30th Annual International IEEE EMBS Conference
Vancouver, British Columbia, Canada, August 20-24, 2008

A Classification-Based Segmentation of cDNA Microarray Images using Support Vector Machines

Nikolaos Giannakeas, *Student member, IEEE,* Petros S. Karvelis, *Student member, IEEE* and Dimitrios I. Fotiadis, *Senior Member, IEEE*

***Abstract*—Microarray technology provides a tool for the simultaneous analysis of the expression level for an amount of genes. Microarray studies have been shown that image processing techniques can significantly influence microarray data precision. In this paper we propose a supervised method for the segmentation of microarray images based on classification techniques. Support Vector machine is employed to classify each pixel of the image into signal, background or artefacts. In addition, a preprocessing step is applied in order to filter the initial image. The proposed method is applied both to real and simulated images. The pixels of the image are classified in two classes for the real images and three classes for the simulated one. For this task, an informative set of features is used from both green and red channels. The results obtained indicate high accuracy (~99%).**

I. Introduction

DNA microarray technology allows the comprehensive measurement of the expression level of many genes simultaneously on a common substrate [1]. Typical applications of microarrays include the quantification of RNA expression profiles of a system under different experimental conditions, or expression profile comparisons of two systems under one or several conditions.

During the biological experiment, the mRNA of two biological tissues of interest (i.e normal and tumour) is extracted. Each of the mRNA samples are reverse transcribed into complementary DNA (cDNA) copy and labelled with two different fluorescent dyes resulting in two fluorescence-tagged cDNA (red Cy5 and green Cy3). The tagged cDNA copies, called the sample probe, are hybridized to a slide containing an array of single-stranded cDNAs called probes. Probes are usually known genes of interest which were printed on a glass microscope slide by a robotic arrayer. According to the hybridization principles, a sample probe will only hybridize with its complementary probe. After the hybridization, the microarray is scanned twice, at different wavelengths corresponding to the different dyes used in the assay. The digital image scanner records the intensity level at each microarray location producing two greyscale images. The intensity level is correlated with the absolute amount of RNA in the original sample, and by extension, the expression level of the gene associated with this RNA.

The processing of microarray images [2] includes three stages: initially, spots and blocks are preliminarily located from the images with gridding. Second, using the available gridding information, each microarray spot is individually segmented into foreground and background. Finally, intensity extraction, calculates foreground fluorescence intensity and background intensities. Morphological [3] or Fuzzy vector filter [4] have been employed to enhance the raw image. In addition, a denoising step has been developed by Wang et al. [5] using Stationary Wavelet Transform.

The methods for microarray image segmentation can be divided into four categories: (i) Fixed and adaptive circle, (ii) histogram-based, (iii) adaptive shape, (iv) clustering. Fixed circle segmentation had been implemented by Eisen et al [6] and it is included in ScanAnalyze software package. Histogram-based techniques [7-8] estimate a threshold such that pixels with intensity lower than the calculated threshold are characterized as background pixels, whereas pixels with higher intensity as signal pixels. More sophisticated image processing techniques which have been used for the microarray image analysis comprise the adaptive shape segmentation. These methods do not include any assumption about the size and the shape of the spot. Algorithms such as Seed Region Growing [9] and Watershed Transform [10] have been employed in this category. Nowadays, clustering is the most common technique that is used for the segmentation of the microarray images. Clustering algorithms, such as K-means [11], Fuzzy C means [12], Expectation-Maximization [13] etc., have been used by several microarray imaging studies.

In this paper we introduce a novel segmentation method which classifies the pixels of the image into two categories (foreground and background) or three (signal, background and artefacts) using Support Vector Machines. Using clustering techniques different clusters are generated but have no distinction between them unless a set of rules is applied to separate them. We propose a classification-based approach which directly classifies each pixel to the designated category. Three classes of pixels are produced from the simulated images. Apart from signal and background pixels, a third class includes pixels of artefacts,

Manuscript received April 7, 2008.

N. Giannakeas is with Laboratory of Biological Chemistry, Medical School, University of Ioannina, Ioannina, Greece, GR 45110. (e-mail: me01310@cc.uoi.gr).

