Computerized Medical Imaging and Graphics 34 (2010) 418–425

Contents lists available at ScienceDirect

Computerized Medical Imaging and Graphics

journal homepage: www.elsevier.com/locate/compmedimag

Unsupervised SVM-based gridding for DNA microarray images

Dimitris Bariamis [a,*], Dimitris Maroulis [a], Dimitris K. Iakovidis [b]

[a] Dept. of Informatics and Telecommunications, University of Athens, Panepistimiopolis, Illisia, 15784 Athens, Greece
[b] Dept. of Informatics and Computer Technology, Technological Educational Institute of Lamia, Greece

ARTICLE INFO

Article history:
Received 31 March 2009
Received in revised form 26 August 2009
Accepted 24 September 2009

Keywords:
cDNA microarray images
Gridding
Spot detection
Rotation estimation
Support vector machines

ABSTRACT

This paper presents a novel method for unsupervised DNA microarray gridding based on support vector machines (SVMs). Each spot is a small region on the microarray surface where chains of known DNA sequences are attached. The goal of microarray gridding is the separation of the spots into distinct cells. The positions of the spots on a DNA microarray image are first detected using image analysis operations and then a set of soft-margin linear SVM classifiers is used to estimate the optimal layout of the grid lines in the image. Each grid line is the separating line produced by one of the SVM classifiers, which maximizes the margin between two consecutive rows or columns of spots. The classifiers are trained using the spot locations as training vectors. The proposed method was evaluated on reference microarray images containing more than two million spots in total. The results illustrate its robustness in the presence of artifacts, noise and weakly expressed spots, as well as image rotation. The comparison to state of the art methods for microarray gridding reveals the superior performance of the proposed method. In 96.4% of the cases, the spots reside completely inside their respective grid cells.

1. Introduction

Complementary DNA (cDNA) microarray devices are a valuable tool of biotechnology, enabling monitoring of the expression levels for thousands of genes in each experiment. The first step of such an experiment is the isolation of two messenger RNA (mRNA) samples to be compared. The two samples are reverse-transcribed into cDNA, amplified using polymerase chain reaction and labeled with distinct fluorescent dyes, commonly Cy5 and Cy3. Subsequently the samples (targets) are hybridized on a microarray, which is a slide that includes a large number of probes, i.e. chains of known DNA sequences, on a solid surface. The hybridized microarray is scanned at the wavelength of each dye and 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. Each spot is an area in the image which represents the level of the hybridization between a single probe and the samples. The intensity of each spot signifies the degree of hybridization of the targets to each probe, which is usually a distinctive part of a gene, thereby indicating the expression level of the respective gene.

The quantification of gene expression levels from microarray images is usually performed in three steps, namely gridding, segmentation and intensity extraction. Gridding involves partitioning the spots into distinct cells in the image, as well as 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 and used to quantify the expression levels of the corresponding genes. Since gridding is the first step in the microarray image processing, its results significantly affect the accuracy of the following steps, as well as the final results. Even though the process of gridding a noiseless image would be quite simple, there are several issues that have to be addressed for real images, such as image 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. A robust algorithm should be unsupervised and able to automatically perform accurate microarray gridding under these circumstances, as any user input or intervention would introduce variation into the results. Moreover, unsupervised gridding allows high-throughput processing of large amounts of data.

Several methods have been proposed for microarray gridding; they either rely on some user input and adjustments or do not achieve a high enough accuracy. Such methods are implemented in ScanAlyze [1], ImaGene [2] or SpotFinder [3] that require several parameters to be set by the user. Only a few state of the art methods address the problem of unsupervised gridding based on methods such as mathematical morphology [4], Markov random fields [5], Voronoi diagrams [6,7], Bayesian grid matching [8], Gaussian mixture model [9], genetic algorithms [10] or a combination of

* Corresponding author. Tel.: +30 2107275317.
E-mail addresses: d.bariamis@di.uoa.gr (D. Bariamis), dmaroulis@di.uoa.gr (D. Maroulis), dimitris.iakovidis@ieee.org (D.K. Iakovidis).

0895-6111/$ – see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compmedimag.2009.09.005

approaches [11]. However, there are still drawbacks that have to be resolved before fully automatic gridding can take place. For example, the method proposed in [4] requires that grid rows and columns are strictly aligned with the x- and y-axes, the region segmentation approach proposed in [5] fails to detect many weak signal spots and in [11] the number of rows and columns of spots per grid is required. The method presented in [8] employs an iterative algorithm to solve a complex deformable model for microarray gridding, but simple linear models such as [10] have been shown to achieve high accuracy. The approach proposed in [6,7] requires the introduction of artificial spots in place of the spots that are very weakly expressed. It is worth noting that the use of Voronoi diagrams is equivalent to the use of an 1 NN (nearest neighbor) classifier. The method proposed in [9] is quite accurate, but the evaluation is performed visually on a small number of spots, without comparison to a ground truth reference. Genetic algorithms [10] have the potential to achieve high accuracy, but are very time-consuming as they have to evaluate a large number of possible solutions in order to converge. Our preliminary version [12] of the proposed method is not entirely unsupervised, as the gridding accuracy depends on the successful selection of a few parameters that have to be experimentally determined.

