Segmentation of complementary DNA microarray images using the Fuzzy Gaussian Mixture Model technique

Emmanouil I. Athanasiadis, Dionisis A. Cavouras, Panagiota P. Spyridonos, Dimitris Th. Glotsos, Ioannis K. Kalatzis, and George C. Nikiforidis

***Abstract*—The objective of this work was to investigate the segmentation ability of the Fuzzy Gaussian Mixture Models (FGMM) clustering algorithm, applied on complementary DNA (cDNA) images. A Simulated Microarray image of 200 cells, each containing one spot, was produced following standard established procedure. An automatic gridding process was developed and applied on the microarray image for the task of locating spot borders and surrounding background in each cell. The FGMM and the Gaussian Mixture Model (GMM) algorithms were applied to each cell, with the purpose of discriminating foreground from background. The segmentation abilities of both algorithms were evaluated by means of the segmentation matching factor in respect to the actual classes (foreground-background pixels) of the simulated spots. The FGMM was found to perform better and with equal processing time, as compared to the GMM, rendering the FGMM algorithm an efficient alternative for segmenting cDNA microarray images.**

I. Introduction

Microarray imaging is used for the concurrent identification of thousands of genes in the field of bioinformatics [1]. By finding the location of the spots in a complementary DNA (cDNA) microarray experiment, calculations of the mean fluorescence intensity value are obtained, that are closely related to the expression of a specific gene. Thus, a precise localization and outlining of a spot are essential to obtain a more accurate intensity measurement, leading to a more precise expression measurement of a gene.

For the task of measuring spot intensity values, three major steps were followed [1]-[2]: 1/ the gridding step, for the precise localization of the cells, 2/ the segmentation step, for distinguishing each cell's foreground from background and 3/ the intensity extraction step, for calculating the mean fluorescence value of each spot.

In the past, several techniques and software packets have been developed for the task of processing microarray images [3]-[7]. In the ScanAlyze [3] software, a fixed circle segmentation method is used, where all spots are considered to be circular with a fixed predefined radius. In the GenePix [4] software, an adaptive circle segmentation technique is employed. According to that method, the radius of each spot is not constant but adapts to each spot separately. In the Spot [6] software, an adaptive shape segmentation technique is performed. In the latter technique, the most representative algorithms employed are the watershed [8] and the seeded region growing [9]. In the ImaGene [7] software, a histogram based segmentation method is applied, in which the values between the 80th and the 95th histogram percentile contribute to the calculation of the mean intensity value. Nevertheless, in all those techniques, the major disadvantages are either that spots are considered to be circular in shape or a-priori knowledge of the precise position of the spots' centers is a pre-requisite [10].

The Fuzzy Gaussian Mixture Model (FGMM) clustering algorithm is an effective clustering technique that has found application in many areas of pattern recognition, such as in voice recognition [11]-[12], but it has not been applied so far in microarray images.

In the present study, the FGMM clustering algorithm was developed for processing microarray images with purpose to investigate the segmentation ability of the algorithm. Additionally, a comparison of the FGMM algorithm with the Gaussian Mixture Model (GMM) algorithm, which has been shown to be effective in microarrays segmentation [13], was performed. Both methods, FGMM and GMM, were developed in MATLAB® [23]. Evaluation was performed by calculating the segmentation matching factors [14] of each algorithm in respect to the actual classes (foreground-background pixels) of the simulated spots. Additionally, the spots' mean intensity values were also calculated, from the segmented images, and were compared with the actual mean intensities values of the simulated spots.

Manuscript received June 30, 2006. E. I. Athanasiadis thanks the Greek State Scholarships Foundation (I.K.Y.), under Grand 4690/2005-2006, for funded the above work.

E. I. Athanasiadis, P. P. Spyridonos, D. Th. Glotsos and G. C. Nikiforidis are with the Medical Image Processing and Analysis (*M.I.P.A.*) Group, Laboratory of Medical Physics, School of Medical Science, University of Patras, 26500 Rion - Patras, Greece. (phone: 0030-2610-997745 e-mail: mathan@upatras.gr).

D. A. Cavouras and I. K. Kalatzis are with Medical Image and Signal Processing (*MED.I.S.P.*) Laboratory, Department of Medical Instruments Technology, Technological Institute of Athens, Ag. Spyridonos Street, Aigaleo, 122 10, Athens, Greece. (phone: 0030-210-5385375; fax: 0030-210-5910975; e-mail: cavouras@teiath.gr.).