P.S. Karvelis is with the Unit of Medical Technology and Intelligent Information Systems, Dept. of Computer Science, University of Ioannina, Ioannina, Greece, GR 45110. (e-mail: pkarvel@cs.uoi.gr).

D.I. Fotiadis is with the Unit of Medical Technology and Intelligent Information Systems, Dept. of Computer Science, University of Ioannina, Ioannina, Greece, GR 45110 (0030-26510-98803; fax: 0030-26510-97092; e-mail: fotiadis@cs.uoi.gr).

pixels of the contour of the spot, and finally pixels of inner holes which are presented in donut spots [14]. To deal with this third class, a set of features of each pixel is used as input for the classification. In addition, microarray image is processed as a multichannel image using a set of multichannel filters in the preprocessing step. The classification step also processes both the image channels simultaneously.

II. Materials and Methods

The proposed method begins with the preprocessing step. The multichannel Fuzzy Vector Filters [4] are applied to the whole raw microarray image. Next, the preprocessed and the raw images are segmented using a classification-based technique. Support Vector Machine is employed for the pixel-by-pixel classification. The flowchart of the proposed method is shown in Fig. 1.

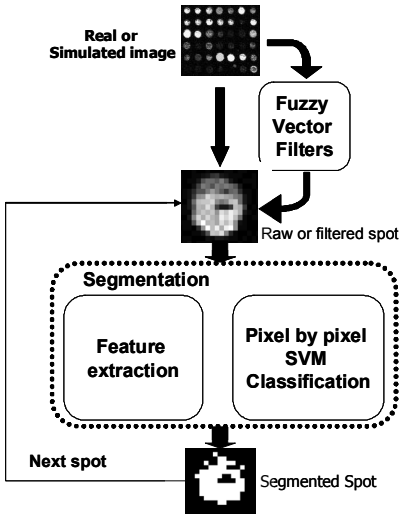

Fig.1: The flowchart of the proposed method.

A. Fuzzy Vector Filters

Due to the fact that microarray image is a dual channel image, multichannel filtering techniques must be applied. Multichannel image filtering, takes into account the correlation between the channels of the image, considering each pixel as a vector of components associated with the intensities of the channels. The output of these filters is defined as the lowest ranked vector according to a specific ordering technique.

Let us consider the dual channel microarray image $I(x,y) \in \mathbb{Z}^2$. In addition, we suppose a square filter window with a set of input multichannel samples such that $X = \{x_i : i = 1,2,\ldots,N\}$, $x_i \in \mathbb{Z}^2$ where N is an odd integer, which represents the size of the window. Suppose an input sample $x_i : 1 \le i \le N$. This sample is associated with the distance measures L_i and A_i defined as:

$$L_i = \sum_{j=1}^{N} \left\| x_i - x_j \right\|_\gamma, \tag{1}$$

$$\mathrm{A}_i = \sum_{j=1}^{N} \cos^{-1}\left(\frac{x_i x_j^T}{|x_i||x_j|} \right), \tag{2}$$

where $|x_i|$ is the magnitude of the vector x_i, and γ is the order of the employed norm, i.e. for $\gamma = 2$ the Euclidean distance is used.

The products of the distance measures can be used as the ordering criterion:

$$\Omega_i = L_i A_i = \sum_{j=1}^{N} \left\| x_i - x_j \right\|_\gamma \sum_{j=1}^{N} \cos^{-1}\left(\frac{x_i x_j^T}{|x_i||x_j|} \right), \quad 1 \le i \le N, \tag{3}$$

Then, the ordered set is given by $\Omega_1 \le \Omega_2 \le \ldots \le \Omega_N$. The same ordering scheme applied to the input set results in the ordered sequence, $x^{(\Omega_1)} \le x^{(\Omega_2)} \le \ldots \le x^{(\Omega_N)}$. The sample $x^{(\Omega_i)}$ associated with Ω_i represents the output of the Directional Distance Filter (DDF). Suppose the DDF with the power parameter p so that the power $1-p$ is associated with the sum of vector distances and the power $p \in [0,1]$ is associated with the sum of vector angles. Thus, Eq. (3) can be simply rewritten as:

$$\Omega_i = L_i^{1-p} A_i^p = \left(\sum_{j=1}^{N} \left\| x_i - x_j \right\|_\gamma \right)^{1-p} \left(\sum_{j=1}^{N} \cos^{-1}\left(\frac{x_i x_j^T}{|x_i||x_j|} \right) \right)^p, \tag{4}$$

According to the value of p there are several types of filters. The DDF operates as the Vector Median Filter (VMF) for $p = 0$, or as the Basic Vector Directional Filter (BVDF) when $p = 1$.

