

SEGMENTATION OF MICROARRAY IMAGES USING GRADIENT VECTOR FLOW ACTIVE CONTOURS BOOSTED BY GAUSSIAN MIXTURE MODELS

Emmanouil Athanasiadis¹, Dionisis Cavouras², Panagiota Spyridonos¹, Dimitris Glotsos¹,
Ioannis Kalatzis², and George Nikoforidis¹

¹ Medical Image Processing and Analysis (MIPA) Group, Laboratory of Medical Physics,
University of Patras, 26500 Patras, Greece
e-mail: mathan@upatras.gr, web page: <http://mipa.med.upatras.gr>

² Medical Image and Signal Processing (MEDISP) Laboratory, Department of Medical Instruments Technology,
Technological Educational Institute of Athens, Ag. Spyridonos Street, Aigaleo, 122 10, Athens, Greece
e-mail: cavouras@teiath.gr, web page: <http://medisp.bme.teiath.gr>

Keywords: Microarrays, Gaussian Mixture Models (GMM), Active Contours.

Abstract. *In this paper, a new methodology for the segmentation of cDNA microarray images is proposed, based on the combination of Gaussian Mixture Models (GMM) with Gradient Vector Flow (GVF) active contours. A simulated microarray image of 1000 spots was produced using a standard procedure. 5 real microarray images were used to evaluate the performance of our algorithm. GMM was firstly applied in all individual cells (spot with each background). The output was used to initialize a GVF active contour. The major advance of our method is that it overcomes limitations of both GMM and active contours when used individually. Segmentation matching factors and mean intensity values were calculated for every cell using GMM, GVF, and the combination of GMM and GVF in the simulated data. Pairwise correlations and mean absolute errors were also calculated by using real microarrays. Numerical experiments using both simulated and real images showed that our method was more accurate in measuring intensity values and detecting actual boundaries of spots, compared with GMM and active contours used individually. Results concerning the segmentability and the mean intensity value of the proposed algorithm were more accurate, as compared with those methods when used individually.*

1 INTRODUCTION

DNA microarray images are used by molecular biologists to identify and measure the expression level of thousands of genes simultaneously [1]. On complementary DNA (cDNA) Microarray experiments [2], the fluorescence intensity value of each spot is related to the expression abundance of the corresponding DNA sequence [4].

The main process for measuring spot intensity values involves three steps [1, 4]: the gridding step, where localization of each spot is achieved, the segmentation step, where identification of the foreground and background of each spots is accomplished, and finally, the intensity extraction step, where calculation of the fluorescence intensity is performed for each spot. However, various sources of noise, during image acquisition [1], degrade image quality. This distorts the shape of the spots leading to inaccuracies in intensity measurements [5]. Additionally, the locations of the arrayer and its sub-arrays may vary from image to image due to imperfections, introducing an error in the scanning procedure [21].

In the past, several methods have been proposed for the segmentation of microarray images [6, 7, 8, 9, 10]. In the ScanAnalyze [6], a fixed circle segmentation method has been implemented with the assumption that spots are considered to be circular with a fixed, predefined radius. In the GenePix [7], an adaptive circle segmentation technique is used. According to this technique, the radius of each spot is not constant but adjusted for every spot separately. The most commonly used methods in adaptive shape segmentation technique are the watershed [11] and seeded region growing [12], which are used by the Spot method [23]. In all previous techniques, the major disadvantage is that a-priori knowledge of the exact location of centers is crucial [3]. In ImaGene [10], a histogram based segmentation method is used, in which the 80th and 95th percentile values of the histogram are defined as spot intensities to compute the mean intensity value.

Intensity based segmentation techniques have also been developed, such as k-means and k-medoids [13]. The major drawbacks of those techniques are that they do not adapt well to irregular based clusters and do not utilize all the available prior knowledge about the data [13]. Other researchers have also introduced more sophisticated methods for the segmentation process, such as mixture model analysis [13] and deformable models [14 - 16]. In the mixture model analysis, the major drawback is the identification of spots with intensity close to background [14]. In methods based on multiple snakes, the initialization points of the active contour are defined by the borders of the cell. Cell is denoted as the rectangular area that includes the spot and its surrounding area. In noisy cells, however, active contours do not succeed in delineating the spot efficiently [15].