II. Methods And Material

A. Automatic Gridding Process

A cDNA image consists of an arrayer, several sub-arrays, and thousands of spots corresponding to specific genes (see Fig. 1). Gridding is the procedure of segmenting each sub-array into numerous cells, each cell containing one spot and its background. Gridding renders the procedure of spot finding easier, since segmentation may be applied within each individual cell automatically. The gridding algorithm adopted in the present study consisted of the following steps [13].

1/Determination of sub-array regions. First average intensities were calculated along image rows and columns for both Red (‘R’) and Green (‘G’) channels (R for Cyanine Cy3 and G for Cyanine Cy5), thus forming two signals for each channel. Second, noise suppression was performed by means of a low pass filtering mask [13]. Third, sub-array regions were determined by finding the local minima of either the R or the G signals in the horizontal and vertical axes [15]-[16]; multiple experimentation showed that by choosing either R, or G channels, no significant differences were observed in sub-array boundary localization.

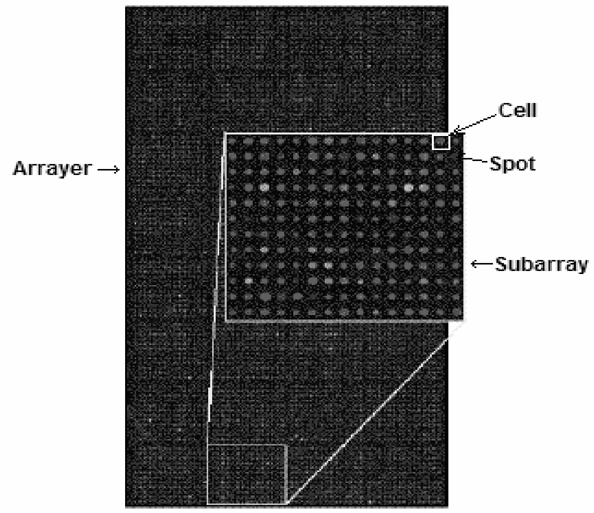
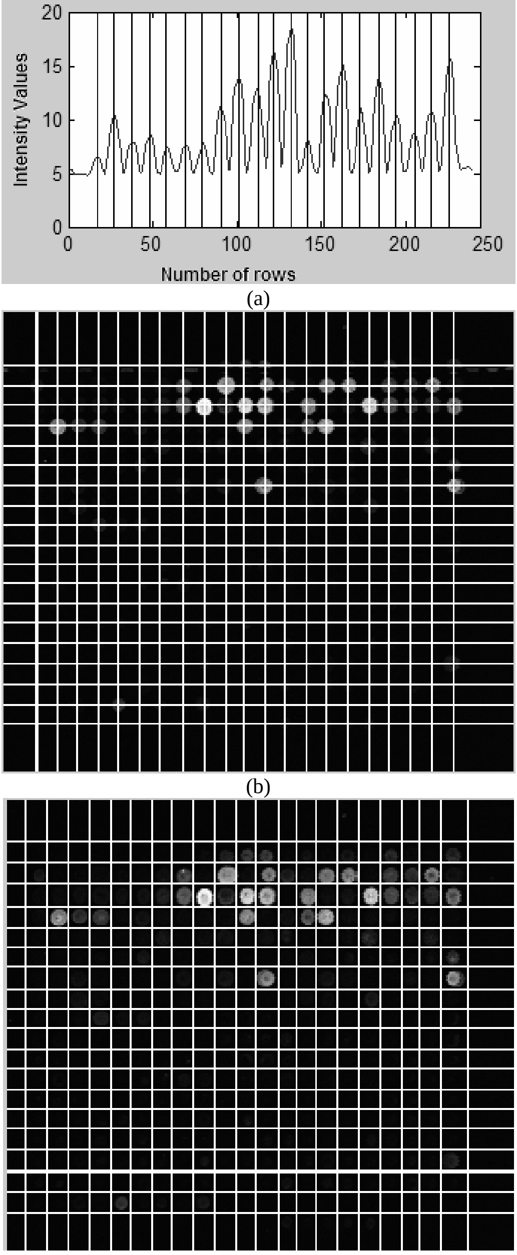


Fig. 1: An Arrayer consists of 8x4 sub-arrays, each sub-array of 19x21 cells and spots respectively.

