

Sub-grid Detection in DNA Microarray Images

Luis Rueda

Department of Computer Science
University of Concepción

Edmundo Larenas 215, Concepción, 4030000, Chile

Phone: +56 41 220-4305, Fax: +56 41 222-1770

lrueda@udec.cl

Abstract. Analysis of DNA microarray images is a crucial step in gene expression analysis, as it influences the whole process for obtaining biological conclusions. When processing the underlying images, accurately separating the sub-grids is of supreme importance for subsequent steps. A method for separating the sub-grids is proposed, which aims to first, detect rotations in the images independently for the x and y axes, corrected by an affine transformation, and second, separate the corresponding sub-grids in the corrected image. Extensive experiments performed in various real-life microarray images from different sources show that the proposed method effectively detects and corrects the underlying rotations and accurately finds the sub-grid separations.

Keywords: Microarray image gridding, image analysis, image feature and detectors.

1 Introduction

One of the most important technologies used in molecular biology are microarrays. They constitute a way to monitor gene expression in cells and organisms under specific conditions, and have many applications. Microarrays are produced on a chip (slide) in which DNA extracted from a tissue is hybridized with the one on the slide, typically in two channels. The slide is then scanned at a very high resolution generating an image composed of sub-grids of spots (in the case of cDNA microarrays) [1,2]. Image processing and analysis are two important aspects of microarrays, since the aim of obtaining meaningful biological conclusions depends on how well these stages are performed. Moreover, many tasks are carried out sequentially, including gridding [3,4,5,6,7], segmentation [8,9], quantification [2], normalization and data mining [1]. An error in any of these stages is propagated to the rest of the process. When producing DNA microarrays, many parameters are specified, such as the number and size of spots, number of sub-grids, and even their exact location. However, many physical-chemical factors produce noise, misalignment, and even deformations in the sub-grid template that it is virtually impossible to know the exact location of the spots after scanning is performed, at least with the current technology. The first stage in the analysis is to find the location of the sub-grids (or gridding), which is the focus of our paper. Roughly speaking, gridding consists of determining the spot locations in a microarray image (typically, in a sub-grid). The problem, however, is that microarray images are divided in sub-grids,

done to facilitate the spot locations. While quite a few works have been done on locating the spots in a sub-grid, they all assume the sub-grids are known, and this is the problem considered in this paper, more formally stated as follows.

Consider an image (matrix) $A = \{a_{ij}\}$, $i = 1, \dots, n$ and $j = 1, \dots, m$, where $a_{ij} \in \mathbb{Z}^+$, and A is a sub-grid of a cDNA microarray image¹ [1] (usually, a_{ij} is in the range [0.65, 535] in a TIFF image). In what follows, we use $A(x, y)$ to denote a_{ij} . The aim is to obtain a matrix G (grid) where $G = \{g_{ij}\}$, $i = 1, \dots, n$ and $j = 1, \dots, m$, $g_{ij} = 0$ or $g_{ij} = 1$ (a binary image), with 0 meaning that g_{ij} belongs to the grid. This image could be thought of as a “free-form” grid. However, in order to strictly use the definition of a “grid”, our aim is to obtain vectors \mathbf{v} and \mathbf{h} , $\mathbf{v} = [v_1, \dots, v_m]^t$, $\mathbf{h} = [h_1, \dots, h_n]^t$, where $v_i \in [1, m]$ and $h_j \in [1, n]$. Each vertical and horizontal vectors are used to separate the sub-grids. As seen later, rotation correction facilitates finding this template.

Many approaches have been proposed for spot detection and image segmentation, which basically assume that the sub-grids are already identified. That is, they typically work on a sub-grid, rather than on the entire microarray image. The Markov random field (MRF) is a well known approach that applies different application specific constraints and heuristic criteria [3,10]. Another gridding method is mathematical morphology, which represents the image as a function and applies erosion operators and morphological filters to transform it into other images resulting in shrinkage and area opening of the image, and which further helps in removing peaks and ridges from the topological surface of the images [11]. Jain’s [8], Katzer’s [12], and Stienfath’s [13] models are integrated systems for microarray gridding and quantitative analysis. A method for detecting spot locations based on a Bayesian model has been recently proposed, and uses a deformable template model to fit the grid of spots in such a template using a posterior probability model which learns its parameters by means of a simulated-annealing-based algorithm [3,5]. Another method for finding spot locations uses a hill-climbing approach to maximize the energy, seen as the intensities of the spots which are fit to different probabilistic models [7]. Fitting the image to a mixture of Gaussians is another technique that has been applied to gridding microarray images by considering radial and perspective distortions [6].