In the present study a new method for spot identification is proposed, based on the combination of gaussian mixture models (GMM) clustering with gradient vector flow (GVF) active contours. The GMM is used to estimate the foreground and background pixels of each cell and that information is employed to initialize a GVF snake to delineate the spot's boundary. These two techniques were combined in order to enhance the performance obtained by employing each algorithm separately, especially for spots with intensity values close to background.

2 METHOD

2.1 Automatic gridding for cell localization.

A cDNA microarray image usually consists of an arrayer, several subarrays, and thousands of spots corresponding to specific genes (see Figure 1). Gridding is the process of segmenting each subarray of the arrayer into cells that contain one spot with its background. After gridding, segmentation may be applied easier within each individual cell. The gridding algorithm adopted in the present study consisted of the following steps.

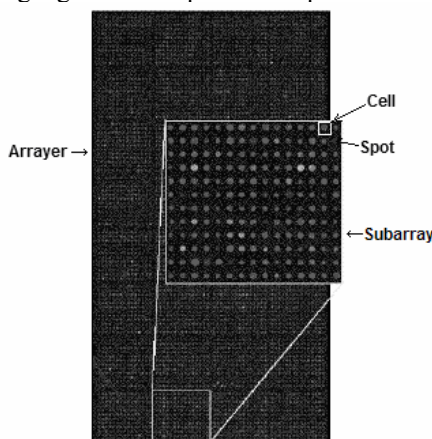


Figure 1: An Arrayer consists of 8x4 subarrays, each subarray of 19x21 cells and spots respectively [19].

1) Firstly, average intensities were calculated along rows and columns for either Red ('R') or Green ('G') channel (R for Cyanine Cy3 and G for Cyanine Cy5). Second, noise was suppressed by means of a low pass filtering mask [5]. Sub-array regions were calculated by finding the local minima of either the R or the G signals in the horizontal and vertical axes [13 and 17]; it should be noted that by choosing R or G channel for grid localization, no significant differences were observed.

2) A similar procedure was used for the identification of 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. Localization of local maxima is illustrated in Figure 2. Additionally, result of the gridding step, applied to a sub-array region of a microarray image that contains 19x21 spots are illustrated in Figure 3.

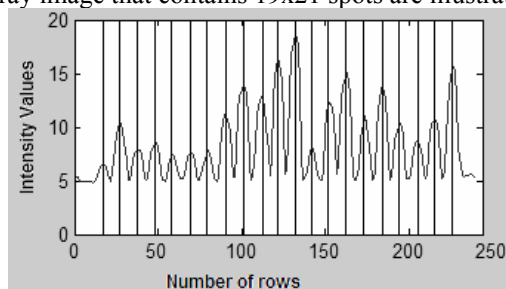


Figure 2: localization of the maxima by Matlab's 'imregionalmax' function.

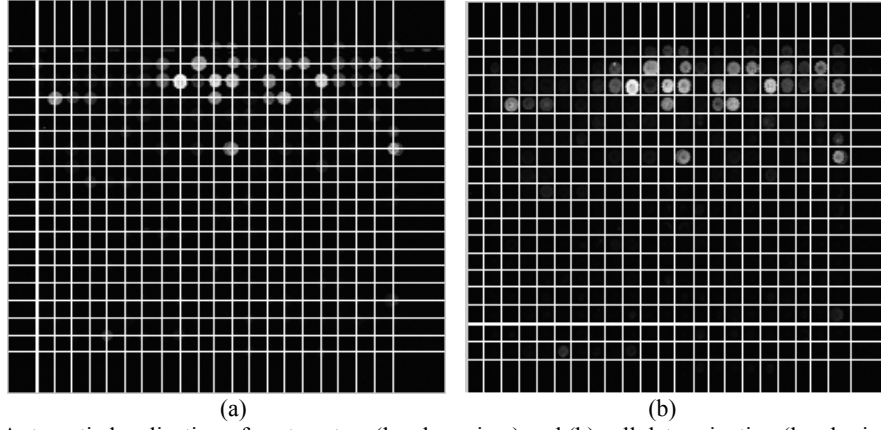


Figure 3: (a) Automatic localization of spot centers (local maxima) and (b) cell determination (local minima) applied on a 19x21 subarray.