In this paper we propose the use of soft-margin linear support vector machine (SVM) classifiers [13] for DNA microarray gridding that overcomes the aforementioned issues. Several improvements in various steps of the methodology lead to a more robust solution, where the optimal operating parameter values are determined automatically. Extensive experiments were performed, which lead to the conclusion that any changes to the operating parameters induce negligible variations in the accuracy of the results. The more efficient spot detection and filtering, as well as the use of additional data in the SVM training process, contribute to the increased accuracy and robustness of the proposed method. The results of the proposed method are supported by a thorough exploration of the parameter space, the use of an extensive data set and the comparison of the gridding results to the ground truth gridding of the reference images. Prior to the use of the SVM classifiers, the distance between rows and columns of spots is estimated, as presented in Section 2.1, and then a spot detection step selects spots that have specific properties, filtering out any irregularities and artifacts. The remaining spots are then separated into rows and columns and the SVM classifiers set the separating lines between consecutive rows or columns so as to maximize the margin between the spots, without any user intervention. The motivation for the using the linear SVM classifier in a gridding application was its well-known geometric properties as a maximum-margin classifier [14], as well as its tolerance to outliers, in the case of the soft-margin support vector machines. These features provide robustness in the presence of weakly expressed spots and in the presence of irregularities or artifacts.

2. Methodology

In the proposed methodology, the distance between consecutive rows and columns of spots is first estimated and then the locations of the spots are discovered. Once extracted, that information is used to separate the detected spots into rows and columns, which are used as training data for a set of linear SVM classifiers. Each classifier produces one grid line of the microarray image grid. In short, the proposed methodology consists of the following steps (Fig. 1):

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

2.1. Distance estimation between consecutive rows and columns

In the first step of the proposed gridding methodology, the distance between consecutive rows and columns is estimated. Even though the image dimensions are known and the number of spots in each row and column might also be known, the row height and column width cannot reliably be estimated due to image rotation or possibly inaccurate cropping of the scanned image. Furthermore, such an estimation would depend on user input and reduce the potential for high-throughput microarray image analysis. Instead, in order to find the optimal row height, the image is segmented into horizontal stripes with a height of d_r pixels, which are then averaged. If d_r is equal to the distance between the rows, the spots of all rows will be highly overlapping in the resulting averaged subimage, producing well defined white areas that are well separated from the black background, as shown on the left side of Fig. 2b. In the case of a suboptimal value of d_r, the spots will partly blend with the background (Fig. 2b, right side), producing numerous gray areas instead of distinct black and white areas. In order to select the optimal value of d_r, the standard deviation of the pixel intensities of the averaged subimage is used as an effective measure of spot overlap. A scheme based on the maximization of the standard deviation will result in the determination of the optimal row height d_r, whereas the optimal column width d_c is likewise estimated.

In more detail, 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. These subimages are then averaged into a single $x \times d_r$ image. Such images for several values of d_r are illustrated in Fig. 3.

The range of d_r values tested can be specified by the user as a parameter, but a wide range ensures successful estimation without user intervention and is thus preferred. The standard deviation of the averaged subimages is calculated for all values of d_r within that range, using a small step in the order of a fraction of a pixel. The values of d_r for which the standard deviation is a local maximum are selected as candidates for the optimal distance estimation, as denoted by the arrows in Fig. 4. The local maxima are most often located on multiples of the optimal d_r value (points a and d of Fig. 4), as a distance estimation of $n \cdot d_r$ also results in highly overlapping spots. Other local maxima (points b, c and e) may be present, depending on the rotation of the image. For each one of the selected d_r values, the average value of the standard deviation in their neighborhood is calculated. The resulting value of d_r is the one that exceeds its neighborhood average by a greater ratio. In the case shown, the greatest ratio is observed for point a, which exceeds the average of its neighborhood by 19.61% and is thus selected.

2.2. Rotation estimation

By analyzing the averaged $x \times d_r$ subimage for the estimated distance d_r, it is possible to calculate the angle of rotation of the original

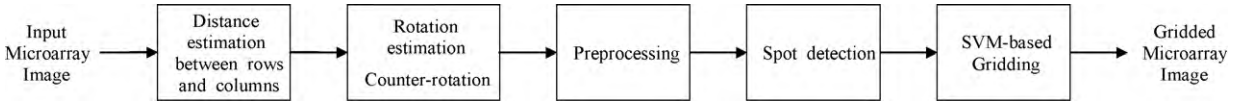

Fig. 1. Block diagram of the proposed methodology.

