ScanAlyze Nov. 1999 [Online]. Available: http://rana.lbl.gov/EisenSoftware.htm

[2] Biodiscovery, Inc., ImaGene 2005 [Online]. Available: http://www.biodiscovery.com/imagene.asp

[3] J. Angulo, J. Serra, "Automatic analysis of DNA microarray images using mathematical morphology", *Bioinformatics* 19 (5), pp. 553–562, 2003.

[4] M. Katzer, F. Kummert, G. Sagerer, "A Markov random field model of microarray gridding", *Proceedings of the SAC*, ACM, New York, 2003.

[5] N. Giannakeas, D. I. Fotiadis, Anastasia S. Politou, "An Automated Method for Gridding in Microarray Images", *Proceedings of the 28th IEEE EMBS Annual International Conference*, pp. 5876-5879, New York City, USA, 2006.

[6] K. Hartelius, J.M. Cartstensen, "Bayesian grid matching", *IEEE Transactions on PAMI* 25 (2), pp. 162–173, 2003.

[7] E. Zacharia, D. Maroulis, "An Original Genetic Approach to the Fully Automatic Gridding of Microarray Images", *IEEE Transactions on Medical Imaging*, vol. 27, no. 6, 2008

[8] G. Antoniol, M. Ceccarelli, "Microarray image gridding with stochastic search based approaches", *Image and Vision Computing*, 25 (2), pp. 155-163, 2007.

[9] C. Cortes, V. Vapnik, "Support-vector networks", *Machine Learning*, 20 (3), pp. 273-297, 1995.

[10] D. Juric, N. J. Lacayo, M. C. Ramsey, J. Racevskis, P. H. Wiernik, J. M. Rowe, A. H. Goldstone, P. J. O'Dwyer, E. Paietta and B. I. Sikic, "Differential gene expression patterns and interaction networks in bcr-abl-positive and -negative adult acute lymphoblastic leukemias", *J. Clin. Oncol.*, vol. 25, no. 11, pp. 1341–1349, Apr. 2007.