All these approaches, though efficient, assume the sub-grids have already been identified, and hence they proceed on a single sub-grid, which has to be specified by the user. A method used for gridding that does not use this assumption has been proposed in [14,15]. It performs a series of steps including rotation detection based on a simple method that compares the running sum of the topmost and bottommost parts of the image. It performs rotations locally and applies morphological opening to find sub-grids. This method, which detects rotation angles wrt one of the axes, either x or y , has not been tested on images having regions with high noise (e.g. bottommost 1/3 of the image is quite noisy).

In this paper, we focus on automatically detecting the sub-grids given the *entire* microarray image. The method proposed here uses the well-known Radon transform and an information-theoretic measure to detect rotations (wrt the x and y axes), which

¹ The aim is to apply this method to a microarray image that contains a template of rows and columns of sub-grids.

are corrected by an affine transformation. Once corrected, the imaged is passed through a mathematical-morphology approach to detect valleys, which are then used to separate the sub-grids. Section 2 discuss the details of the proposed method, while Section 3 presents the experiments on real-life images, followed by the conclusions to the paper.

2 Sub-grid Detection

The proposed sub-grid detection method aims to first correct any rotation of the image by means of the Radon transform [4,16]. After this, the (x or y -axis) running sum of pixels is passed through morphological operators to reduce noise, and then the detected valleys denote the separation between sub-grids. Note, however, that in order to process the microarray image in subsequent steps (i.e. gridding and segmentation) the image does not have necessarily to be rotated. Although, the method that we proposed herein performs the rotation correction, this can be avoided by generating the horizontal and vertical lines, \mathbf{v} and \mathbf{h} , for the corrected image, applying the inverse affine transformation to \mathbf{v} and \mathbf{h} , obtaining the sub-grids in the original image, and hence not degrading the quality of the image.

Rotations of the image are seen in two different directions, wrt the x and y axes, in the aim at finding two independent angles of rotation for an affine transformation, and for this the Radon transform is applied. Roughly speaking, the Radon transform, which can be seen as a particular case of the Hough transform, is the integral of an n -dimensional function over all types of functions of dimension smaller than n . In two dimensions, like in the case of images, the Radon transform is the integral of the 2D function over all types of real-valued functions, e.g. polynomials, exponentials, etc. In particular, when the latter function is a line, the Radon transform can be seen as the projection of the two-variable function (the integral) over a line having a direction (slope) and a displacement (wrt the origin of the coordinate system); this is the case considered in this work. The Radon transform has been found quite useful in the past few decades in many applications, including medicine (for the computed axial tomography, or CAT), geology, and other fields of science. In two-dimensional functions projected onto lines, it works as follows. Given an image $A(x, y)$, the Radon transform performs the following transformation:

$$R(p, t) = \int_{-\infty}^{\infty} A(x, t + px) dx, \quad (1)$$

where p is the slope and t its intercept. The rotation angle of the image with respect to the slope p is given by $\phi = \arctan p$. For the sake of the notation, $R(\phi, t)$ is used to denote the Radon transform of image A . Each rotation angle ϕ gives a different one-dimensional function, and the aim is to obtain the angle that gives the best alignment with the rows and columns. This will occur when the rows and columns are *parallel* to the x or y -axis. There are many ways to state this as an optimization problem, and different objective functions have been used (cf. [3]). In this work, an entropy-based function is used. Assuming the sub-grids are (or should be²) aligned wrt the y -axis (and x -axis),

² The aim is to detect the correct alignment. While the assumption made here is to formalize the model, such an alignment is indeed what is found in the proposed approach.

the one-dimensional projected function will show well-accentuated peaks, each corresponding to a column (row) of spots and deep valleys corresponding to the background separating the spots and sub-grids. Assuming the experimental setup in the microarray fabrication considers a reasonable separation between the sub-grids (otherwise it would not be possible to detect such a separation), deeper and wider valleys will be expected between sub-grids, and which are then used to detect the corresponding sub-grids.