2/Determination of cells. A similar procedure to sub-array determination was followed for identifying automatically the centers of the spots (local maxima) and the boundaries of the cells (local minima), employing an algorithm based on regional connectivity properties of pixels. Figure 2 demonstrates the automatic localization of the maximum signal values of an R-signal, as well as the result of the gridding step, applied to a sub-array region of a microarray image that contains 19x21 spots.



(a)

(b)

(c)

Fig. 2: (a) and (b), automatic localization of spot centers (local maxima using Matlab’s ‘imregionalmax’) and (c) cell determination (local minima using Matlab’s ‘imregionalmin’), applied on a real cDNA 19x21 sub-array microarray image collected from the Davidson College [21].

B. Gaussian Mixture Models

Let $X=\{x_1, x_2,..., x_T\}$ be a sequence of T vectors with x_i intensity values, and $\theta = \{p_i, \mu_i, C_i\}$ for $i = 1,..,c$, be a set of parameters to be maximized, using the Expectation Maximization technique [13], where parameters p_i, μ_i and C_i correspond to mixture weights, mean vectors, and covariance matrixes respectively of a mixture of c Gaussian Distributions.

The major task of the Gaussian Mixture Model [13]-[15] algorithm is to maximize the likelihood function that is described by equation (1):

$$p(X \mid \theta) = \prod_{t=1}^{T} p(x_t \mid \theta) \tag{1}$$

where the likelihood function of each vector $p(x_t| \theta)$ is computed by equations (2) and (3) as a probability density function of multiple Gaussians:

$$p(x_t \mid \theta) = \sum_{i=1}^{c} p(x_t, i \mid \theta) = \sum_{i=1}^{c} p_i p(x_t \mid i, \theta) \tag{2}$$

$$p(x_t \mid i, \theta) = \frac{\exp\left\{-\frac{1}{2}(x_t - \mu_i)' C_i^{-1} (x_t - \mu_i)\right\}}{(2\pi)^{d/2} \mid C_i \mid^{1/2}} \tag{3}$$

where $(x_t$-$\mu_i)'$ is the transpose matrix of $(x_t$-$\mu_i)$, C_i^{-1} is the inverse matrix and $|C_i|$ is the determinant of the covariance matrix C, for each class i, and d is the number of features used.

For the maximization of $p(X_t| \theta)$, an auxiliary function Q, named expected log-likelihood function [22], may be estimated employing equation (4):

$$Q(\theta, \bar{\theta}) = \sum_{t=1}^{T} p(i \mid x_t, \theta) \log[\overline{p_i} p(x_t \mid i, \bar{\theta})] \tag{4}$$

where $\bar{\theta} = \{\overline{p_i}, \overline{\mu_i}, \overline{C_i}\}$ are computed by using equations (5)-(7) (Maximization Step).

$$\overline{p_i} = \frac{1}{T} \sum_{t=1}^{T} p(i \mid x_t, \theta) \tag{5}$$

$$\overline{\mu_i} = \frac{\sum_{t=1}^{T} p(i \mid x_t, \theta) x_t}{\sum_{t=1}^{T} p(i \mid x_t, \theta)} \tag{6}$$

$$\overline{C_i} = \frac{\sum_{t=1}^{T} p(i \mid x_t, \theta)(x_t - \mu_i)(x_t - \mu_i)'}{\sum_{t=1}^{T} p(i \mid x_t, \theta)} \tag{7}$$

The posterior probabilities $p(i|x_t, \theta)$ are then recomputed by using equation (8) (Estimation Step).

$$p(i \mid x_t, \theta) = \frac{\overline{p_i} p(x_t \mid i, \bar{\theta})}{\sum_{k=1}^{c} \overline{p_k} p(x_t \mid i, \bar{\theta})} \tag{8}$$

Equation (4) is then calculated, and the process is repeated until there is no significant change in Q.

