Applied Medical Informatics
Vol. 37, No. 3 /2015, pp: 21-30

Automated cDNA Microarray Segmentation using Independent Component Analysis Algorithm

Elahe AMINIAN [1], Shayan YAZDANI[2], Hamidreza SABERKARI[3,*]

[1] Department of Electrical Engineering, Ahar Branch, Islamic Azad University, Tabriz, Iran
[2] Department of Electrical Engineering, Sahand University of Technology, Tabriz, Iran
[3] Department of Electrical Engineering, Rasht Branch, Islamic Azad University, Rasht, Iran
E-mails: elaheaminian@gmail.com; sh_yazdani@sut.ac.ir; h_saberkari@sut.ac.ir

* Author to whom correspondence should be addressed; Tel.: +98-911-2362692

Received: 22 June 2015 / Accepted: 30 August 2015 / Published online: 5 September 2015

Abstract
There is a most useful method in order to simultaneously study thousands of gene expression levels of a simple experiment, called DNA microarray. The average value of fluorescent intensity can be calculated through a microarray experiment. The calculated intensity values are close to the expression levels of a specific gene. Therefore, the appropriate determination of every spot position (specific gene) will lead to the accurate measurement of the intensity amount, and as a result, accurate classification of normal and abnormal gene expression levels within the microarray image. In this paper, first, a preprocessing step is performed in order to cancel the noise and artifacts in DNA microarray images using the nonlinear diffusion filtering method. Then, the coordinate center of each spot is determined utilizing mathematical morphology operations. Finally, pixel classification is performed using the independent component analysis (ICA) algorithm. Performance of the proposed algorithm has been assessed on the microarray images of the Stanford Microarray Database (SMD). Realization results illustrate that the classification accuracy of the proposed algorithm of the noisy microarray cells is close to 98%, while this amount is equal to 100% for noiseless cells.

Keywords: Gene expression; Microarray; Independent components; Noise

Introduction

The possibility for researchers to analyze simultaneously thousands of gene expression levels was provided through microarray technology discovery in 1995. This can help facilitate genetic diseases identification at a molecular level. Due to this huge change in microarray, this technology has been utilized in many medical applications in recent decades. Some of these applications are; (i) Cancer: determining the differences between normal and abnormal cells, classifying tumors, and identifying risk factors, (ii) Pharmaceutical: determining the relation between gene expression profiles and their response to various drugs, and (iii) Toxicology: determining the relation between response to various toxins and deviations made in different tissues of genetic profiles in facing different toxins.

There are thousands of individual DNA strands in cDNA microarray, which are printed by a robotic arrayer on the highly pure array (often on a glass slide). cDNA microarray raw image is stored in a 16 bit image in Tagged Image Format (TIFF) for both of the dyes. Images include various blocks (called sub-networks). These blocks contain various spots, which are located in columns and rows. Generally, there are four steps to analyze the microarray image as follows:

1. Preprocessing: This step is used to eliminate the background noise and artifacts,
2. Gridding: This step is used to determine the position of each spot and also the center of the coordinates is carried out in this step,
3. Segmentation: In this step, pixels in the microarray images are classified into foreground (spot) and background,
4. Determining the gene expression levels: In this step, assigned pixels to each spot are used to determine the gene expression levels.

Image segmentation is one of the most important steps in analyzing of these images. In recent decades, different software were introduced in literatures for segmentation. Some of these packages are known as: ScanAlyze [1], Dapple [2], ImaGene [3], SpotFinder [4]. The main problem of these softwares is that all the parameters should be set manually as well as the center of spot should be situated by human intervention, which has a negative influence on the analysis of gene expression levels. In order to overcome these drawbacks, different methods are proposed, as provided bellow [5]:

- **Shape-based segmentation methods:** Shape-based methods are based on the specific shape of spots. There are two conventional methods for this approach; fixed circle segmentation algorithm implemented by ScanAlyze [1] software, and adaptive circle segmentation algorithm presented by Buhler [2]. These two methods place a circular template on each spot. After this, Sarder et al. [6] proposed the parametric circular technique with the elliptic centers for segmentation of spots in noisy microarrays. The main problem of the shape-based methods is to not be able to segment the non-circular spots.
- **Shape-independent segmentation methods:** These methods relieve the main problem of shape-based methods. Seeded growing regions (SGR) algorithm used in spot software for segmentation of irregular spots is a common approach proposed in [7]. In this algorithm, a collection of seeds are considered for each cell in the first step. The similar neighborhood pixels are considered as spots in the iterative procedure. The main restriction of SRG algorithm is its highly sensitive performance on the suitable seed selection. In [8-10], two methods have been proposed in order to assign the pixels belonging to the spots and backgrounds, named k-means and the hybrid k-means method, respectively. These methods will show a poor performance if their spots have a low contrast. The other method, which is based on clustering the pixel values, utilizes elimination method in order to eliminate the non-connected clusters. In this approach, small clusters are considered as artifacts. However, the dimension of each cluster should be manually adjusted, which is the main problem of this method. In [11-15], active contour and multiple snake methods are proposed. But, both of them suffer from their inappropriate performance in the noisy images.

In this paper, an optimum algorithm is proposed for microarray cells segmentation based on independent component analysis (ICA) algorithm.

Material and Method

Dataset

In order to evaluate the performance of proposed algorithm in this paper, the Stanford Microarray Database (SMD) [16] is used. The sub-networks of the microarray contain 576 spots, which form an image with 24×24 rows and columns (the total pixels are equal to 112896). In addition, annotation images are extracted using a constant radius circle and it is used as a ground truth image. The binary versions are produced for all images. Inner/outer pixels of the binary images represent the signal pixels (spot)/ background, respectively.

Proposed Algorithm

An optimum algorithm for microarray images segmentation based on the Independent Component Analysis (ICA) approach has been proposed. First, a preprocessing step is applied for noise and artifacts elimination, which leads to an improved image quality. Then, mathematical morphology operations are applied in order to perform image segmentation. Finally, independent component analysis (ICA) algorithm is utilized for segmentation of each cell. Figure 1 shows the general block diagram of proposed algorithm for segmentation and it will be explained in details in this section.

A. Preprocessing

In this paper, a method based on a model has been proposed for background noise suppression in both of the red and green channels of microarray image. This model is known as nonlinear diffusion filtering method. This method has a physical inspiration, originated from mass and heat transfer rules, which is used to make a balance in concentration deviation between two environments. The spots and also the noise pixels of a microarray image can be modeled as a concentration and a little inhomogeneity in density, respectively. Noise inhomogeneity is smoothed by utilizing the diffusion law and this phenomenon leads to reduction of noise of the microarray image [17].

Originally, the Gaussian representation introduces a scale dimension by convolving the original image with a Gaussian mask. This is similar to solve the linear diffusion equation as below:

$$\begin{aligned} &I(t)=c\nabla^2 I \\ &I(t=0)=I_0 \\ &c>0\in \mathrm{R} \end{aligned} \qquad (1)$$

where c is a constant named diffusion coefficient and the symbol represents the vector differential operator.

Peroan and Malil [18] proposed anisotropic diffusion for adaptive smoothing in order to formulate the problem in terms of the non-linear heat equation. The major advantage of this approach is the edge preservation by introducing the coefficient function c(x). The basic relation of anisotropic diffusion is as below:

$$\frac{\partial I_t(x)}{\partial t}=div\{c(x)\nabla I_t(x)\} \qquad (2)$$

B. Gridding

After de-noising, the coordinate of each spot should be determined and then, a gridded image should be created. Some gridding techniques are proposed in the literature as [19-21]. In this paper, we used a useful method which is based on the projection of microarray image in length of rows and columns using the morphology reconstruction operations used for determining the spot location. Assume that a sub-network of cDNA microarray image is shown as $f=\{a_{xy}\}$ in which $x\in[1,h]$ and $y\in[1,w]$. Gridding procedures based on the morphology operation are as follows [26]:

1. Calculation of vertical and horizontal projection signals:

$$\begin{aligned} &H(y)=\sum_{x=0}^{w\ \ 1} f(x,y) \\ &V(x)=\sum_{y=0}^{h\ \ 1} f(x,y) \end{aligned} \qquad (3)$$

2. Filtering the horizontal projection signal using the morphological reconstruction operations:

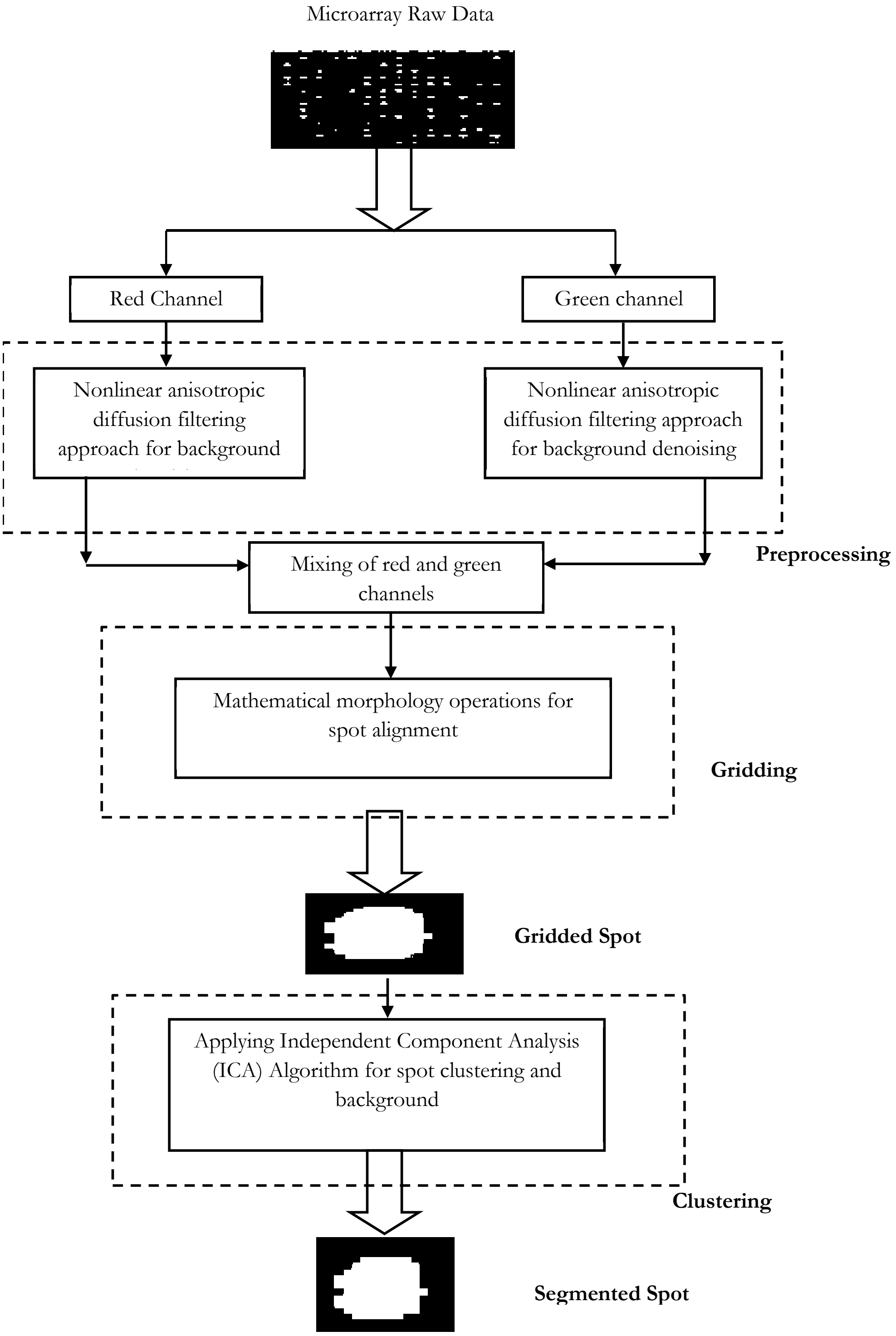
Microarray Raw Data
Red Channel
Green channel
Nonlinear anisotropic diffusion filtering approach for background
Nonlinear anisotropic diffusion filtering approach for background denoising
Preprocessing
Mixing of red and green channels
Mathematical morphology operations for spot alignment
Gridding
Gridded Spot
Applying Independent Component Analysis (ICA) Algorithm for spot clustering and background
Clustering
Segmented Spot