2.2 Spot Segmentation

A hybrid segmentation method was developed and applied to every cell, determined by the gridding process. It comprised an unsupervised classification algorithm (GMM) [13, 22], for a first estimation of spot and background pixels, and a GVF Active Contour [18] procedure, for determining the exact boundaries of each spot. The GVF-Snake functioned automatically using as starting points the segmentation results of the GMM algorithm.

2.2.1 The Gaussian Mixture Model

As a first step, prior to GMM clustering, a K-means algorithm [13, 20, 22] was used to estimate two clusters ($K=2$), the foreground and background classes. Based on the probability density functions (PDFs) of those two classes, the GMM algorithm refined the two classes, employing Bayes' rule according to Equation 1.

$$P(C_j|X) = p(X|C_j) \frac{P(C_j)}{p(X)} \quad \text{Equation 1}$$

Where, $p(X|C_j)$ is the PDF of class j evaluated at X , $P(C_j)$ is the prior probability of for class j , and $p(X)$ is the overall PDF evaluated at X .

GMM estimates $p(X|C_j)$ as a weighted average of multiple Gaussians. The estimation of the PDF of each class j is calculated by Equation 2.

$$p(X|C_j) = \sum_{k=1}^{N_c} \pi_k G_k \quad \text{Equation 2}$$

where π_k is a weighting factor of the k -th Gaussian G_k , equal to 0.5 for 2 classes, described in Equation 3, and N_c is the number of features belonging to class j .

$$G_k = \frac{1}{(2\pi)^{n/2} |V_k|^{1/2}} e^{\left[-\frac{1}{2}(x-\mu_k)^T V_k^{-1} (x-\mu_k)\right]} \quad \text{Equation 3}$$

where, n is the number of features, V_k and μ_k are the cluster's k covariance matrix and mean value respectively, and T stands for transpose.

The GMM follows an Expectation-Maximization procedure (E-Step/M-Step) [5]. In the E-Step, estimation is made of the prior probability of every pixel value (pattern X_p) to belong to cluster i , according to Equation 4. In the M-Step, the parameters π , μ , and V (represented by θ in equation 4) are recomputed, according to Equations 5, 6 and 7.

$$t_{ip} = P(G_i|X_p) = \frac{\pi_i p(X_p|\theta_i, C_i)}{p(X_p)} = \frac{\pi_i p(X_p|\theta_i, C_i)}{\sum_{j=1}^{N_c} \pi_j p(X_p|\theta_j, C_j)} \quad \text{Equation 4}$$

$$\pi_i(t+1) = \frac{1}{N_c} \sum_{p=1}^{N_c} t_{ip}(t) \quad \text{Equation 5}$$

$$\mu_i(t+1) = \frac{\sum_{p=1}^{N_c} t_{ip}(t) X_p}{\sum_{p=1}^{N_c} t_{ip}(t)} \quad \text{Equation 6}$$

$$V_i(t+1) = \frac{1}{N_c \pi_i(t)} \sum_{p=1}^{N_c} t_{ip}(t) \left[(X_p - \mu_i(t)) (X_p - \mu_i(t))^T \right] \quad \text{Equation 7}$$

The two steps are repeated until either $V_i(t+1)=V_i(t)$ or the number of iterations t equals a predefined value. For each pattern X_p , the confidence (Equation 8) for each class j is computed and the pattern is classified to the class with the highest confidence.

$$P(C_j) p(X | \theta_x, C_j) \quad \text{Equation 8}$$

The output of the algorithm is a binary image of each cell, black pixels for background and white pixels for foreground that correspond to the actual spot image. Employing a canny edge detection algorithm [24], the boundary pixels of each spot were estimated and were used to initialize the GVF-Snake that determined the exact outline of each spot.