Finally, the Weighted Vector Median Filter (WVMF) could be defined through a set of weights. Let us denote of nonnegative integer weights as $w_1, w_2, \ldots, w_N$ so that each weight $w_j, 1 \le j \le N$ is associated to each input sample x_j. The weighted vector distance J_i is given as:

$$J_i = \sum_{j=1}^{N} w_j \left\| x_i - x_j \right\|_\gamma, \quad 1 \le i \le N. \tag{5}$$

TABLE I
FILTERING RESULTS

	Raw	VMF	WVMF	VBDF
Real				
Simulated Good				
Simulated Normal				

The sample $x^{(J_i)} \in \{x_1, x_2, \ldots, x_N\}$ associated with the minimal combined weighted distance J_1 is the sample which minimizes the sum of weighted vector distances and the output of the WVFM filter.

As it shown in Table I, the three filters perform well in a spot with artefacts and in inner holes of donuts spot. The background noise is eliminated too.

B. Segmentation

After the preprocessing of the image using the fuzzy vector filters, the features of each pixel are extracted and the segmentation technique is applied. More specifically, each pixel in the area of each spot is represented by a feature vector which contains the several features.

1) Feature extraction

The classification algorithm is fed with an informative set of 11 features [15]. These features can be categorized in three main categories: (i) intensity features which include intensity of the pixel, the mean value of the neighborhood of the pixel and the standard deviation of the neighborhood of the pixel, are used as intensity features. (ii) Spatial features, which are the coordinates of the pixel and the Euclidean distance between the position of the pixel and the center of the corresponding spot. (iii) Features related to the shape of the theoretical spot which contains the correlation of the pixel's 11x11 neighbourhood with an 11x11 2D-gaussian template. This is a measure of similarity between the neighborhood and the shape of a typical spot.

2) Support Vector Machines

Support Vector Machines (SVM) [16] composes a powerful learning system which simultaneously minimizes the empirical classification error and maximizes the geometric margin. Suppose a set of training vectors belonging to two separate classes, $\{(y_1, c_1), (y_2, c_2), \ldots, (y_k, c_k)\}$, where each y_i is the feature vector of each pixel of the image. c_i is either 0 (background) or 1 (signal), indicating the class

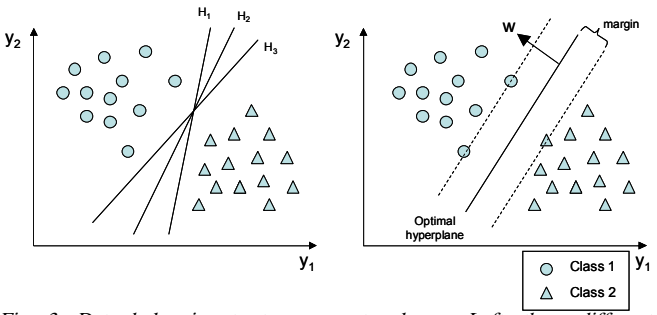

Fig. 3: Data belonging to two separate classes. Left: three different hyperplanes which divide the dataset. Right: the maximum-margin hyperplane

to which the point y_i belongs. The concept is to find a hyperplane which separates the data and can be written as:

$$wx - b = 0, \tag{6}$$

where vector w is the perpendicular vector to the hyperplane and the parameter b determines the offset of the hyperplane from the origin along the vector w. In our case SVMs produce the maximal-margin hyperplane which divides the background from the signal. Margin refers to the distance between the hyperplane and the nearest data point in each class.