Figure 1. The block diagram of proposed algorithm

$$H^{rec}(i)=\gamma^{rec}\left(H(i),H_{\eta}(i)\right) \tag{4}$$

where

$$H_{\eta}(i)=H(i) \quad \overline{H}$$

$$\overline{H}=\frac{1}{w}\sum_{i=1}^{w}H(i) \tag{5}$$

And γ shows opening morphology operation, which is defined as:

$$\begin{aligned}&Opening: \gamma_B\left(f(x)\right)=\delta_B\left[\varepsilon_B\left(f\right)\right]\\&Erosion: \varepsilon_B\left(f(x)\right)=inf_{y\in B}\left\{f\left(x \quad y\right)\right\} \qquad f: D_f \rightarrow T\\&Dilation: \delta_B\left(f(x)\right)=sup_{y\in B}\left\{f\left(x \quad y\right)\right\}\end{aligned} \tag{6}$$

In Eq. (6), f is the amounts of gray level of image at point (x, y), D_f is a subset of Z^2 and T is a set of gray levels. It should be noted that the morphological reconstruction is implemented on the basis of restricting the iterative dilation of a function marker f by B (B is a subset of Z^2) to a function mask g, $\delta_g^n\left(f\right)=\delta_g^1\,\delta_g^{n \quad 1}\left(f\right)$, where $\delta_g^1\left(f\right)=\delta_B\left(f\right)\wedge g$.

3. Determining the residual signal amount:

$$H_r(i)=H(i) \quad H^{rec}(i) \tag{7}$$

4. Estimating the optimal value of threshold (t_H), which is defined as below:

$$t_H=\frac{1}{2}\cdot\frac{1}{w}\sum_{i=1}^{w}H_r(i) \tag{8}$$

5. Obtaining the square value of the binary signal by t_H and finding the border lines in the right and left sides of each interval.
6. Calculating the middle of each interval of binary signals and drawing the straight lines.

Vertical gridding uses the same method explained above for the horizontal gridding.

In Figure 2(a), a sample of microarray image is shown. Figures 2(b) and (c) are the red and green channels extracted from original image, respectively. Results of horizontal projection by applying mathematical morphology operations and also the gridded network of the microarray image are shown in Figures 2(d) and 2(f), respectively. As it can be seen from Figure 2 (g), the coordinate of each spot in every sub-network is calculated precisely.

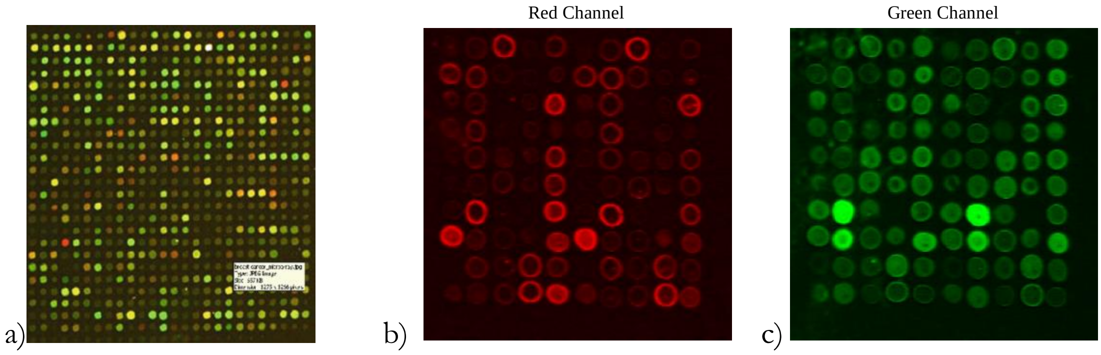

Figure 2. Results from applying the proposed math morphology method in microarray image segmentation. (a) Microarray images containing the green and red channels, (b) and (c) extracting the green and red channels of microarray image