To compute the entropy function, the $R(\phi, t)$ function is normalized and renamed $R'(\phi, t)$, such that $\int_t R'(\phi, t) = 1$. The best alignment will thus occur at the angle ϕ_{min} that minimizes the *entropy* as follows:

$$H(\phi) = - \int_{-\infty}^{\infty} R'(\phi, t) \log R'(\phi, t) dt. \quad (2)$$

One of the problems, however, the entropy function has is that, depending on the rotation angle ϕ , the sides of the one-dimensional function tend to diminish the “uniformity” of the function, and hence bias the entropy measure. This occurs when ϕ is near $\pi/4$. Since reasonable small rotations are expected to occur, small angles are considered (no more than 10 degrees of rotation). Also, the resulting signal function is on a discrete domain, i.e. ϕ takes discrete values, and hence the entropy is computed as follows:

$$H(\phi) = - \sum_{t=-\infty}^{\infty} R'(\phi, t) \log R'(\phi, t) dt. \quad (3)$$

Note that $R(\phi, t)$ is normalized into $R'(\phi, t)$, such that $\sum_t R'(\phi, t) = 1$.

The image is checked for rotations in both directions, wrt the x and y axes, obtaining two different angles of rotation ϕ_{min_x} and ϕ_{min_y} respectively. The positions of the pixels in the new image, $[uv]$, are obtained as follows:

$$[uv] = [xy1]T, \quad (4)$$

where T is the following 3×2 matrix:

$$T = \begin{bmatrix} \alpha & \beta \\ \beta & \alpha \\ \gamma_1 & \gamma_2 \end{bmatrix} \quad (5)$$

The first two parameters, α and β , are given by the best angles of rotation found by the Radon transform, ϕ_{min_x} and ϕ_{min_y} , and computed as follows:

$$\alpha = s \cos \phi_{min_x} \quad (6)$$

$$\beta = s \sin \phi_{min_y} \quad (7)$$

where s is a scaling factor (in this work, s is set to 1), and γ_1 and γ_2 are translation factors, which are set to 0. The transformed image, A' , is reconstructed by interpolating the intensities of pixels x and y and their neighbors; in this work, *bicubic* is used.

