2004 8th International Conference on
Control, Automation, Robotics and Vision
Kunming, China, 6-9th December 2004

"COPASETIC ANALYSIS": AUTOMATED ANALYSIS OF BIOLOGICAL GENE EXPRESSION IMAGES

Karl Fraser, Paul O'Neill*, Zidong Wang and Xiaohui Liu*
DISC, Brunel University, Uxbridge, Middlesex, UB8 3PH, UK
Contact: xiaohui.liu@brunel.ac.uk
Tel/Fax: +44 (0)1895 203397 / +44 (0)1895 251686

Abstract

In the past decade computational biology has come to the forefront of the public's perception with advancements in domain knowledge and a variety of analysis techniques. With the recent completion of projects like the human Genome sequence, and the development of microarray chips it has become possible to simultaneously analyse expression levels for thousands of genes. Typically, a slide surface of less than 24cm^2, receptors for 30,000 genes can be printed, but currently the analysis process is a time consuming semi-autonomous step requiring human guidance. The paper proposes a framework which facilitates automated processing of these images. This is supported by real world examples which demonstrate the technique's capabilities along with results which show a marked improvement over existing implementations.

Introduction

With the development of microarray technology it has become possible for biologists to analyse many thousands of genes simultaneously. A single microarray chip, can contain the genome of a whole organism, treated under different conditions. The chip can be analysed using various techniques, such as clustering [1], and modelling [2]. This is accomplished by a technique called 'competitive hybridisation', which is conducted on a microscopic scale [3]. Here, we provide a brief review of relevant background material, for a more detailed explanation readers may find references [4] and [5] of interest.

Once the genes have been identified, samples of their DNA can be printed onto a specially treated glass slide which is similar in dimensions to a standard microscope slide. This chip can then be used to detect the presence of these genes in the RNA extracted from cells which were treated under varying conditions. The chip is then digitised using a dual laser scanning device, producing a two-channel 16bit grey-scale image. The gene receptor locations in this image (typically 16~20 pixels in diameter) are identified; their median intensity values are then measured and summarised as $\log_2$ ratios across both channels.

This type of analysis has many uses, an example of which is the comparison between cells for a patient before and after infection by a disease. If particular genes are used more after infection (highly expressed) then it can be surmised that these genes may play an important role in the life cycle of this virus.

Figure 1, shows a typical microarray slide, with a zoomed section across two full blocks of genes. Each spot on this image is about two tenths of a millimetre in diameter and represents a specific gene. The image measures approximately 5000×2000 pixels and requires 40MB of storage memory. Due partly to the size of the image produced and partly because of the inherent variation which can be expected of any biological process; the images produced are extremely noisy. This leads to a complex yet very interesting computer vision problem, whereby so far, complete automation of this analysis process has proved to be elusive.

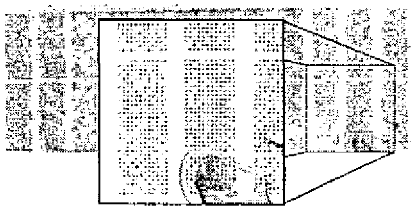


Figure 1: Example slides from test dataset illustrating varying structure and noise elements.

This paper proposes a framework which is designed around the flexible analysis of these images using no prior knowledge of the slide. Real microarray images are used to verify the performance of the system and also comparisons are made against analysis conducted by trained biologists using one of the dominant analysis packages.

Background

Feature detection is the process whereby either an algorithm or an operator categorises the pixels in the image as belonging to either a specific gene spot or the background. This consists of two distinct stages; the first 'spotting', such as the Bayesian approach proposed by Hartelius and Carstensen [6] which divides the imagery into manageable blocks. The second involves segmentation [6, 7] which classifies pixels in a region immediately surrounding a gene as belonging to either the foreground or background domains.

Once the pixels for each spot have been identified, they can then be summarised as $\log_2$ ratios. For a detailed comparison of many standard techniques used for this purpose, refer to [8].

0-7803-8653-1/04/$20.00 © 2004 IEEE

The large amount of time that has to be spent on manually processing the microarrays has lead to the recent interest in trying to fully automate the process. Bozinov and Rahnenfuhrer [9] proposed clustering the full image area in one step; however this is not computationally feasible with current processing power. To overcome these issues, Bozinov [10] proposed an abstraction of the k-means [11] technique whereby pre-defined centroids were chosen for both foreground and background, to which all pixel intensities could be assigned. Although the proposed abstraction [10] is portrayed as a clustering technique, it would be more accurately described as a simple assignment process between two classes. Unfortunately although k-means is able to choose centroids according to the dataset's characteristics, this approach is inherently biased towards outlying values (saturated pixels for example), and not the true region of interest (the foreground pixels). Other methods, such as the application of wavelets [12] and Markov random fields [13] show great promise, however, at this time they have only been attempted on what would be classified as 'good slides' (whereby the noise is not of an extreme nature). If these techniques fail to determine the location of just one gene, the system will fail, thus having to fall back on user intervention in order to recover.