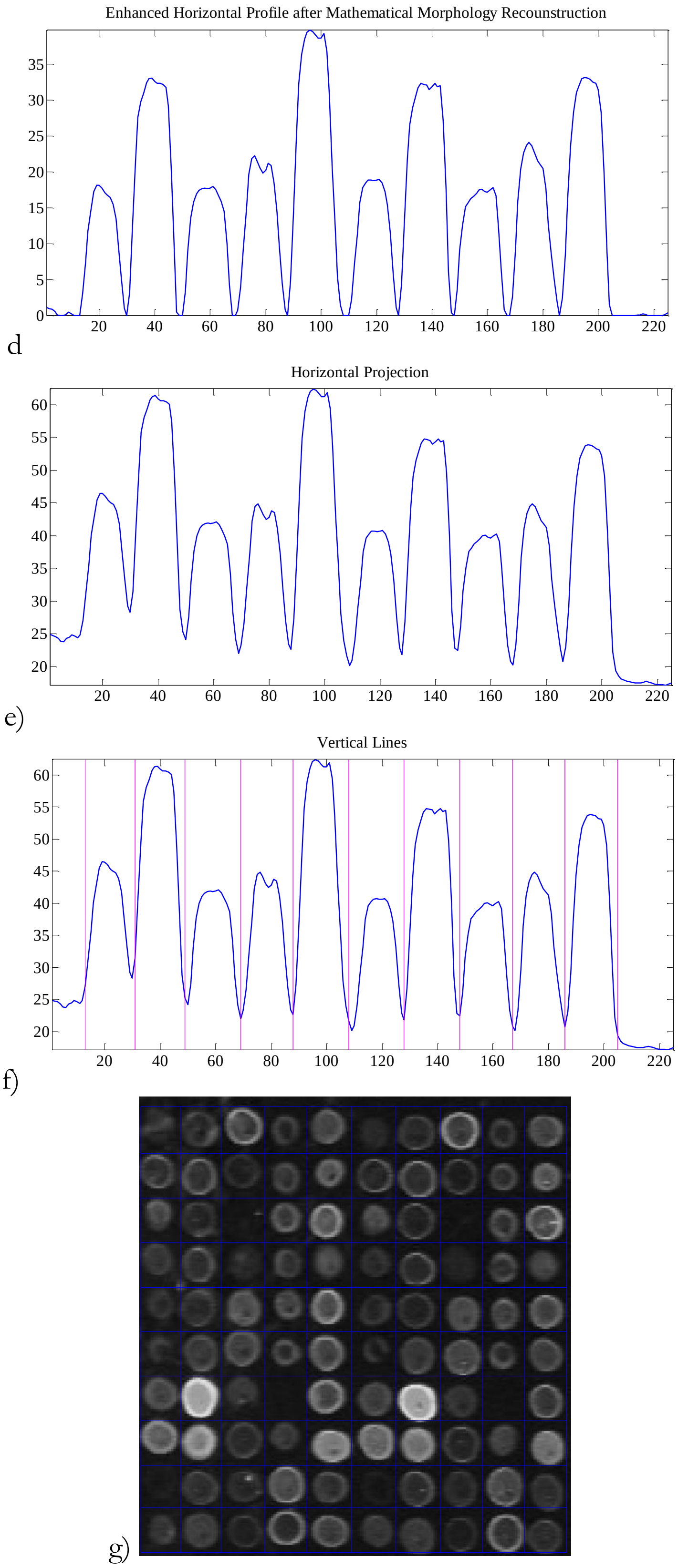

Figure 2. (d) extraction of a 1D horizontal projection signal, (e) Restoration of the horizontal projection signal using the mathematical morphology operations, (f) Drawing the vertical lines. (g). Final result of the microarray image segmentation.