C. Fuzzy Gaussian Mixture Models

In the Fuzzy Gaussian Mixture Models FGMM [11]-[12] algorithm, a modification of the Fuzzy C Means FCM clustering technique [24] is performed for the estimation of parameters θ of GMM. In the FCM algorithm, the objective is to minimize equation (9).

$$J(U, \theta) = \sum_{t=1}^{T} \sum_{i=1}^{c} u_{it}^{m} d_{it}^{2} \tag{9}$$

where $U=\{u_{it}\}$ is a fuzzy c-partition of the initial vector X, u_{it} is the probability of vector x_i of belonging to class i, m is the fuzzy membership of u_{it}, called the degree of fuzziness, and d is a dissimilarity measurement [17], defined by equation (10).

$$d_{it}^{2} = -\log p(x_t, i \mid \bar{\theta}) = -\log \overline{p_i} p(x_t \mid i, \bar{\theta}) \tag{10}$$

Substituting equation (10) to (9), equation (11) is derived:

$$J(U, \bar{\theta}) = -\sum_{t=1}^{T} \sum_{i=1}^{c} u_{it}^{m} \log \overline{p_i} p(x_t \mid i, \bar{\theta}) \tag{11}$$

Minimization of equation (11) may be accomplished by using Lagrange multiplier methods [18]. Calculation of the new fuzzy parameters is achieved by using equations (12) – (14) (Minimization Step).

$$\overline{p_i} = \frac{\sum_{t=1}^{T} u_{it}^{m}}{\sum_{i=1}^{c} \sum_{t=1}^{T} u_{it}^{m}} \tag{12}$$

$$\overline{\mu_i} = \frac{\sum_{t=1}^{T} u_{it}^{m} x_t}{\sum_{t=1}^{T} u_{it}^{m}} \tag{13}$$

$$\overline{C_i} = \frac{\sum_{t=1}^{T} u_{it}^{m} (x_t - \mu_i)(x_t - \mu_i)'}{\sum_{t=1}^{T} u_{it}^{m}} \tag{14}$$

After the calculation of the new parameters, matrix U is recomputed according to equation (15) (Estimation Step).

$$u_{it} = \left[\sum_{k=1}^{c} \left(\frac{d_{it}}{d_{kt}} \right)^{\frac{2}{m-1}} \right]^{-1} \tag{15}$$

Equation (11) is then calculated, and the process is repeated until there is no significant change in J.

D. Segmentation Process

For every cell determined by the gridding process, the following procedure was performed:

1/ The NxM cell image, considering R and G channels separately, was converted into a vector X, with dimensions $1xNM$. A typical example of the conversion of a $3x3$ image cell into 1-dimensional vector is illustrated in Fig.3.

$$\begin{bmatrix} 1 & 2 & 1 \\ 3 & 5 & 3 \\ 2 & 3 & 2 \end{bmatrix} \longrightarrow \begin{bmatrix} 1 & 2 & 1 & 3 & 5 & 3 & 2 & 3 & 2 \end{bmatrix}$$

Fig. 3. Conversion of an initial cell (3x3 dimensions) into a vector (1x9 dimensions). The numbers inside the figure indicate random intensity values.

2/ The clustering algorithms (GMM and FGMM) were separately applied to the vector *X*, in order to discriminate the data into two categories or classes (*c=2*), the foreground (FG) and background (BG) class, denoted by zeros and ones, respectively. Next, the binary vector was transformed into a binary cell, following the inverse procedure, as illustrated in Figure 4.

$$\begin{bmatrix} 1 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 1 \end{bmatrix} \longrightarrow \begin{bmatrix} 1 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 1 \end{bmatrix}$$

Fig. 4. Conversion of a binary vector (1x9 dimensions) into a binary cell image (3x3 dimensions). Zeros indicate foreground and ones indicate background, as these values were the outcome of a clustering algorithm.

E. Intensity Extraction

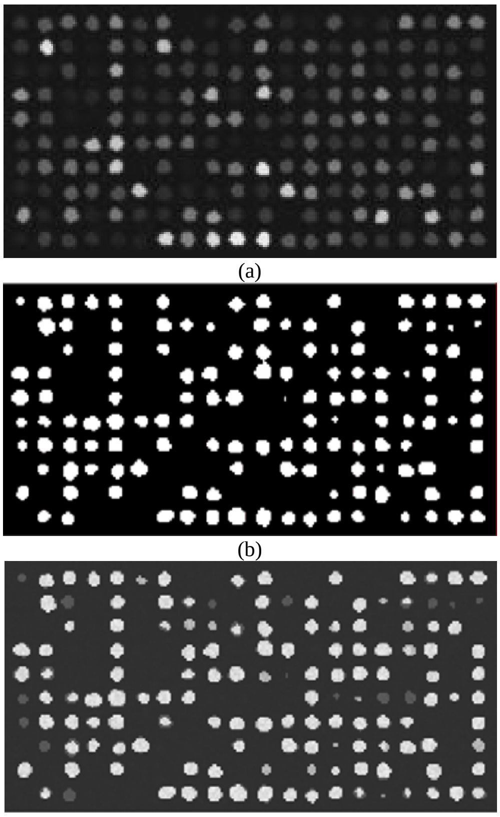
Representative spot intensity was obtained by subtracting the mean of the FG from the BG, according to equation (16).