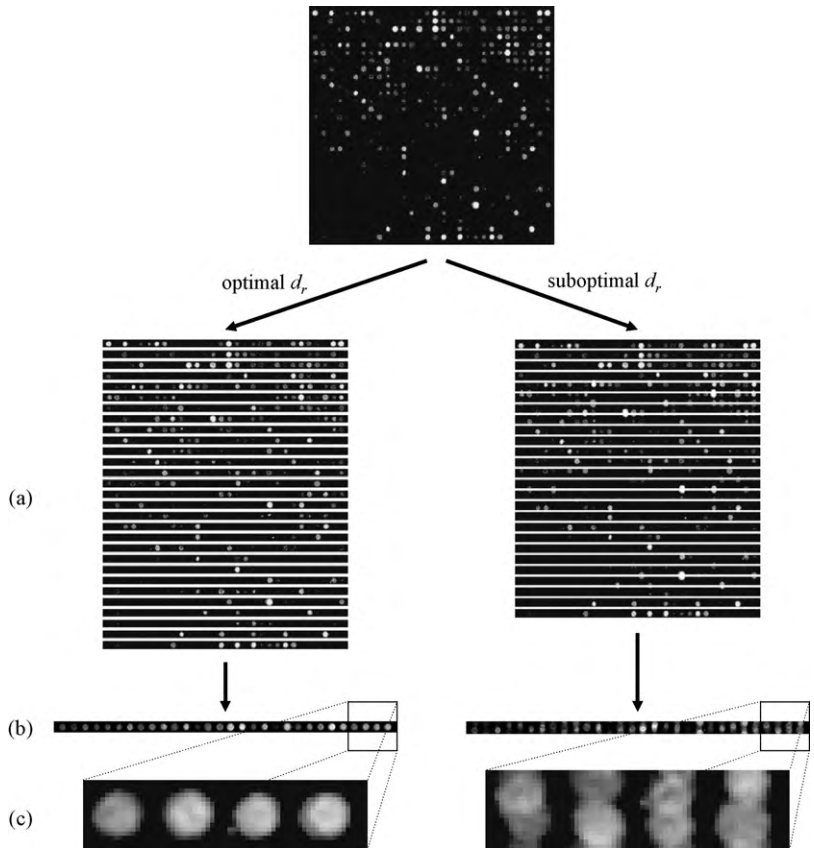

Fig. 2. Production of (a) horizontal subimages, (b) averaged subimage for optimal d_r and suboptimal d_r and (c) detail of averaged subimages.

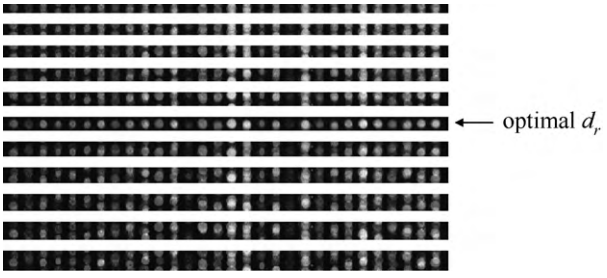

Fig. 3. The averaged row subimages produced for various values of d_r.

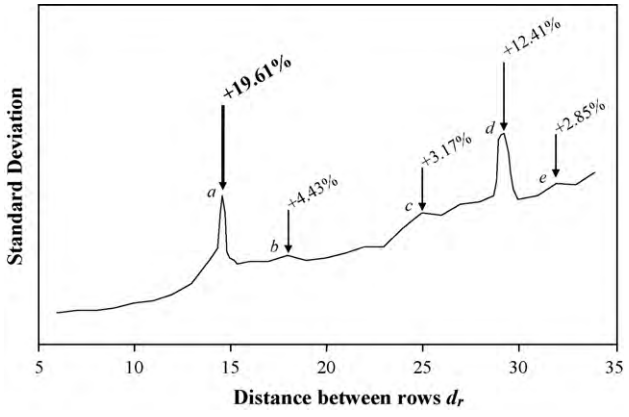

Fig. 4. Standard deviation of pixel intensity as a function of distance between rows d_r. The selected point a is indicated in bold.

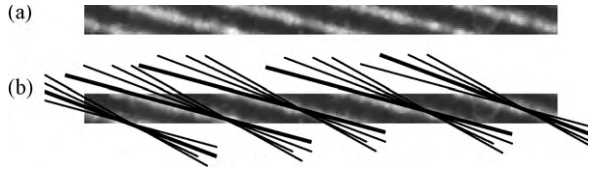

Fig. 5. (a) The averaged row subimage produced by a rotated microarray image and (b) the directions of highest average intensity.

microarray image. Fig. 5a depicts an $x \times d_r$ averaged subimage produced from a microarray image that has been manually rotated. In order to estimate the rotation angle of the image, a large number of the brightest pixels of the subimage are randomly selected. Starting from each of these pixels, the average pixel intensity over all directions ranging from -45° to $+45^\circ$ is calculated. The direction that results in the highest average intensity is chosen, as shown in Fig. 5b. The rotation estimated from the averaged subimage is the median of the chosen directions of all the selected pixels. This procedure is repeated for the averaged $d_c \times y$ subimage generated using the column distance d_c estimation. The final result is the arithmetic mean of the two image rotation angle estimations. Finally, the input image is counter-rotated so as to realign the rows and columns of spots to the x- and y-axes. The values of d_r and d_c are recalculated for the counter-rotated image.

2.3. Image preprocessing

This step involves the normalization of the microarray image by adjusting the intensity histogram into the range 0–255. This results in effective use of the full dynamic range of the 8-bit image. The edges of the spots are detected by the application of the Sobel operator on the normalized image. A threshold T is used to isolate the

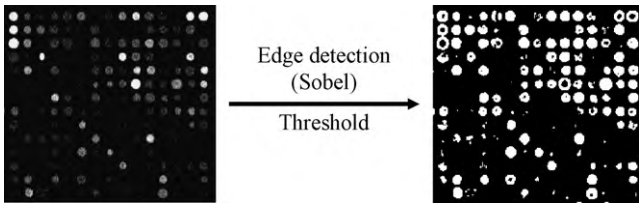

Fig. 6. The result of edge detection and thresholding.

sharpest edges, which correspond to prevalent spots, as shown in Fig. 6.

2.4. Spot detection

The thresholded image (Fig. 7a) is analyzed, in order to locate pixel groups that contain consecutive white pixels. The pixels of a group reside on the same spot edge. Each group is represented as a rectangle that circumscribes the pixels of the group, as illustrated in Fig. 7b. Ideally, each rectangle should contain the edge of a single microarray spot, however, depending on the threshold used and the noise present in the image, it might also include artifacts or multiple merged spots. Subsequently, only the rectangles that have specific shape and size characteristics should be considered valid, therefore a method for filtering the spots is employed.

The rectangles should be quasi-square in order to contain only one microarray spot, therefore the ratio of the smaller to the larger side of each rectangle must be close to unity. Also, each spot should belong to exactly one row and one column, therefore its size should not exceed the distance between rows or columns in the image. Hence, any pixel group that has a diagonal longer than $\sqrt{d_r^2 + d_c^2}$ is discarded. The output of the pixel group filtering is shown in Fig. 7c.