C. Segmentation

The Independent Component Analysis (ICA) algorithm was used to segment each cell into foreground and background regions. Independent component analysis is a method to process signals, based on high order statistical information. It decomposes multipath signals into independent statistical components, source signals. ICA can be modeled perceiving below assumptions:

- Source signals are independent statistically,
- The number of source signals is lower than or equal to the number of observed signals, and
- The number of source signals with Gaussian distribution is 0 or 1, and Gaussian combinational signals are inseparable.

Perceiving upper assumptions ICA model for $X(t)$ is expressed as below:

$$X(t) = A * S(t) \tag{9}$$

where $X(t) = [X_1(t), X_2(t), \cdots, X_p(t)]^T$ is a data matrix with $p \times n$ dimensions, and its rows correspond with observed signals and its columns correspond with the number of samples. X represents the log ratios of red (experiment) and green (reference) intensities ($x_{ij} = log_2 \; {R_{ij}}/{G_{ij}}$). The mixing matrix $A = \lfloor a_1, a_2, \cdots, a_m \rfloor$ is combination matrix with $p \times m$ dimensions which contains the linear coefficients a_{ij} where a_{ij} is the activity level of process *j* in condition *i*. $S(t) = [S_1(t), S_2(t), \cdots, S_m(t)]^T$ is source signal matrix with $m \times n$ dimensions as its rows are independent statistically. Variables found in $S(t)$ rows are called ICs and $X(t)$ observed signals form a linear combination with these ICs. ICs estimation is made with finding linear relation of observed signals. In other words, with estimating a *W* matrix, satisfying the equation below, this objective can be reached.

$$S(t) = A^{-1} * X(t) = W * X(t) \tag{10}$$

The fast-ICA (FICA) algorithm was used to achieve IC components with equal variable number as the dimension of samples. Generally, when the number of source signals is equal to observation, reconstructed observed signals can contain comprehensive information. FICA is based on a fixed-point iteration scheme for finding a maximum of the non-Gaussianity of the sources. It is to be noted that this algorithm performs best if the distribution of the sources is “generalized normal distribution” which means that $p(x) = C \times e^{-\alpha|x|^{\gamma}}$, $\gamma \neq 2$. In this formula, $\gamma \neq 2$ gives Gaussian distribution.

Results and Discussion

In Figures 3, the results from segmentation of the proposed algorithm are shown. In this paper, Fuzzy C-Means (FCM) and Otsu algorithms have been implemented in order to compare with the performance of the proposed algorithm. Also, in order to evaluate the quantitative performance of the proposed algorithm in the segmentation of microarray image, the Segmentation Matching Factor is utilized. This parameter considers the pixels that have been segmented incorrectly and is defined as below [22]:

$$SMF = \frac{A_{segment} \cap A_{actual}}{A_{segment} \cup A_{actual}} \tag{11}$$

where $A_{segment}$ and A_{actual} are the binary versions of the real and segmented images, respectively. For SMF parameter we have:

- If SMF=100%, we have perfect matching of images.
- If SMF>50%, the result of segmentation is acceptable.
- If SMF<50%, the result of segmentation is weak.