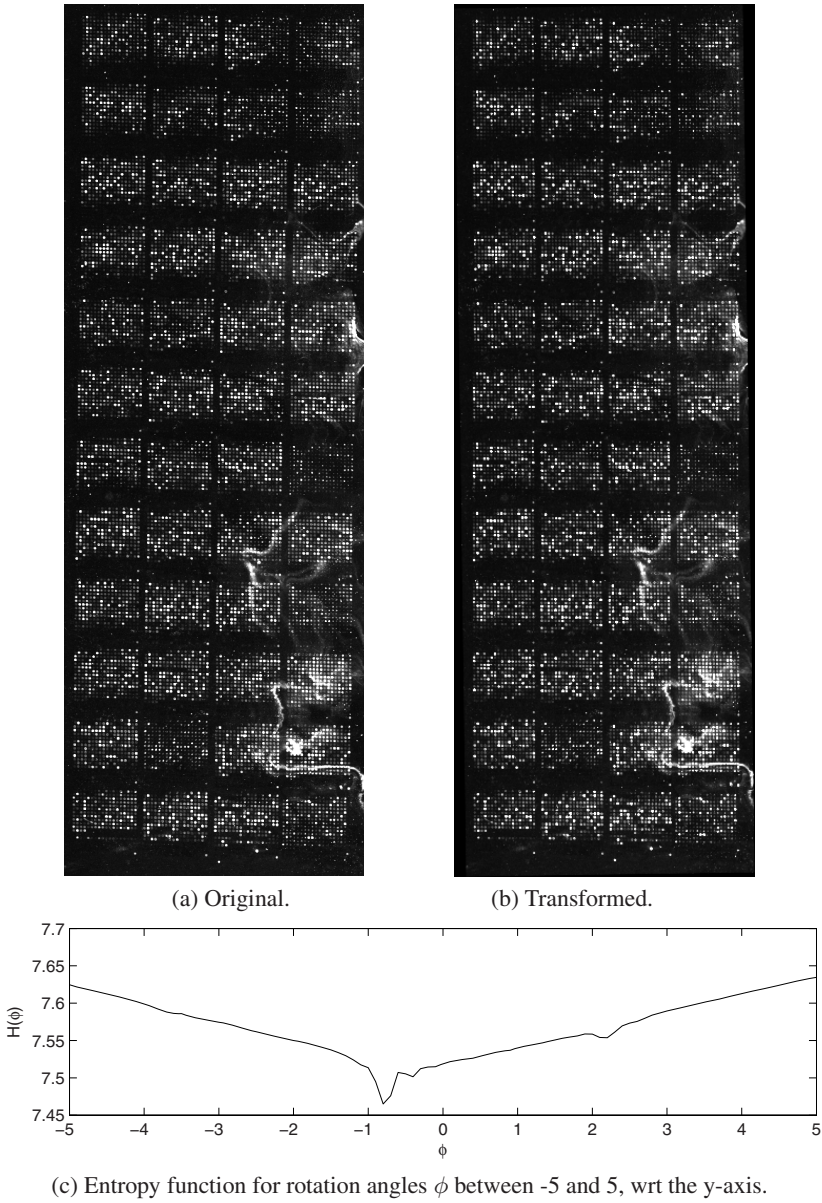


Fig. 1. A sample DNA microarray image (AT-20385-ch1) drawn from the Stanford microarray database, along with the transformed (by means of the affine transformation) image, and the entropy function wrt the y -axis

A sample image from the Stanford microarray database is shown in Fig. 1(a), namely image AT-20385-ch1. This image has been reduced in size, and the whole image can

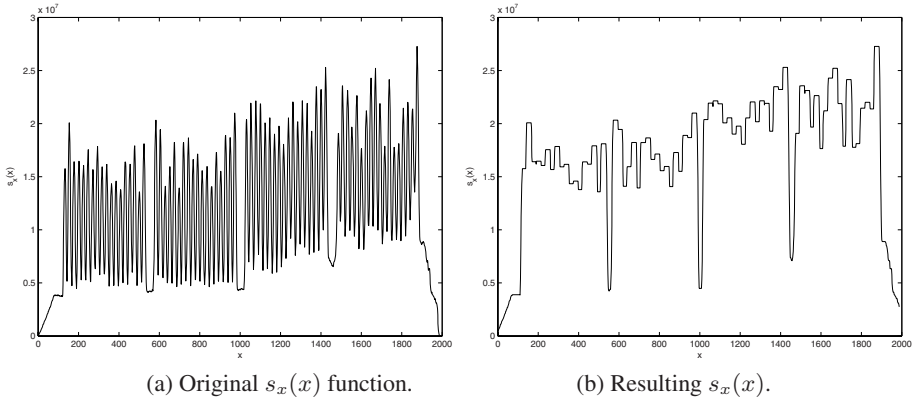


Fig. 2. The original running sum function, $s_x(x)$, before and after applying the morphological operators, for image AT-20385-ch1

be found in the database³. This image contains 48 sub-grids arranged in 12 rows and 4 columns. This image is rotated -0.8 degrees wrt the y -axis and 1.5 degrees wrt the x -axis (the latter not easily visualized by eye). These rotations are accurately detected by means of horizontal and vertical Radon transforms, which are performed independently, and the resulting image after applying the affine transformation as per (4) is shown in Fig. 1(b). Fig. 1(c) depicts the entropy function as per (3) for all angles ϕ between -5 and 5 degrees wrt the y -axis. The global minimum at $\phi_{min_y} = -0.8$ is clearly visible in the plot.

The next step consists of finding the lines separating the sub-grids. For this, it is assumed that the angles that give the optimal affine transformation are ϕ_{min_x} and ϕ_{min_y} , and the “transformed” image is A' . To detect the vertical lines⁴, the running sum of pixel intensities of A' is computed for all values of y , obtaining the function $s_x(x) = \sum_y A'(x, y)$. To detect the lines separating the sub-grids, the n deepest and widest valleys are found, where n is the number of columns of sub-grids (parameter given by the user). The function $s_x(x)$ is passed through morphological operators (dilation, $s_x(x) \oplus b$, followed by erosion, $s_x(x) \ominus b$, with $b = [0, 1, 1, 1, 1, 0]$) to in order to remove noisy peaks. After this, n potential centers for the sub-grids are found by dividing the range of $s_x(x)$ into nearly-equal parts. Since $s_x(x)$ contains many peaks (each representing a spot), and the aim is to detect sub-grids, the function is passed, again, by morphological operators (dilation, $s_x(x) \oplus b$, followed by erosion, $s_x(x) \ominus b$), where b depends on the spot width in pixels (scanning resolution), and computed as follows. The number of pixels p for each spot is found by means of a “hill-climbing” procedure that finds the lowest valleys (in this paper, six are found) around the potential centers for each sub-grid. Averaging the distances between the valleys found gives a resolution r (width of each spot), and the morphological operand b is set as follows: $b = [01_r0]$,

³ The full image is electronically available at smd.stanford.edu, in category “Hormone treatment”, subcategory “Transcription factors”, experiment ID “20385”, channel “1”.