2.2.2 The GVF active contour

The GVF-Snake attempts to minimize the energy E_{snake} , computed by Equation 9 [17].

$$E_{snake} = \int_0^1 E_{int} [v(s)] ds + \int_0^1 E_{ext} [v(s)] ds \quad \text{Equation 9}$$

where $v(s)$ is a curve with spatial coordinates $[x(s), y(s)]$ $s \in [0,1]$, $E_{int}[v(s)]$ is the internal energy, defined by Equation 9, and $E_{ext}[v(s)]$ is the external energy, defined by Equation 10.

$$E_{int} [v(s)] = \frac{1}{2} \left[a(s) \left| \frac{\partial v(s)}{\partial s} \right|^2 + \beta(s) \left| \frac{\partial^2 v(s)}{\partial s^2} \right|^2 \right] \quad \text{Equation 10}$$

$$E_{ext} [v(s)] = - \left| \nabla (G_\sigma(x, y) * I(x, y)) \right|^2 \quad \text{Equation 11}$$

where, $G_\sigma(x,y)$ is a two-dimensional Gaussian function with σ standard deviation, ∇ is the gradient operator of image spot I and $\alpha(s)$ and $\beta(s)$ are variables that control the internal energy. In the present study, the GVF-Snake toolbox [17] was used.

2.3 Intensity Extraction

The final step of the process was to compute the foreground and background mean intensity values I for each spot in the Red and Green channels separately. The intensity value of each spot was computed by subtracting the mean foreground from mean background value [1].

The proposed methodology is illustrated in Figure 4.

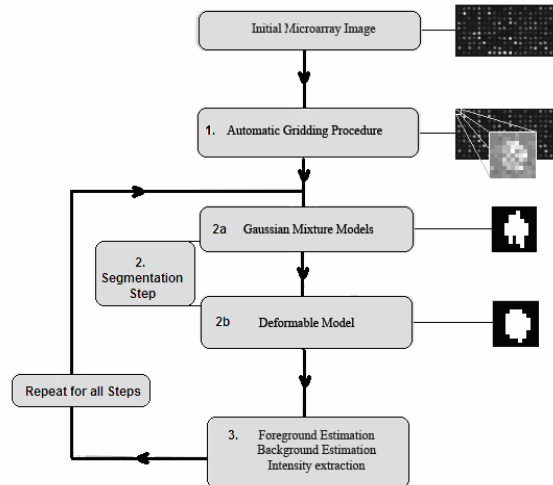


Figure 4: The basic scheme of the proposed methodology.

3 MATERIAL

For the validation of the proposed method, two simulated cDNA images were produced [19, 20]. In order to generate spots with realistic characteristics, the following procedure was followed. A cDNA image, consisting of 1000 spots, was used as a template, and its binary version was produced employing a thresholding technique [19] (see Figure 5). The second simulated binary image was produced with the difference that in that image all spots were alike, that is they have the same shape and dimension. This replicated spot was chosen randomly from the first binary image. In both binary images, the location as well as the area of each spot was known. The mean intensity value of each spot was pre-defined, ranging between 0 and $2^{16}-1$ for both the R and G channels. Spot intensities were produced using an exponential distribution with mean value the pre-defined mean intensity value. Background intensities were drawn from a single exponential distribution, with mean value determined from the original image's mean intensity background [19].

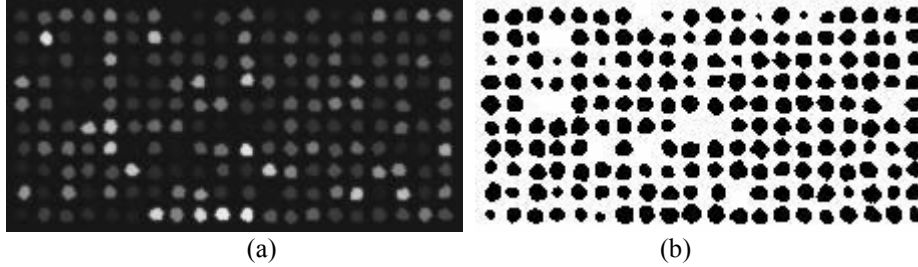


Figure 5: (a) Original simulated microarray image and (b) the binary image used to produce the simulated data.

The accuracy of segmentation was numerically calculated using the segmentation matching factor (Equation 12) [22].

$$accuracy = \frac{A_{calc} \cap A_{real}}{A_{calc} \cup A_{real}} \times 100 \quad \text{Equation 12}$$

where A_{calc} 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 [22] while a score of less than 50% indicates poor segmentation [22].