$$I = \mu_{FG} - \mu_{BG} \tag{16}$$

Where μ_{FG} and μ_{BG} are the mean foreground and mean background respectively, both calculated from the corresponding labeled cell pixels.

F. Material

For the numerical evaluation of the clustering ability of the two techniques, a simulated cDNA image was produced as described in literature [19]-[20]. In order to generate spots with realistic characteristics, the following procedure was followed. A true cDNA image, consisting of 200 spots, was used as a template, and its binary version was produced by employing a thresholding technique [19] (see Fig. 5). In the simulated image, the location as well as the area of each spot was a-priory known. The mean intensity value of each spot was pre-defined, ranging between 0 and 255 for both the R and G channels [19]. Spot intensities were produced using an exponential distribution with mean value the pre-defined mean intensity value (using Matlab's 'exprnd' and 'expfit' functions). Background intensities were drawn from a single exponential distribution, with mean value determined from the true cDNA image's mean intensity background [19]. Figure 5c shows the simulated cDNA image.



(a)

(b)

(c)

Fig. 5: (a) Original real cDNA microarray image [21] and (b) the binary image used as a template to produce the simulated cDNA microarray image, (c) the simulated image.

The accuracy of segmentation was numerically calculated using the segmentation matching factor (equation (17)) [14] for every binary cell, produced by the clustering algorithm.

$$Accuracy = \frac{A_{cal} \cap A_{real}}{A_{cal} \cup A_{real}} \tag{17}$$

where A_{cal} is the area of the spot, as determined by the proposed algorithm, and A_{real} is the actual spot area. A perfect match is indicated by a 100% score, any score higher than 50% indicates reasonable segmentation [14] whereas, a score of less than 50% indicates poor segmentation [14].

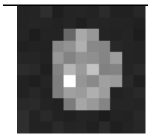
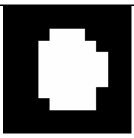
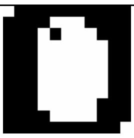
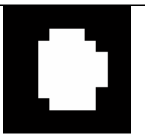
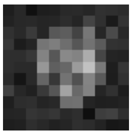
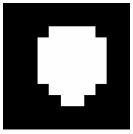
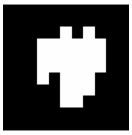
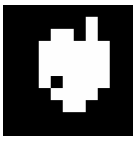
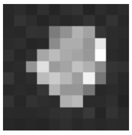
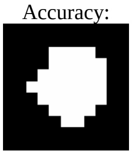
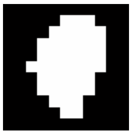
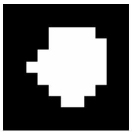
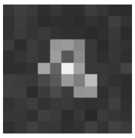
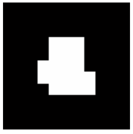
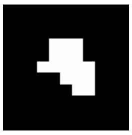
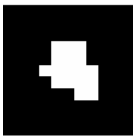
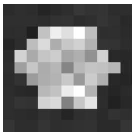
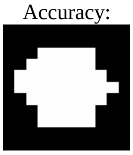
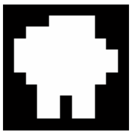
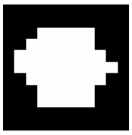
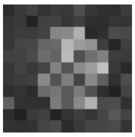
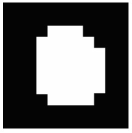
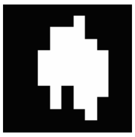
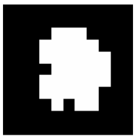
III. Experimental Results and Discussion

The gridding procedure was first performed and for every cell produced, the GMM and FGMM algorithms were applied identifying two classes, the foreground and background pixels. Thus, a set of two binary images, one for each clustering algorithm, were produced. The segmentation matching factor (equation (17)) was then calculated for each binary image, in order to estimate the accuracy of each method quantitatively.