2.5. SVM-based gridding

In general, an SVM classifier [13] is provided with a training set $D = \{(\overline{x_i}, c_i) | \overline{x_i} \in \Re^2, c_i \in \{-1, +1\}\}$, which consists of vectors $\overline{x_i}$ and their respective class labels c_i. It produces the normal vector $\overline{w}$ and parameter b of the separating hyperplane $\bar{w} \cdot \bar{x} - b = 0$, which maximizes the margin between vectors $\overline{x_i}$ of different classes. The width of the margin is equal to $2/||w||$, therefore the widest margin is found by minimizing $||w||$ under the constraints $c_i(\bar{w} \cdot \overline{x_i} - b) \geq 1$, i.e. requiring that all the vectors in the training set are correctly classified. Fig. 8 presents an example of two possible lines for the separation of two classes of vectors. Although line l_2 is a valid separating line, line l_1 maximizes the margin ($m_1 > m_2$) and would therefore be chosen by the SVM.

The support vector machine described above is called a "hard-margin" SVM and does not take into account any outliers. One of its properties is that the separating hyperplane is determined by the support vectors, which are the ones that lie on the edges of the margin. Thus, in the case of outliers present inside the margin, the separating hyperplane will be placed suboptimally. Fig. 9 illustrates this case, where an outlier (denoted by the arrow)

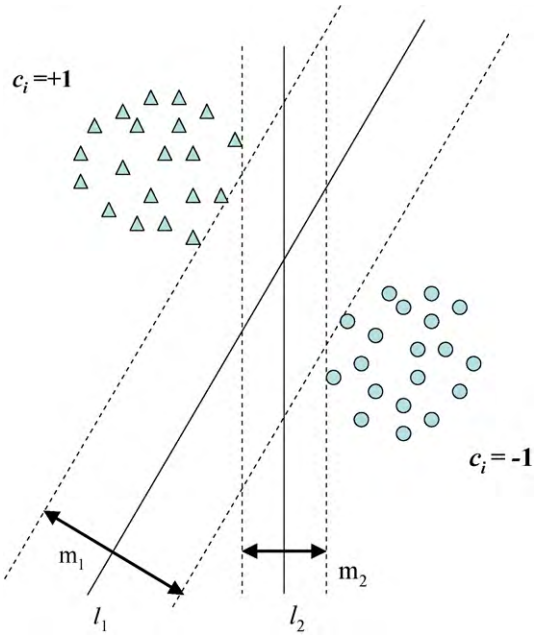

Fig. 8. Separating hyperplanes and their respective margins.

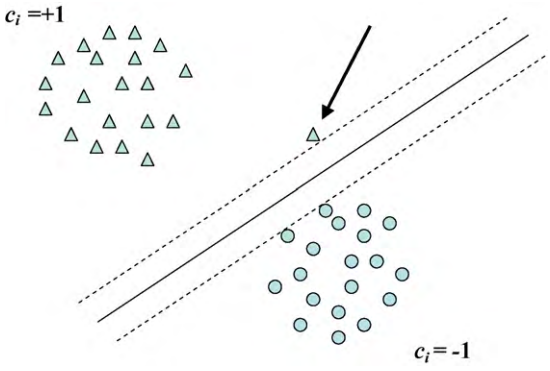

Fig. 9. Reduction of margin width due to an outlier, in the case of hard-margin linear SVM.

forces the SVM to position the separating hyperplane significantly closer to vectors with a class label of −1, reducing the width of the margin. This problem can be solved using the "soft-margin" SVM, where a slack variable ξ_i is introduced for each vector $\overline{x_i}$. The constraints are then formulated as $c_i(\bar{w} \cdot \overline{x_i} - b) \geq 1 - \xi_i$ and the separating hyperplane can be found by minimizing:

$$\min \left[\frac{1}{2} ||w||^2 + C \sum_i \xi_i \right] \quad (1)$$

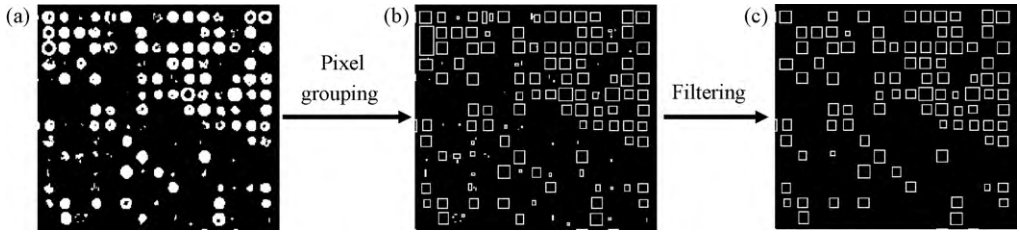

Fig. 7. Grouping and filtering of white pixels.

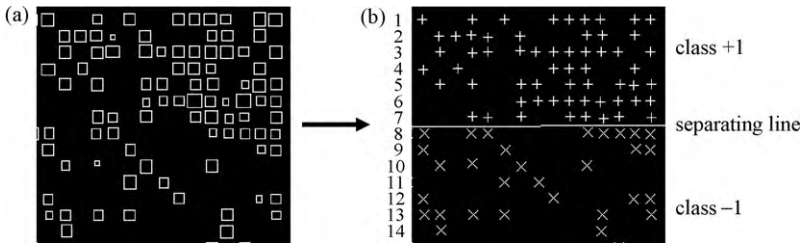

Fig. 10. (a) The valid spots and (b) the training set and resulting separating line produced by the SVM classifier for the separation of rows 7 and 8.

where C is a cost parameter that determines the effect of outliers on the resulting hyperplane. Large values of C result in a separating hyperplane that is mostly determined by any outliers, while on the other hand, if a smaller value of C is used, the separating hyperplane follows the general trend of the training set given to the classifier, ignoring any outliers. The hard-margin classifier is equivalent to a soft-margin classifier with an infinitely large C [14].