⁴ The details for detecting the horizontal lines are similar and omitted to avoid repetition.

where 1_r is a run of r ones. Once the morphological operators are applied, the lines separating the sub-grids are obtained as the centers of the deepest and widest valleys between the potential centers found previously.

Fig. 2 shows the running sum of pixel intensities along the x -axis, for image AT-20385-ch1. The original $s_x(x)$ function is plotted in Fig. 2(a), which contains many sharpened peaks corresponding to each column of spots, and hence making it difficult to detect the separation between grids (the widest and deepest valleys). The resulting function after applying the morphological operators is depicted in Fig. 2(b), in which the sharpened peaks tend to “disappear”, while the deepest valleys are preserved. The three deepest and widest valleys, which can be easily visualized by eye, correspond to the three lines separating the four columns of sub-grids. Note that it is not difficult to detect these three valleys despite the image does not clearly show the separation between columns of sub-grids.

3 Experimental Results

For the experiments, two different kinds of cDNA microarray images have been used. The images have been selected from different sources, and have different scanning resolutions, in order to study the flexibility of the proposed method to detect sub-grids under different spot sizes.

The first set of images has been drawn from the Stanford Microarray Database (SMD), and corresponds to a study of the global transcriptional factors for hormone treatment of *Arabidopsis thaliana*⁵ samples. Ten images were selected for testing the proposed method, and they correspond to channels 1 and 2 for experiments IDs 20385, 20387, 20391, 20392 and 20395. The list of the images used for the testing are listed in Table 1. The images have been named using AT (which stands for *Arabidopsis thaliana*), followed by the experiment ID, and the channel number (1 or 2). The images have a resolution of 1910×5550 pixels and are in TIFF format. The spot resolution is 24×24 pixels per spot, and the separation between sub-grid columns is about 40 pixels, which is very low. Each image contains 48 sub-grids, arranged in 12 rows and 4 columns. Also, listed in the table are for each image, the angles of rotation wrt to the x and y axes, ϕ_{min_x} and ϕ_{min_y} respectively, found by maximizing (3). The last column lists the accuracy in terms of percentage, which represents the number of sub-grids correctly detected. All the images are rotated with respect to both x and y axes. Also, the angles of rotation are different for the two axes, x and y , for all the images. These rotations are detected and corrected by the proposed method. Note that even when the angles of rotation are small, e.g. 0.5 and 0.2 for AT-20395 ch1 and ch2, it is critical to detect these angles and correct them, since the resolution of the images is very high and a small angle variation will produce a displacement of a vertical line by a large number of pixels. For example, a rotation angle of 0.8 degrees wrt to the y -axis will produce a displacement of 25 pixels (for images AT-20385 ch1 and ch2). Since the separation between sub-grids is about 40 pixels, it is quite difficult, though possible, to detect the vertical lines separating the sub-grids, while after detecting and correcting the rotations,

⁵ The images can be downloaded from smd.stanford.edu, by searching “Hormone treatment” as category and “Transcription factors” as subcategory.

Table 1. Test images drawn from the SMD, angles of rotation and percentage of sub-grids detected

Image	ϕ_{min_x}	ϕ_{min_y}	Accuracy
AT-20385-ch1	1.5	-0.8	100%
AT-20385-ch2	1.5	-0.8	100%
AT-20387-ch1	0.8	-0.1	100%
AT-20387-ch2	0.8	-0.1	100%
AT-20391-ch1	0.9	-0.2	100%
AT-20391-ch2	0.9	-0.2	100%
AT-20392-ch1	1.0	-0.2	100%
AT-20392-ch2	1.0	-0.2	100%
AT-20395-ch1	0.5	0.2	100%
AT-20395-ch2	0.5	0.2	100%

it is rather easy to separate the sub-grids – this is observed by the 100% accuracy the method yields on all the images of the SMD.