Fig. 3(a) shows an example and several possible hyperplanes, but there is only one which maximizes the margin, as it shown in Fig. 3(b). The proposed approach employs the polynomial kernel. We set the c parameter equal to $c = 1$.

III. DATASET AND RESULTS

The proposed method is evaluated using both real and simulated microarray images. For both cases an annotation image, which includes pixel by pixel information for the whole image, is extracted

A. Real images

Real microarray images from the Stanford microarray Database SMD [17] are employed for the evaluation of the proposed approach. We have randomly chosen a block of the image "lc4b007rex2". The blocks of this image consists of 576 spots, forming 24x24 rows and columns (~112896 pixels). Given the annotation file of the SMD pixel by pixel information is extracted from the annotation, simulating the fixed circle segmentation. For this task, the known radius of the fixed circle and the coordinates of the centres of each spot are used. A binary map is generated for the whole block, characterizing the pixels inside the circle as signal pixels and the pixel outside of the circle as background.

Table II presents the classification accuracy results performed not only for the raw microarray image but also for the filtered images using the three types of Fuzzy Vector Filters.

TABLE II
CLASSIFICATION RESULTS FOR REAL IMAGES

Image	ACC %	Sens %	Spec %	Confusion Matrix (in pixels)	
				background	*signal*
Raw	99.81	98.18	99.96	53973	177
				21	53081
				background	*signal*
VMF	99.98	99.97	100	54134	16
				0	53102
				background	*signal*
WVMF	99.97	99.94	99.99	54120	30
				2	53100
				background	*signal*
BVDF	99.88	99.76	100	54025	125
				0	53102

B. Simulated Images

Apart from the real microarray images the proposed approach is evaluated using simulated images which are

produced by a simulation model [18]. To produce the simulated image we use the data from the annotation file of SMD of the real image lc4b007rex2. The mean intensities of each spot are given as an input in the model in order to provide a block, similar with the real one. The simulator can include a large number of parameters concerning the noise, the hybridization and scanner options etc. According to the values of these parameters the model categorizes the produced images into good, normal, and bad.

The annotation image is built during the simulation. When the model adds a spot or an artefact in the image we mark the same pixels in the annotation image as signal or artefacts, respectively. Then, we proceed to the filtering and classification-based segmentation of the image. Table III presents the accuracy results for good and normal images.

TABLE III
CLASSIFICATION RESULTS FOR SIMULATED IMAGES

Image	**ACC %**	**Confusion Matrix** (in pixels)		
		background	*signal*	*artefacts*
Good	97.30	81260	36	2149
		8	55026	170
		1680	215	17597
		background	*signal*	*artefacts*
Normal	87.51	110834	98	507
		2283	17804	2428
		8396	1269	8420

As we mentioned above, fuzzy vector filters perfectly effects in spots with artefacts and donuts spots. Thus, it is no sense to perform accuracy experiments for the third class in a filtered image. However, the classification results for the background and the foreground classes can be computed (Table IV).

TABLE IV
CLASSIFICATION RESULTS FOR FILTERED SIMULATED IMAGES

	Good			Normal		
	Acc %	Sens %	Spec %	Acc %	Sens %	Spec %
Raw	99.92	99.94	99.89	98.42	99.78	91.63
VMF	99.69	99.73	99.64	99.01	99.60	96.11
WVMF	99.89	99.92	99.85	98.80	99.75	94.14
VBDF	98.54	98.96	97.92	97.42	99.24	88.43

IV. DISCUSSION

In this paper, a supervised classification-based method for the segmentation of microarray images is presented. It is a multichannel approach consisting of two steps, the pre-processing step where fuzzy vector filters are applied to the initial image, and the classification step where all the pixels of the image are classified into signal, background or artefacts.

According to the results presented in Tables II-IV the proposed method detects efficiently the signal and the background pixels. Both for real and simulated images the results for these two classes are high. For the normal simulated images a large number of artefact pixels (third class) are classified as background pixels. This could be explained because the simulator adds several low-intensity artefacts i.e. bubbles and scratches. Another point that we have to address is the effectiveness of Fuzzy Vector Filters. Apart from the case of good simulated images, in which the quality of the image is extremely high, the accuracy results are optimised by the use of the filters. Unfortunately, none of the already developed segmentation methods present pixel by pixel accuracy results. Thus, a direct comparison of our method with existing methods is not possible.