In order to use the SVM classifier for microarray gridding, the valid spots (Fig. 10a) that have been produced by the previous steps are first assigned into distinct rows and columns with respect to the distances d_r and d_c. For each pair of consecutive rows numbered k and $k+1$, the respective grid line that separates the spots of these rows is calculated by a soft-margin linear SVM classifier. Every valid spot in the image is represented by a two-dimensional vector $\overline{x_i}$ that consists of the coordinates of the center pixel of the valid spot, and these vectors comprise the training set D. The class label c_i of each valid spot is determined as a function of the row that it belongs to. More specifically if the spot belongs to any row with number ranging from 1 to k, it is assigned to class +1, else it belongs to the rows with numbers greater than or equal to $k+1$ and is thus assigned to class -1, as shown in Fig. 10b. The classifier is then trained and produces the separating line that maximizes the margin between the vectors $\overline{x_i}$, which is also the resulting grid line. It is only the training phase of the classifier that is used for the determination of the grid lines and not the testing phase.

If the successful detection of all spots in the image could be guaranteed, the training set would consist of only the necessary spots, i.e. those residing on rows k and $k+1$. However, in real microarray images, there are cases where several consecutive spots might be weakly expressed and therefore not detected, so adding spots from rows above k and below $k+1$ to the training set provides more useful data to the classifier for successful gridding.

In the case that row k contains less than two detected spots, the two grid lines that separate this row from rows $k-1$ and $k+1$ cannot be determined by the use of the SVM classifier. This is a rather rare case considering that the image is normalized during the preprocessing step. To cope with this limitation, the endpoints of the two grid lines are positioned equidistantly between the endpoints of the first neighboring grid lines above and below them. In the case where the top or bottom rows of spots contain less than two spots, the endpoints of the grid lines that cannot be determined are positioned d_r pixels further from the nearest grid lines.

Furthermore, the outliers that result from misdetected spots due to artifacts and noise require the use of the soft-margin SVM to diminish their effects. In Fig. 11, an outlier has been introduced into the SVM training set. It is evident that in the case of a small C (Fig. 11a), the margin is determined by the other spots in the row and the outlier is virtually ignored, whereas in the case of a large C (Fig. 11b), a single outlier determines the positioning of the separating line, resulting in a line that is significantly closer to most of the vectors of the top row, reducing the margin and rendering it suboptimal for gridding. The microarray gridding is completed after the application of the above procedure for the determination of the grid lines that separate each pair of consecutive columns of spots.

3. Results

The dataset used for the evaluation of the proposed method consists of 54 DNA microarray images, from the Stanford Microarray Database [15]. The images have 1900 × 5500 pixels and 16-bit gray level depth. The images include 48 blocks of about 870 spots each, for a total of 2,255,040 spots in the data set. They have been produced for the study of the gene expression profiles of 54 specimens of acute lymphoblastic leukemia, which span 37 positive and 17 negative to BCR-ABL [16], a fusion gene product resulting from translocation between the 9th and the 22th chromosomes. The dataset is accompanied by ground truth annotations regarding the positions and sizes of the spots.

In order to enhance the reliability of the results, the data set used for evaluation is a superset of the one used in [10] and [12], as it includes all 54 images instead of only 25 used in the previous studies. The statistical analysis is performed correspondingly, in order to produce directly comparable results. It is important to note that [10] presents a comparison to the state of the art methods [1,3,9], which it surpasses significantly with regards to microarray gridding accuracy. Therefore the evaluation of the proposed methodology is performed in comparison to [10]. For the statistical analysis, each spot was evaluated as being perfectly gridded when all its pixels reside completely 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 1. Out of more than two million spots

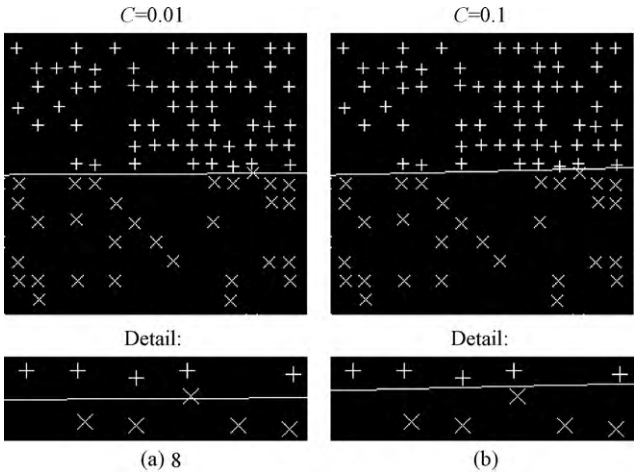

Fig. 11. The effect of an outlier as a function of the SVM cost parameter C. (a) Small value of C and (b) large value of C.

Table 1
Comparison of gridding results.

	Perfect	Marginal	Incorrect
Proposed method	96.4%	3.2%	0.4%
Bariamis et al. [12]	95.1%	4.5%	0.4%
Zacharia and Maroulis [10]	94.6%	4.8%	0.6%

Table 2
Percentage of correctly gridded spots as a function of the SVM cost parameter C and the threshold T.