To observe visually how the method performs, Figs. 3 and 4 show two images, AT-20385-ch1 and AT-20387-ch1, in their original form, and the resulting images obtained after applying the proposed method (Figs. 3(b) and 4(b)). For AT-20385-ch1, the rotation wrt to the y -axis is clearly visible in the picture and it is seen how it is corrected. It is clear also how the sub-grids are accurately detected, specially the vertical lines separating the grids, despite the image contains many noisy artifacts resulting from the microarray experimental stages – some sub-grids on the bottom right part of the image are even quite noisy. For AT-20387-ch1 the angle of rotation wrt to the y -axis is very small, $\phi_{min_y} = 0.1$; however, it is detected and corrected by the proposed method. It is clear from Figs. 3 and 4 how the sub-grids are detected and well separated by the vertical and horizontal lines.

The second test suite consists of a set of images produced in a microarray study of a few genes in an experiment where human cultured cell line was used to look at the toxicogenomic effects of two pesticides that were found in the rural drinking water [17], namely the human toxicogenomic dataset (HTD). Ten images were selected for testing the proposed method, which correspond to five different experiments in two channels, Cy3 and Cy5. The images are listed in Table 2, and are named by using HT (which stands for human toxicogenomic), followed by the channel number (Cy3 or Cy5) and the experiment ID. The images have a resolution of 7013×3514 pixels, and are in TIFF format. The spot resolution is 40×40 pixels per spot, and the separation between sub-grid columns and rows is about 400 pixels. Each image contains 32 sub-grids, arranged in 8 rows and 4 columns. The second, third and fourth columns have the same meaning as in Table 1. As in the other set of images, all the sugrids are detected with 100% accuracy, denoting the efficiency of the proposed method. While the angles of rotation for the images of the HTD are quite small, they are detected and corrected by the proposed method. However, a small angle for these images produce a large displacement in terms of pixels, since their resolution is higher than that of the images of the SMD. For example, for image HT-Cy3-12667177, a rotation of -0.2 degrees produces a displacement of 8 pixels, which in turn, affects the process of detecting the sub-grid separation.

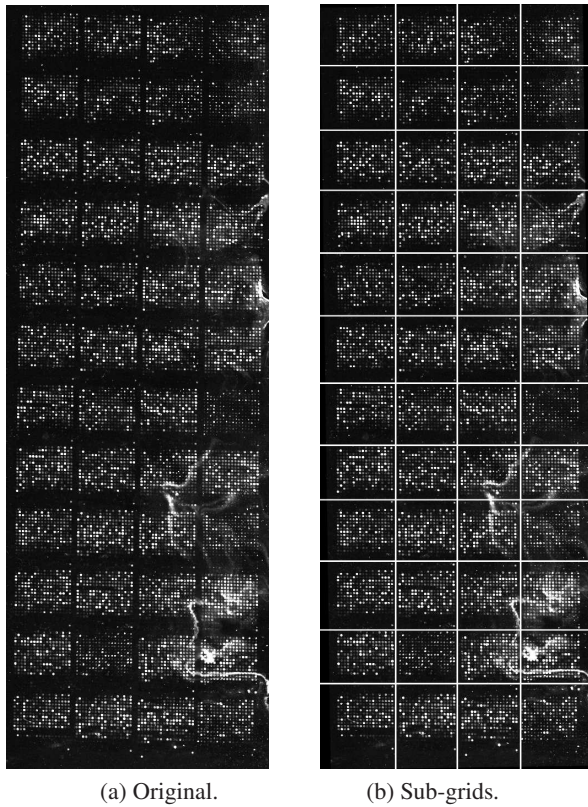


Fig. 3. Original and sub-grids detected by the proposed method, for images AT-20385-ch1 drawn from the SMD

An image, HT-Cy5-12663787, drawn from the HTD is shown in Fig. 5. The original and sub-grids detected are shown in (a) and (b) respectively. Even though the sub-grids are well separated by a large number of pixels, the image contains a lot of noise in the separating area, and thus, making it difficult to detect the sub-grid separation (accurately done by the proposed method). The noise present in the separating area, however, does produce a displacement of the separating lines, but each box perfectly encloses the corresponding sub-grid. All the 20 images tested can be downloaded from the corresponding links given above.