3 RESULTS AND DISCUSSION

The proposed method was applied to both simulated microarray images. Actual mean intensity values and boundaries of each spot were a-priori known. Gridding procedure was performed for both simulated images separately, and for every cell produced by the addressing process, the GMM, the snakes and the GMM-snake algorithm was applied. Thus, a set of 3 binary images, having all spot boundaries delineated, were produced. The Segmentation matching factor (equation 12) was then calculated for each binary image, in order to estimate the accuracy of each method quantitatively. In Table 1, results of 6 different spots obtained from the first simulated image for the R channel are illustrated. Additionally, in Table 2, the G channel results of 5 identical spots, calculated from the second simulated image, are also provided. Active contour parameters α , β and σ were determined by means of multiple experimentation for obtaining optimal results. The same optimal parameters were used in all cases, for our results to be comparable.

	Original Cell	Real Boundaries	GMM	Snake	GMM+Snake
	spot 1	Accuracy:	87.60	94.21	99.17
	spot 2	Accuracy:	95.88	86.78	98.35


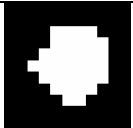


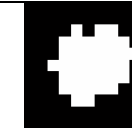
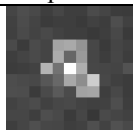
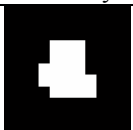

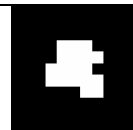
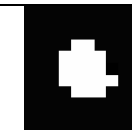
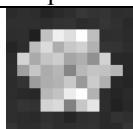
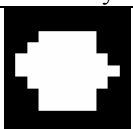

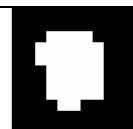
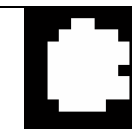
				
spot 3	Accuracy:	91.74	93.24	95.04
				
spot 4	Accuracy:	95.88	96.49	96.69
				
spot 5	Accuracy:	89.26	90.08	91.74

Table 1: Comparative results for 5 different spots for the G channel. The first column indicates the simulated spot, the second column indicates the actual boundaries of the spot and the third, forth, and fifth columns present the results of the GMM, the snakes and the GMM-snake combination algorithms and the corresponding matching factors.

Moreover, an overall accuracy for all simulated spots was calculated. Standard deviations of matched and mismatched pixels for every cell were calculated by using each of the three different techniques. Results are illustrated in Table 3.

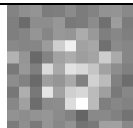



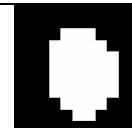
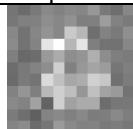


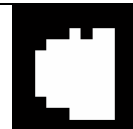
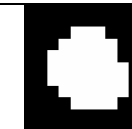



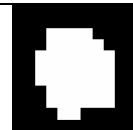




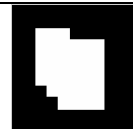




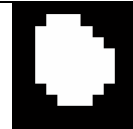
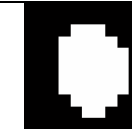
Original Cell	Real Boundaries	GMM	Snake	GMM+Snake
				
spot 1	Accuracy:	90.08	91.74	97.52
				
spot 2	Accuracy:	90.91	86.78	95.87
				
spot 3	Accuracy:	90.91	92.56	94.22
				
spot 4	Accuracy:	97.52	85.95	99.17
				
spot 5	Accuracy:	95.04	91.74	97.52

Table 2: Comparative results for 5 identical spots for the R channel. The first column indicates the simulated spot, the second column indicates the actual boundaries of the spot and the third, forth and fifth columns present the results of the GMM, the snakes and the GMM-snakes combination algorithms and the corresponding matching factors.

	GMM		Snake		GMM + snake	
	Match	Mismatch	Match	Mismatch	Match	Mismatch
Overall Accuracy (%)	93.88	6.11	90.12	9.88	95.04	4.96
Standard Deviation	7.61	7.61	3.75	3.75	3.63	3.63

Table 3: The Overall accuracy and standard deviation achieved by each of the three different techniques

It is clear that the best overall accuracy for matching pixels (95.04%) was accomplished by the proposed methodology, followed by GMM (93.88%). Significant was also the matching pixel accuracy achieved by snakes (90.12%). According to table 3, the standard deviation of matched pixels was the lowest (3.63) in the case of our methodology, rendering it more robust than the other two techniques.