Threshold T	SVM cost C			
	0.005	0.01	0.05	0.1
8	96.22%	96.20%	95.06%	94.74%
9	96.27%	96.27%	95.11%	94.80%
10	96.29%	96.30%	95.15%	94.84%
11	96.32%	96.35%	95.19%	94.88%
12	96.34%	**96.41%**	95.25%	94.93%
13	96.29%	96.38%	95.22%	94.91%
14	96.22%	96.34%	95.17%	94.86%
15	96.15%	96.29%	95.14%	94.82%
16	96.09%	96.25%	95.10%	94.79%
17	96.00%	96.18%	95.05%	94.74%
18	95.90%	96.12%	95.02%	94.70%
19	95.81%	96.04%	94.93%	94.62%
20	95.69%	95.95%	94.86%	94.54%
21	95.55%	95.84%	94.78%	94.46%
22	95.41%	95.73%	94.66%	94.35%
23	95.26%	95.62%	94.57%	94.25%
24	95.12%	95.50%	94.46%	94.14%

present in the data set, 96.4% spots were perfectly gridded, whereas 3.2 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 [10], and consequently it is also superior to [1,3,9]. In comparison to the preliminary version presented in [12] which displayed promising results, the achieved accuracy is increased as several changes have been included in the proposed method, such as the automatic determination of valid spot sizes based on the distance between rows d_r and columns d_c, as well as the inclusion of the valid spots from the whole image into the training set of each SVM classifier.

The gridding performance of the proposed method was evaluated using $C = 0.1$, 0.05, 0.01 and 0.005 and T ranging from 8 to 24. The SVM cost parameter C determines the effect that outliers or noise might have on the separating lines that the SVM produces, therefore a small value of C should be selected for successful gridding. The threshold T affects the sensitivity of the spot detection step, as well as its susceptibility to noise. The choice of $C = 0.01$ is supported by the results shown in Table 2, where it produces the most accurately gridded spots compared to the other values of C evaluated. Lowering the value of C results in negligible changes of accuracy, but the choice of a larger value would reduce the achieved accuracy. Even though the optimal value of C is usually application and data dependent, in the proposed method the choice of a value lower than the optimal results in comparable accuracy. Table 2 also illustrates that the proposed method is highly accurate for a wide range of thresholds T, as the greatest percentage of correctly gridded spots is 96.41% for $T = 12$ (denoted in bold), but the accuracy remains higher than 96% for T ranging from 8 to 19. The results illustrate that the effect of threshold selection only marginally affects the achieved accuracy.

Although the dataset only includes microarray images with rotation of up to a few degrees, an evaluation method was needed to assess the performance of the rotation detection step of Section 2.2 for a wider range of image rotation angles. We have therefore manually rotated the images of the dataset by angles θ_{real} ranging from -25° to $+25^\circ$ and used the proposed rotation detection method to compute an estimate θ_{est} of the rotation for each image. Based on that estimate, the images were counter-rotated and gridded. Table 3 presents the results of the rotation detection as a function of the rotation angle θ_{real}. The evaluation was made based on the mean and standard deviation of $\Delta\theta = \theta_{est} - \theta_{real}$, denoted as $m_{\Delta\theta}$ and $\sigma_{\Delta\theta}$, respectively. The mean difference was less than 1.3° for all cases, which resulted in negligible variation of the gridding accuracy compared to the original images. The variation of the accuracy was less than 0.3% in all cases. An example of an image rotated by 15°, as well as the counter-rotated image and the gridding result are illustrated in Fig. 12. In this case, $\Delta\theta$ was equal to 0.9°.

Fig. 13 illustrates the gridding resulting from the application of the proposed method in the presence of artifacts. More specifically, in Fig. 13a–c, several bright artifacts are present, whereas in Fig. 13d the top right part of the image has been affected by noise during the wet lab process. Despite the presence of these artifacts and noise, the proposed method achieves successful gridding in all those cases. Fig. 14 illustrates a microarray image area that includes a large and bright artifact. Even in the vicinity of the artifact, the gridding is not affected by its presence.

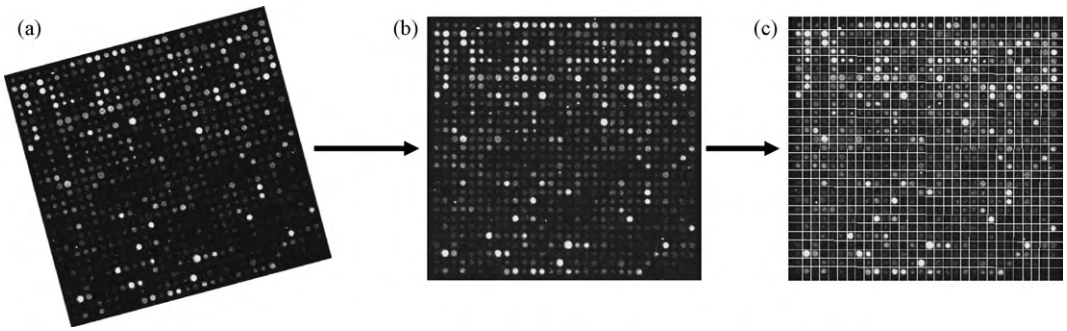

Fig. 12. (a) A microarray image rotated by 15°, (b) the counter-rotated image ($\Delta\theta = 0.9^\circ$) and (c) the resulting gridding for this image.

Table 3
Mean and standard deviation of difference between actual and detected rotation angles $\Delta\theta$.