To conclude the paper, the advantages of using the proposed method are summarized as follows. First, the proposed method allows to automatically detect angles of rotation (independently for the x and y axes), and performs a correction based on an affine transformation. Second, rotations are detected by mathematically sound principles involving the Radon transform and information-theoretic measures. Third, once the affine transformation is performed, the method allows to detect the sub-grids accurately, as shown in two sets of images from different sources and having different parameters (resolution,

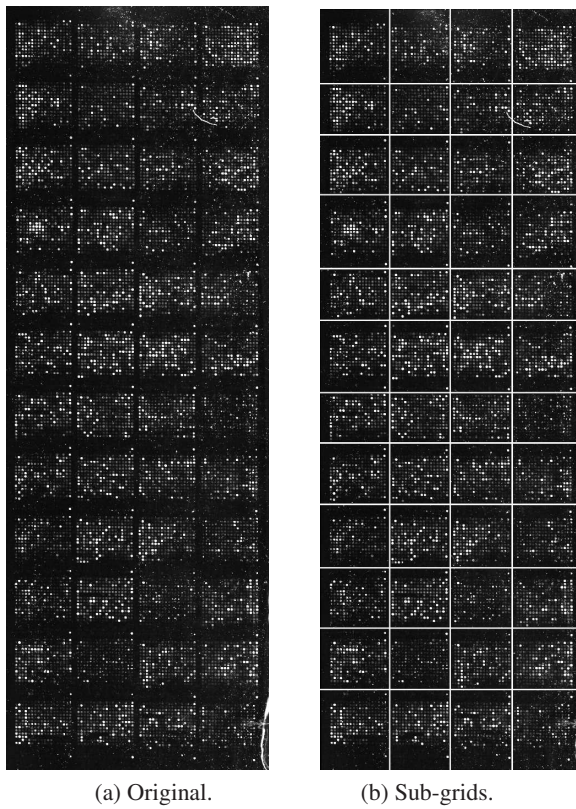


Fig. 4. Original and sub-grids detected by the proposed method, for images AT-20387-ch1 drawn from the SMD

Table 2. Test images drawn from the HTD, angles of rotation and percentage of sub-grids detected

Image	ϕ_{min_x}	ϕ_{min_y}	Accuracy
HT-Cy3-12663787	0.3	-0.1	100%
HT-Cy5-12663787	0.3	-0.1	100%
HT-Cy3-12667177	0.3	-0.2	100%
HT-Cy5-12667177	0.3	-0.2	100%
HT-Cy3-12667189	0.3	-0.1	100%
HT-Cy5-12667189	0.3	-0.1	100%
HT-Cy3-12667190	0.4	-0.2	100%
HT-Cy5-12667190	0.4	-0.2	100%
HT-Cy3-12684418	0.0	0.0	100%
HT-Cy5-12684418	0.0	0.0	100%

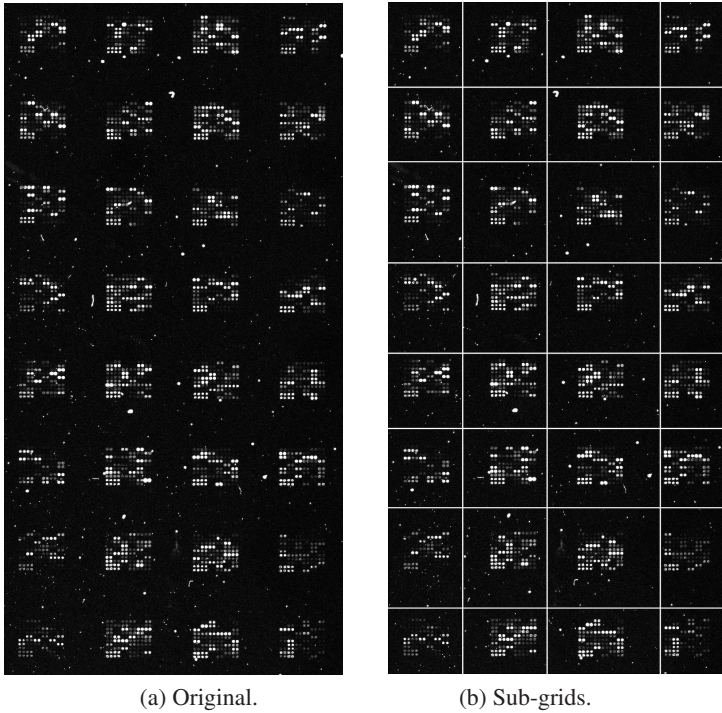


Fig. 5. Original and sub-grids detected for image HT-Cy5-12663787 drawn from the HTD

number of sub-grids, spot width, etc.). Fourth, the method provides the right orientation of the sub-grids detected so that they can be processed in the subsequent steps required to continue the microarray data analysis, namely detecting the spot centers (or gridding), and separating the background from foreground (segmentation).