Furthermore, the comparison of the algorithms was carried out in a set of five real microarray images, concerning *Saccharomyces cerevisiae*, obtained from publicly available database [19]. In common reference channel (G), each spot should have the same gene expression ratio throughout replication experiments, thus correlation between each experiment should be maximal [25]. Moreover, pairwise mean absolute error (MAE) was calculated among the replicates. Box plots for correlation and MAE are illustrated in fig. 6(a) and 6(b) respectively. It should be noted that the higher the correlation and the lower the MAE values, the better the performance of the segmentation algorithm [25].

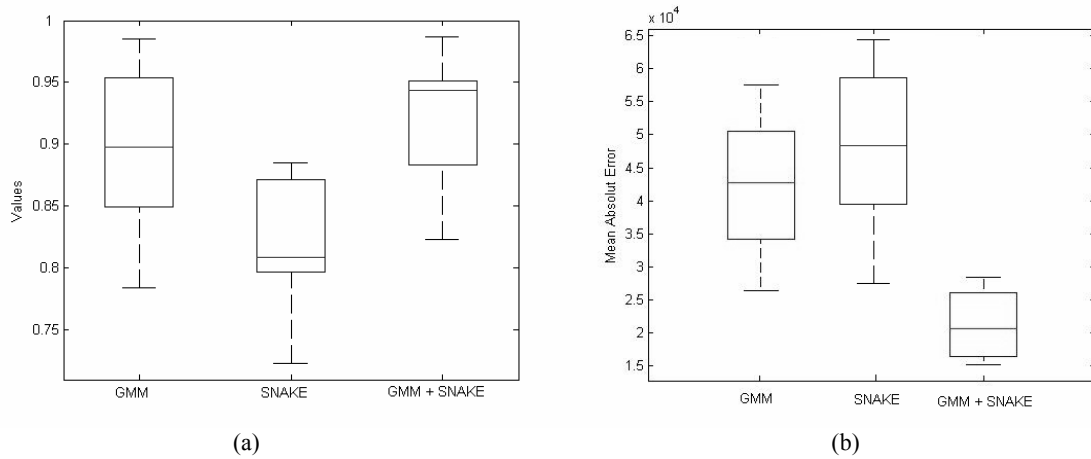


Figure 6: Box plots that illustrate (a) the correlation and (b) the MAE among the replicates.

The GMM+snakes algorithm achieved a correlation of 0.95 and MAE close to 2.5×10^4 . GMM and snake algorithms resulted to a lower correlation value, that of 0.90 and 0.82, and a higher MAE that of 4.5×10^4 and 4.9×10^4 respectively. The results obtained by the real microarray images verify the simulated experiments and support the performance superiority of the FGMM algorithm in segmenting DNA images.

4 CONCLUSIONS

In the present study, a GMM-snakes combination method is proposed for the segmentation of cDNA microarray images. This approach shows that limitations of GMM and snakes used individually could be overcome. It was also revealed that using the proposed methodology, results concerning the segmentability and the mean intensity value for each spot were more accurate, as compared with those methods when used individually.

5 ACKNOWLEDGEMENT

We would like to thank the Greek State Scholarships Foundation (I.K.Y) for funding the above work.

REFERENCES

- [1] Yang H., Buckley J., Duboit S. and Speed P. (2002), "Comparison of methods for Image Analysis on cDNA Microarray Data", *Journal of Computational and Graphical Statistics*, Vol. 11, pp 108-136.
- [2] Schena M., Shalon D., Davis R.W. and Brown P. O. (1995), "Quantitative monitoring of gene expression patterns with a complementary DNA microarray", *Science* 270, pp. 467-470.