According to our results, FGMM was more accurate than the conventional GMM clustering algorithm. Results of 6 different cells of the Green channel with the corresponding segmentation matching factors are illustrated in Table I. It should be noted that the number of Gaussians was set to c=2 in both clustering algorithms and that the degree of fuzziness in FGMM was set to m=1.5.

TABLE I

SEGMENTATION RESULTS FOR 6 DIFFERENT CELLS

Original Cells	Actual Boundaries	GMM Result	FGMM Result
			
Cell 1	Accuracy:	87.60%	100.00%
			
Cell 2	Accuracy:	95.87%	97.52%
			
Cell 3	Accuracy:	91.74%	100.00%
			
Cell 4	Accuracy:	95.87%	96.69%
			
Cell 5	Accuracy:	89.26%	100.00%
			
cell 6	Accuracy:	95.04%	98.35%

Comparative results for 6 different cells obtained from the G channel of the simulated microarray. The first column indicates the simulated spot with the surrounding area, the second column indicates the actual boundaries of the spot and the third and the forth columns present the segmentation results of the GMM and FGMM algorithms as well as the corresponding matching factors.

Additionally, the overall accuracy for all simulated spots for both Red and Green channels was calculated. The total number of pixels was 26448 (228 x 116 pixels) for each channel. The standard deviation of matched pixels for each cell was determined for both the FGMM and the GMM. The results are illustrated in Table II.

It is clear that the best overall accuracy for matching pixels (95.04%) was accomplished by the FGMM algorithm. According to Table II, the standard deviation of matched pixels was the lowest (3.63), in the case of FGMM algorithm, rendering it more robust than the conventional GMM technique.

TABLE II

OVERALL SEGMENTATION ACCURACY AND STANDARD DEVIATION CALCULATIONS

	GMM		FGMM	
	Match	Mismatch	Match	Mismatch
Overall Accuracy	93.88%	6.12%	95.04%	4.96%
Standard Deviation	7.61		3.63	

The Overall accuracy and standard deviation for matched and mismatched pixels achieved by each of the two different techniques in both R and G channels for all 200 spots (52896 pixels total).

Intensity values for each cell were also calculated by using equation (16). The intensity values of the 6 cells, shown in Table I, were calculated and are illustrated in Table III. Moreover, percentage differences between the actual and the calculated by GMM and FGMM intensity values are also presented in Table III.

TABLE III

INTENSITY VALUE CALCULATIONS FOR 6 DIFFERENT SPOTS

Cells	I_1 (Actual)	I_2 (GMM)	I_3 (FGMM)	$(\|I_1-I_2\|/I_1)$ x100	$\|(\|I_1-I_3\|/I_1)$ x100
Cell 1	114	85	114	25.44	0.00
Cell 2	72	76	74	5.56	2.78
Cell 3	128	101	128	21.10	0.00
Cell 4	122	92	110	24.60	9.84
Cell 5	146	105	146	28.08	0.00
Cell 6	65	64	64	1.54	1.54

I_1 column denotes the actual intensity values of the spots, *I_2* the intensity values of the spots by using the GMM, and *I_3* the intensity values of the spots by using the FGMM algorithm. The percentage differences of the calculated, in respect to the actual, intensity values of the two different clustering techniques are also presented in columns five and six.

According to Table I, it is clear that in cases were the segmentation matching factor was 100% (e.g. cell 3-FGMM column 4 Table I), the calculated intensity values were identical with the actual ones (see corresponding cell 3 column 6 in Table III). In the cases where the segmentation matching factor was close to 90%, the percentage difference between the actual and the calculated intensity values was close to 21%. Segmentation accuracy lower than 90% (e.g. cell 5-GMM column 3 in Table I) resulted in percentage differences higher than 25% (see Column 5 Table III). Thus, for precise spot intensity estimation it is essential to achieve accurate segmentation.

Finally, it was found that the computation time required and the iterations involved for the FGMM were equal to that of the GMM.

IV. CONCLUSIONS

In the present study, the FGMM clustering technique is proposed for improving the segmentation of cDNA microarray images. This fuzzy GMM approach proved more accurate in spot intensity computation than the conventional GMM algorithm, thus, providing a more reliable means for estimating gene expression on cDNA microarray images.