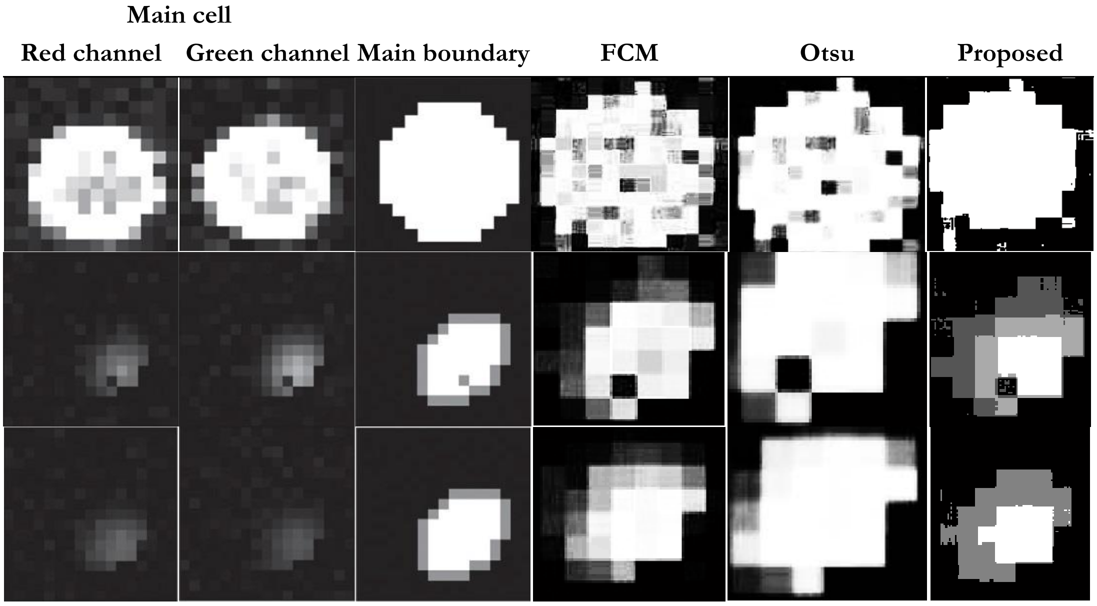

Figure 3. Comparison between results of proposed algorithm and FCM and Otsu methods in microarray cells

In order to evaluate the amount of stability of the proposed algorithm against the microarray image noise, the microarray images are corrupted with additive white Gaussian noise by the signal to noise (SNR) ratio of 1, 3, 5, 7 and 9 (dB). Tables 1 to 3 show the *SMF* quantities for the proposed algorithm and other segmentation methods for the cells shown in Figure 3. As it is obvious, the *SMF* of the proposed algorithm increases by increasing the amount signal to noise (SNR) ratio. Also, the proposed algorithm has a good stability against the noise.

Table 1. Qualitative comparison of SMF values between proposed method and other algorithms for cell 1 versus signal to noise ratio deviation

Signal to noise ratio (dB)	FCM method	Otsu method	Proposed method
1	89.2	85.4	**90.1**
3	89.9	87.3	**91.2**
5	91.2	89.5	**93.6**
7	93.1	90.6	**94.8**
9	94.6	91.4	**95.7**

In this paper, segmentation of noisy and noiseless cells is performed by extracting the intensity value of pixels of each spot in microarray image. First, the artifacts in microarray images were reduced using non-linear anisotropic diffusion method. Then, mathematical morphology operators were utilized to grid the microarray image. Finally, the intensities of each spot were accurately calculated using ICA algorithm.

Table 2. Qualitative comparison of SMF values between proposed method and other algorithms for cell 2 versus signal to noise ratio deviation

Signal to noise ratio (dB)	FCM method	Otsu method	Proposed method
1	78.2	68.0	**84.6**
3	76.5	70.6	**85.2**
5	79.1	72.4	**88.9**
7	85.3	75.9	**90.5**
9	88.9	79.1	**92.8**

Table 3. Qualitative comparison of SMF values between proposed method and other algorithms for cell 3 versus signal to noise ratio deviation

Signal to noise ratio (dB)	FCM method	Otsu method	Proposed method
1	72.4	64.8	**82.6**
3	73.6	67.5	**83.0**
5	75.8	69.5	**85.4**
7	79.4	70.9	**88.1**
9	80.9	72.4	**90.5**

Performance of the proposed algorithm was compared with the other methods based on a criteria measure and results have been proved the superiority of our algorithm. This method suffers from two problems. First, the noise, especially noise of the fluorescent, has a negative effect on this method. Secondly, threshold level is considered as the mean value of the filtered signal, which is not an optimum value for the threshold level. Our aims in the future will be choosing an adaptive threshold value using intelligent algorithms for gridding, and also combining the Gaussian kernels besides the spatial fuzzy clustering algorithm to reach a more improvement in the segmentation process.