- [3] Bozinov D. and Rahenführer J. (2002), "Unsupervised technique for robust target separation and analysis of DNA Microarray spots through adaptive pixel clustering", *Journal of Bioinformatics*, Vol 18, pp 747-756.
- [4] Yang H., Buckley J., Dubois S. and Speed P. (2001), "Analysis of cDNA Microarray images", *Briefing in Bioinformatics*, vol.2, No.4, pp. 341-349.
- [5] Blekas K., Galatsanos N.P. and Georgiou I. (2003), "An unsupervised Artifact Correction Approach for the Analysis of DNA Microarray Images", in *Proc. IEEE International Conf. on Image Processing (ICIP)*, vol 2, pp 165-168.
- [6] Eisen M.B., ScanAlyze (1999). <http://rana.lbl.gov/EisenSoftware.htm>
- [7] Axon Instruments, Inc. (1999): *GenPix 4000A User's guide*
- [8] GeneSifter data center: <http://www.genesifter.net/web/dataCenter.html>
- [9] Buckley M.J. (2000), *The Spot user's guide*. CSIRO Mathematical and Information Science. <http://www.cmis.csiro.au/IAP/Spot/spotmanual.htm>
- [10] ImaGene, ImaGene 6.1 User Manual <http://www.biodiscovery.com/index/papps-webfiles-action>.
- [11] Beucher S., Meyer F (1993), "The morphological approach to segmentation: The watershed transformation", *Optical Engineering*, Vol. 34, pp. 433-481.
- [12] Adams R. and Bischof L. (1994), "Seeded Region Growing", *IEEE Trans. Pattern Anal. Machine Intell.*, vol 16, pp 641-647.
- [13] Blekas K., Galatsanos N., Likas A., and Lagaris I. (2005), "Mixture Model Analysis of DNA Microarray Images", *IEEE Transactions on Medical Imaging*, vol 24. pp. 901-907.
- [14] Srinark T. and Kambhamettu C. (2004), "A Microarray Image Analysis System Based on Multiple Snakes". *International Conference on Bioinformatics and its Applications (ICBA)*
- [15] Katzer M., Kummert F. and Sagerer G. (2003), "Methods for Automatic Microarray Image Segmentation". *IEEE Transaction on nanobioscience*, vol.2 No.4, pp. 202-214.
- [16] Srinark T. and Kambhamettu C. (2001), "A framework for Multiple Snakes", *Computer Vision and Pattern Recognition*, vol.2 , pp. 202-209.
- [17] Xu C. and Prince J. (1997), "Gradient Vector Flow:A New External Force for Snakes", *IEEE Proc. Conf. on Comp. Vis. Patt. Recog. (CVPR)*.
- [18] Nagarajan R. , Peterson C. (2002), "Identifying Spots in Microarray Images", *IEEE Transactions on nanobioscience*, Vol. 1, No.2, pp. 78-84
- [19] L. J. Heyer, D. Z. Moskowitz, J. A. Abele, P. Karnik, D. Choi, A. M. Campbell, E. E. Oldham and B. K. Akin, (2005) "MAGIC Tool: integrated microarray data analysis", *Bioinformatics*, Vol. 21, no. 9, pp. 2114 – 2115.
- [20] Balagurunathan Y., Dougherty E., Chen Y., Bittner M. and Trent J. (2002), "Simulation of cDNA Microarray via a parameterized random signal model", *Journal of Biomedical Optics*, vol. 7, pp. 507-523.
- [21] Wang Y., Shih F., Ma M. (2005), "Precise Gridding of Microarray Images By Detecting and Correcting Rotation in Subarrays". *Sixth Inter. Conf. on Computer Vision, Pattern Recognition and Image Processing*.
- [22] Athanasiadis E., Cavouras D., Spyridonos P., Glotsos D., Kalatzis I., and Nikofofidis G. (2006) "Segmentation of complementary DNA microarray images using the Fuzzy Gaussian Mixture Model Technique", *The International Special Topic Conference On Information Technology in Biomedicine (ITAB)*.
- [23] Yang Y (2000), "SPOT User Guide" (www.cmis.csiro.au/IAP/spotmanual.htm),.
- [24] Canny, J (1986), "A Computational Approach to Edge Detection", *IEEE Trans. Pattern Analysis and Machine Intelligence*, 8:679-714.
- [25] Lehmann A., Ruusuvoori P. and Yli-Harja O. (2006) "Evaluating the performance of microarray segmentation algorithms", *Bioinformatics*, Vol 22, 2910-2917.