References

[1] Y.H. Yang, M. J. Buckley, S. Duboit and T.P.Speed , "Comparison of methods for Image Analysis on cDNA Microarray Data", *Journal of Computational and Graphical Statistics*, vol. 11, pp 108-136, 2002.

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

[3] M.B. Eisen, ScanAlyze (1999). Available: http://rana.lbl.gov/EisenSoftware.htm

[4] Axon Instruments, Inc. (1999): GenPix 4000A User's guide

[5] GeneSifter data center, Available: http://www.genesifter.net/web/dataCenter.html

[6] M.J. Buckley (2000), The Spot user's guide. CSIRO Mathematical and Information Science. Available: http://www.cmis.csiro.au/IAP/Spot/spotmanual.htm

[7] ImaGene, ImaGene 6.1 User Manual, Available: http://www.biodiscovery.com/index/papps-webfiles-action.

[8] S. Beucher, F. Meyer, "The morphological approach to segmentation: The watershed transformation", *Optical Engineering*, Vol. 34, pp. 433-481, 1993.

[9] R. Adams and L. Bischof, "Seeded Region Growing", *IEEE Trans. Pattern Anal. Machine Intell.*, vol 16, pp 641-647, 1994.

[10] D. Bozinov and J. Rahenfuhrer, "Unsupervised technique for robust target separation and analysis of DNA Microarray spots through adaptive pixel clustering", *Journal of Bioinformatics*, vol. 18, pp 747-756, 2002.

[11] Dat Tran, Michael Wagner, Yee W. Lau and Mitsuo Gen, "Fuzzy Methods for Voice-Based Person Authentication", *IEEJ (Institute of Electrical Engineers of Japan) Transactions on Electronics, Information and Systems*, vol. 124, no. 10, pp. 1958-1963, 2004.

[12] Dat Tran and Michael Wagner, "Fuzzy C-Means Clustering-Based Speaker Verification", *Lecture Notes in Computer Science: Advances in Soft Computing - AFSS 2002*, N.R. Pal, M. Sugeno (Eds.), pp. 318-324, Springer-Verlag, 2002.

[13] K. Blekas, N.P. Galatsanos and I. Georgiou, "An unsupervised Artifact Correction Approach for the Analysis of DNA Microarray Images", *Proc. IEEE International Conf. on Image Processing (ICIP)*, vol 2, pp 165-168, 2003.

[14] Betal D, Roberts N, Whitehouse GH., "Segmentation and numerical analysis of microcalcifications on mammograms using mathematical morphology". *Br. J. Radiol.*;70(837):903-17, 1997.

[15] K. Blekas, N.Galatsanos, A. Likas, and I.E. Lagaris, "Mixture Model Analysis of DNA Microarray Images", *IEEE Transactions on Medical Imaging*, vol 24. pp. 901-907, 2005.

[16] S.Lonardi and Y. Luo, "Gridding and Compression of Microarray Images". *IEEE Computational Systems Bioinformatics Conference CSB*, 2004.

[17] James C. Bezdek , "*Pattern Recognition with fuzzy objective function algorithms*", Plenum Press, New York and London, 1987.

[18] X.D. Huang, Y. Ariki, and M.A. Jack, "*Hidden Markov models for speech recognition*", Edinburgh University Press, 1990.

[19] O. Demirkaya, M. H. Asyali and M.M. Shoukri, "Segmentation of cDNA Microarray Spots Using Markov Radom Field Modeling", *Bioinformatics*, vol. 21 no. 13, pp. 2994-3000, 2005.

[20] Y. Balagurunathan, E. R. Dougherty, Y. Chen, M.L. Bittner and J.M. Trent, "Simulation of cDNA Microarray via a parameterized random signal model", *Journal of Biomedical Optics*, vol. 7, pp. 507-523, 2002.

[21] Laurie Heyer, MicroArray Genome Imaging & Clustering (MAGIC) Tool, Davidson College, Available: http://www.bio.davidson.edu/projects/magic/magic.html

[22] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm" *J. Roy. Statist. Soc. B*, vol. 39, pp. 1–38, 1977.

[23] The MathWorks, Inc. Software, MATLAB®.

[24] James C. Bezdek, "*Pattern Recognition with fuzzy objective function algorithms*", Plenum Press, New York and London, 1987.