V. REFERENCES

[1] M.B. Eisen, and P.O. Brown, "DNA Arrays for Analysis of Gene Expression," *Methods Enzymol,* vol. 303, pp.179-205, 1999.

[2] M. Schena, D. Shalon, R.W. Davis, and P.O. Brown, "Quantitative motoring of gene expression patterns with a complementary DNA microarray," *Science*, vol. 270, pp. 467-470, 1995.

[3] A. Bengtsson and H. Bengtsson, "Microarray image analysis: background estimation using quantile and morphological filters," *BMC Bioinformatics*, vol. 7, pp.:96, 2006

[4] R. Lukac, K.N. Plataniotis, B. Smolka, and A.N. Venetsanopoulos, "cDNA microarray image processing using fuzzy vector filtering framework," *Fuzzy Sets and Systems*, vol. 152 (1), pp. 17-35, 2005.

[5] X.H. Wang, R.S.H. Istepanian, and H.S. Yong, "Microarray image enhancement by denoising using stationary wavelet transform," *IEEE Transactions on NanoBioscience*, vol. 2(4), pp. 184-189, 2003

[6] M.B. Eisen, ScanAlyse, http://rana.Stanford.EDU/software/, 1999.

[7] QuantArray Analysis Software, http://lifesciences.perkinelmer.com.

[8] Y. Chen, E.R. Dougherty, and M.L. Bittner, "Ratio-Based Decisions and the Quantitative Analysis of cDNA Microarray Images," *Journal Of Biomedical Optics* vol.2(4), pp.364–374, 1997.

[9] M.J. Buckley, Spot User's Guide, *CSIRO Mathematical and Information Sciences*, Sydney, Australia, 2000.

[10] K.I. Siddiqui, A. Hero, and M. Siddiqui, "Mathematical Morphology applied to Spot Segmentation and Quantification of Gene Microarray Images," *in Proc. of Asilomar Conf. on Signals and Systems*, 2002.

[11] D. Bozinov, and J. Rahnenfuhrer, "Unsupervised technique for robust target separation and analysis of DNA microarray spots through adaptive pixel clustering," *Bioinform,,* vol. 18, pp. 747–756, 2002.

[12] E. Ergüt, Y. Yardimci, E. Mumcuoglu, O. Konu, "Analysis of microarray imagesusing FCM and K-means clustering algorithm," in Proc IJCI, pp.116-121, 2003.

[13] K. Blekas, N. Galatsanos, A. Likas, and I.E. Lagaris, "Mixture Model Analysis of DNA Microarray Images," *IEEE Transactions on Medical Imaging,* vol. 24(7), pp. 901-909, 2005.

[14] Q. Li, C. Fraley, R.E. Bumgarner, K.Y. Yeung, and A.E. Raftery, "Donuts, scratches and blanks: robust model-based segmentation of microarray images," *Bioinformatics*, vol.21, pp.2875–2882, 2005.

[15] N. Giannakeas, and D.I. Fotiadis, "Multichannel Segmentation of cDNA Microarray Images using the Bayes Classifier," *29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2007.

[16] C.J.C. Burges, "A Tutorial on Support Vector Machines for Pattern Recognition," *D. M. and Kn. Discovery* vol. 2, pp. 121-167, 1998.

[17] J. Gollub, C. A. Ball, G. Binkley, K. Demeter, D. B. Finkelstein, J. M. Hebert, T. Hernandez-Boussard, H. Jin, M. Kaplper, J. C. Matese, M. Schroeder, P. O. Brown, D. Botstein, and G. Sherlock, "The Stanford Microarray Database: data access and quality assessment tools," *Nucleic Acids Res*., vol. 31, pp. 94-96, 2003.

[18] M. Nykter, T. Aho, M. Ahdesmäki, P. Ruusuvuori, A. Lehmussola and O. Yli-Harja, "Simulation of microarray data with realistic characteristics." *BMC Bioinformatics*, vol. 7, 2006.