θ_{real}	-25°	-20°	-15°	-10°	-5°	0°	5°	10°	15°	20°	25°
$m_{\Delta\theta}$	1.23°	0.62°	0.72°	0.72°	-0.21°	0.28°	0.63°	-0.42°	0.20°	-0.3°	-1.02°
$\sigma_{\Delta\theta}$	0.41°	0.83°	0.95°	0.58°	0.70°	0.35°	0.88°	0.61°	0.89°	0.72°	0.57°

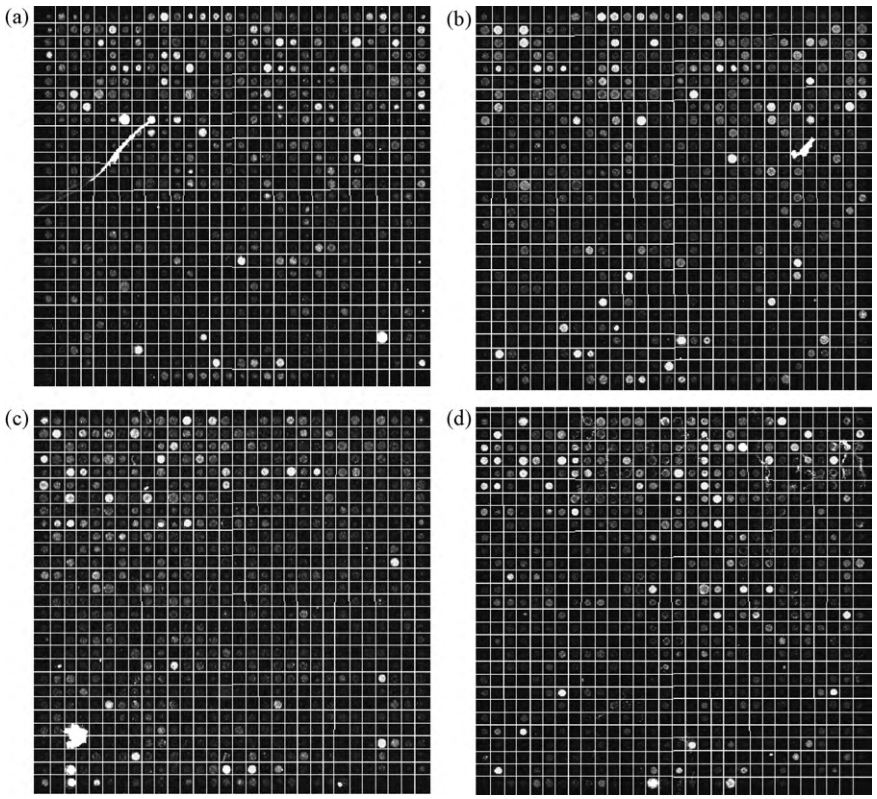

Fig. 13. Gridding examples: (a) large artifact, (b) and (c) small artifacts, and (d) noise at the top of the image.

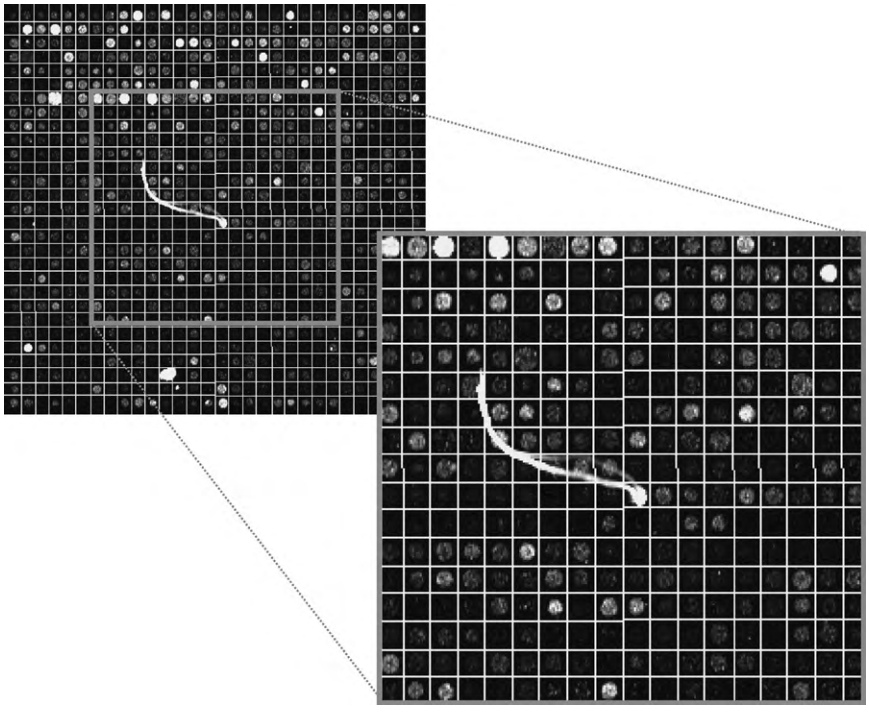

Fig. 14. Detail of successful gridding in the presence of a bright artifact.

4. Discussion and conclusions

In this paper, we presented a novel method for unsupervised microarray gridding, which consists of five steps. In the first step, the distance between rows and columns of spots is estimated. In the second step, the image rotation angle is estimated and the image is counter-rotated to align the rows and columns of spots with the *x*- and *y*-axes. In the third step, the input image is pre-processed, whereas the fourth step involves the spot detection and filtering. In the final step, a set of soft-margin linear support vector machine classifiers determine the positioning of the grid lines. The SVM produces the separating lines of the grid so as to maximize

the margin between the rows and columns of spots, and displays high tolerance to outliers that result from misdetected spots due to artifacts and noise. Furthermore, the proposed method allows high accuracy gridding for a wide range of operating parameters by employing efficient filtering of the detected spots based on their size and shape, in addition to using soft-margin linear SVM classifier with an extended training set.