Conflict of Interest

The authors declare that they have no conflict of interest.

References

1. Scanalyze [Online] [cited March 2015]. Available from: http://graphics.stanford.edu/software/scanalyze/.
2. Buhler J, Ideker T, Haynor D. Dapple: improved techniques for finding spots on DNA microarrays. UW CSE Technical Report UWTR 2000-08-05, 2000; pp. 1-12.
3. BioDiscovery Inc. (2005). ImaGene. [Online] [cited March 2015]. Available from: http://www.biodiscovery.com/software/imagene/
4. 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):548-562.
5. Yang YH, Buckley MJ, Dudoit S, Speed T. Comparison of methods for image analysis on cDNA microarray data. J Comput Graph Stat 2002;11(1):108-136.
6. Sarder P, Nehorai A, Davis PH, Stanley SL. Estimating gene signals from noisy images. IEEE Trans Nanobiosci 2008;7(2):142-153.
7. Beare R, Buckley MJ. (2000). The Spot User's Guide, CSIRO mathematical and information sciences. [Online] [cited March 2015] Available from: http://www.hca-

vision.com/Spot_Documentation/Spot_old.pdf.

8. Bozinov D, Rahnenfuhrer J. Unsupervised technique for robust target separation and analysis of DNA microarray spots through adaptive pixel clustering. Bioinformatics 2002;18(5):747-756.
9. Rahnenfuhrer J, Bozinov D. Hybrid clustering for microarray image analysis combining intensity and shape features. BMC Bioinformatics 2004;5(5):47-58.
10. Nagarajan R. Intensity-based segmentation of microarray images. IEEE Trans Med Imag 2003;22(7):882-889.
11. Nagarajan R, Peterson CA. Identifying spots in microarray images. IEEE Trans Nanobiosci 2002;1(2):78-84.
12. Li Q, Fraley C, Bumgarner RE, Yeung KY, Raftery AE. Donuts, scratches and blanks: Robust model-based segmentation of microarray images. Bioinformatics 2005;21(12):2875-2882.
13. Ho J, Hwang WL. Automatic Microarray Spot Segmentation using a snake-fisher model. IEEE Trans Med Imag 2008;27(6):847-857.
14. Srinark T, Kambhamettu C, Kambhamettu R. A framework for Multiple Snakes. Proc Comput Vision Pattern Recogn 2001:202-209.
15. Katzer M, Kummert F, Sagerer G. Methods for automatic microarray image segmentation. IEEE Trans Nanobiosci 2003;2(4):202-212.
16. Stanford Microarray Database. [Online] [cited March 2015]. Available from: http://smd.princeton.edu/.
17. Chua L. CNN: a vision of complexity. Int J Bifurc Chaos Appl Sci Eng 1997;7:1425-2219.
18. Perona P, Malik J. Scale-space and edge detection using anisotropic diffusion. Proceedings of IEEE Computer Society Workshop on Computer Vision, 1987. pp. 16–22.
19. Bariamis D, Iakovidis DK, Maroulis D. M^{3}G: Maximum Margin Microarray Gridding. BMC Bioinformatics 2010;11:49. doi:10.1186/1471-2105-11-49.
20. Rueda L, Rezaeian I. A fully automatic gridding method for cDNA microarray images. BMC Bioinformatics 2011;12:113.
21. Belean B, Terebes R, Bot A. Low-complexity PDE-based approach for automatic microarray image processing. Med Biol Eng Comput 2015;53(2):99-110.
22. Lehmussola A, Ruusuvuori P, Yli-Harja O. Evaluating the performance of microarray segmentation algorithms. Bioinformatics 2006;22(23):2910-2917.