4 Conclusions

A method for separating sub-grids in cDNA microarray images has been proposed. The method performs two main steps involving the Radon transform for detecting rotations wrt the x and y axes, and the use of morphological operators to detect the corresponding valleys that separate the sub-grids.

The proposed method has been tested on real-life, high-resolution microarray images drawn from two sources, the SMD and the HTD. The results show that (1) the rotations are effectively detected and corrected by affine transformations, and (2) the sub-grids are accurately detected in all cases, even in abnormal conditions such as extremely noisy areas present in the images.

Future work involves the use of nonlinear functions for the Radon transform, in order to detect curvilinear rotations. This is far from trivial as it involves a number of possible nonlinear functions, e.g. polynomials or exponentials. Another topic to investigate is

to fit each sub-grid into a perfect box eliminating any surrounding background, and hence providing advantages for the subsequent steps, a problem that is currently being undertaking.

References

1. Drăghici, S.: *Data Analysis Tools for DNA Microarrays*. Chapman & Hall (2003)
2. Schena, M.: *Microarray Analysis*. John Wiley & Sons, Chichester (2002)
3. Antoniol, G., Ceccarelli, M.: A markov random field approach to microarray image gridding. In: *Proceedings of the 17th International Conference on Pattern Recognition*, Cambridge, UK, pp. 550–553 (2004)
4. Brandle, N., Bischof, H., Lapp, H.: Robust dna microarray image analysis. *Machine Vision and Applications* 15, 11–28 (2003)
5. Ceccarelli, B., Antoniol, G.: A deformable grid-matching approach for microarray images. *IEEE Transactions on Image Processing* 15(10), 3178–3188 (2006)
6. Qi, F., Luo, Y., Hu, D.: Recognition of perspectively distorted planar grids. *Pattern Recognition Letters* 27(14), 1725–1731 (2006)
7. Rueda, L., Vidyadharan, V.: A hill-climbing approach for automatic gridding of cdna microarray images. *IEEE Transactions on Computational Biology and Bioinformatics* 3(1), 72–83 (2006)
8. Jain, A., Tokuyasu, T., Snijders, A., Segraves, R., Albertson, D., Pinkel, D.: Fully automatic quantification of microarray data. *Genome Research* 12(2), 325–332 (2002)
9. Noordmans, H., Smeulders, A.: Detection and characterization of isolated and overlapping spots. *Computer Vision and Image Understanding* 70(1), 23–35 (1998)
10. Katzer, M., Kummer, F., Sagerer, G.: A markov random field model of microarray gridding. In: *Proceedings of the 2003 ACM Symposium on Applied Computing*, pp. 72–77 (2003)
11. Angulo, J., Serra, J.: Automatic analysis of dna microarray images using mathematical morphology. *Bioinformatics* 19(5), 553–562 (2003)
12. Katzer, M., Kummert, F., Sagerer, G.: Automatische auswertung von mikroarraybildern. In: *Proceedings of Workshop Bildverarbeitung für die Medizin*, Cambridge, UK (2002)
13. Steinfath, M., Wruck, W., Seidel, H.: Automated image analysis for array hybridization experiments. *Bioinformatics* 17(7), 634–641 (2001)
14. Wang, Y., Ma, M., Zhang, K., Shih, F.: A hierarchical refinement algorithm for fully automatic gridding in spotted dna microarray image processing. *Information Sciences* 177(4), 1123–1135 (2007)
15. Wang, Y., Shih, F., Ma, M.: Precise gridding of microarray images by detecting and correcting rotations in subarrays. In: *Proceedings of the 8th Joint Conference on Information Sciences*, Salt Lake City, USA, pp. 1195–1198 (2005)
16. Helgason, S.: *The Radon Transform*, 2nd edn. Springer, Heidelberg (1999)
17. Qin, L., Rueda, L., Ali, A., Ngom, A.: Spot detection and image segmentation in dna microarray data. *Applied Bioinformatics* 4(1), 1–12 (2005)