Overall, the advantage of the proposed method is that it manages to perform successful gridding of DNA microarray images in the presence of the following conditions: irregular and weakly expressed spots, noise and artifacts, as well as rotation. The effects of the irregular spots, the noise and the artifacts are diminished by the high tolerance of the soft-margin SVM to outliers, as well as by the spot filtering included in the spot detection step. Furthermore, the generalization performance of the SVM classifier allows it to determine the grid lines in the presence of weakly expressed spots. Lastly, the proposed method estimates the image rotation angle and counter-rotates the input image in order to produce accurate gridding. A potential disadvantage of the proposed method is that the SVM classifiers require several detected spots in each row and column of spots. Rarely, most of the spots in a row or column might be weakly expressed and not detected. In such cases, which account for less than 0.1% of the rows and columns in the data set, the grid line positioning is determined by the nearest grid lines.

Out of more than two million spots present in the data set, 96.4% spots were perfectly gridded, whereas 3.2 and 0.4% were marginally and incorrectly gridded, respectively. These experimental results show that the proposed method achieves higher quality gridding than the state of the art method presented in [10], providing the potential of achieving perfect gridding for the vast majority of the spots.

Acknowledgements

The authors wish to thank the anonymous reviewers for their useful suggestions. 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.

References

[1] M.B. Eisen, ScanAlyze, http://rana.lbl.gov/EisenSoftware.htm; November 1999.
[2] Biodiscovery, Inc., ImaGene, http://www.biodiscovery.com/imagene.asp; 2005.
[3] Hegde P, Qi R, Abernathy K, Gay C, Dharap S, Gaspard R, et al. A concise guide to cDNA microarray analysis. BioTechniques 2000;29(3):548–62.
[4] Angulo J, Serra J. Automatic analysis of DNA microarray images using mathematical morphology". Bioinformatics 2003;19(5):553–62.
[5] Katzer M, Kummert F, Sagerer G. A Markov random field model of microarray gridding. In: Proceedings of the SAC. New York: ACM; 2003.
[6] Giannakeas N, Fotiadis DI, Politou AS. An automated method for gridding in microarray images. In: Proceedings of the 28th IEEE EMBS annual international conference. 2006. p. 5876–9.
[7] Giannakeas N, Fotiadis DI. An automated method for gridding and clustering-based segmentation of cDNA microarray images. Computerized Medical Imaging and Graphics 2009;33(January (1)):40–9.
[8] Hartelius K, Cartstensen JM. Bayesian grid matching. IEEE Transactions on PAMI 2003;25(2):162–73.
[9] Blekas K, Galatsanos NP, Likas A, Lagaris IE. Mixture model analysis of DNA microarray images. IEEE Transactions on Medical Imaging 2005;24(7): 901–9.
[10] Zacharia E, Maroulis D. An original genetic approach to the fully automatic gridding of microarray images. IEEE Transactions on Medical Imaging 2008; 27(6).
[11] Antoniol G, Ceccarelli M. Microarray image gridding with stochastic search based approaches. Image and Vision Computing 2007;25(2):155–63.
[12] Bariamis D, Maroulis D, Iakovidis DK. Automatic DNA microarray gridding based on support vector machines. In: Proceedings of the 8th IEEE international conference on bioinformatics and bioengineering (BIBE). 2008.
[13] Cortes C, Vapnik V. Support-vector networks. Machine Learning 1995;20(3):273–97.
[14] Theodoridis S, Koutroumbas K. Pattern recognition. Fourth ed. New York: Academic Press; 2008.
[15] Stanford Microarray Database, http://smd.stanford.edu/.
[16] Juric D, Lacayo NJ, Ramsey MC, Racevskis J, Wiernik PH, Rowe JM, et al. Differential gene expression patterns and interaction networks in bcr-abl-positive and -negative adult acute lymphoblastic leukemias. Journal of Clinical Oncology 2007;25(April (11)):1341–9.

Dimitris Bariamis received the B.Sc. degree in Informatics and Telecommunications in 2006 and the Ph.D. Degree in biomedical data analysis in 2009 from the Dept. of Informatics and Telecommunications, University of Athens, Greece. His research interests include image and signal analysis, real time systems, hardware architectures, pattern recognition and bioinformatics.

Dimitris Maroulis received the B.Sc. degree in Physics, the M.Sc. degree in Radioelectricity, the M.Sc. in Cybernetics with honors from the University of Athens. He received the Ph.D. degree from the Dept. of Informatics and Telecommunications, University of Athens in 1990. He served as a research fellow at the Meudon Observatory, France. Currently, he is an associate professor in the Dept. of Informatics and Telecommunications of the University of Athens. He has co-authored more than 100 research papers and book chapters, and he is a reviewer in more than 10 international journals. He has been actively involved in more than 10 European and National R&D projects. His research interests include image/signal processing and analysis, data acquisition and real-time systems, pattern recognition with applications on biomedical systems and bioinformatics.

Dimitris K. Iakovidis was born in Athens in 1973. He received his B.Sc. degree in Physics from the University of Athens, Greece. In April 2001, he received his M.Sc. degree in Cybernetics with honors and in February 2004 his Ph.D. degree from the Dept. of Informatics and Telecommunications, University of Athens, Greece. Currently he is Assistant Professor in the Dept. of Informatics and Computer Technology of the Technological Educational Institute of Lamia. Dr. Iakovidis has co-authored more than 80 research papers, he is a reviewer in many international journals, and he has been actively involved in several European and National R&D projects. His research interests include signal/image processing and analysis, data mining, and pattern recognition with applications on biomedical systems and bioinformatics.