Overall our work has been based on slides representing two underlying structures with varying degrees of noise. Combined, these consist of 10 images, which were selected as they contained varying anomalies both in the background intensities and in the printed spot structure. In figure 1, an image from these sets is displayed to highlight the varying structure. This example slide is also suggestive of the problems which are associated with the processing of this type of data, such as background artefacts and gene block misalignment. In the next section this paper will present a framework which has been established to facilitate the processing of these images. This is a challenging and important problem and as far as the authors are aware, it is the first time such a comprehensive framework has been applied to microarray image analysis.

Copasetic Analysis Overview

Copasetic Analysis (CA) is a framework in which automated microarray image analysis can be conducted. Unlike other techniques that have been proposed to this effect, it is not a rigid framework, in fact it is its modularity and adaptability that give it its robustness. In figure 2, a skeletal structure of this process is presented, showing the required stages from the original input images, through to calculating the gene spot $\log_2$ ratios. In this diagram we can see there are four key parts which make up the CA process, which will be described more detail. Importantly, each of these parts and the processes within are goal orientated, which means that the techniques can be 'swapped out' to allow various computational tasks to be conducted. For example, in *image layout* if the method based on periodicity underperforms, an alternative, such as wavelet de-convolution could be utilised. Some stages are composed of combinations of existing and new techniques, such as the 'Data Services' stage, while others are novel algorithms like *Copasetic Clustering* [14] which facilitate the application of existing clustering techniques to a previously infeasible dataset. Another interesting point is the adaptability of the framework when things do not quite go as planned, in the above example where one method underperforms, it could be that the parameter settings were corrupted, in this case it would be preferable to backtrack to a previous step and try again. This backtracking capability is managed by a quality assessment process which is performed at the completion of each component. If the quality assessment process determines that the data itself could be improved, it can request a different view of the data from the *Image Transformation Engine*

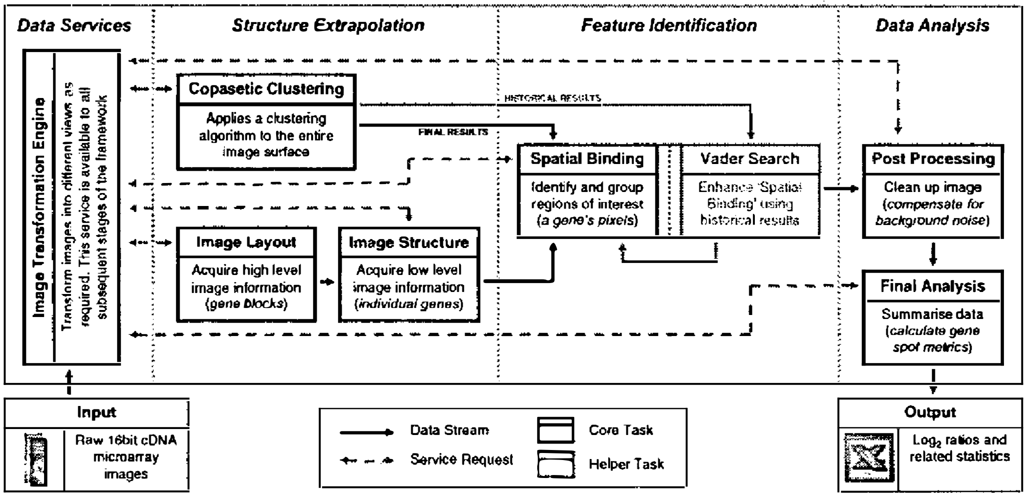

Data Services
Structure Extrapolation
Feature Identification
Data Analysis
Image Transformation Engine
Transform images into different views as required. This service is available to all subsequent stages of the framework
Copasetic Clustering
Applies a clustering algorithm to the entire image surface
HISTORICAL RESULTS
FINAL RESULTS
Spatial Binding
Identify and group regions of interest (a gene's pixels)
Vader Search
Enhance 'Spatial Binding' using historical results
Post Processing
Clean up image (compensate for background noise)
Image Layout
Acquire high level image information (gene blocks)
Image Structure
Acquire low level image information (individual genes)
Final Analysis
Summarise data (calculate gene spot metrics)
Input
Raw 16bit cDNA microarray images
Data Stream
Service Request
Core Task
Helper Task
Output
Log2 ratios and related statistics

Figure 2: Copasetic Analysis Workflow Diagram.

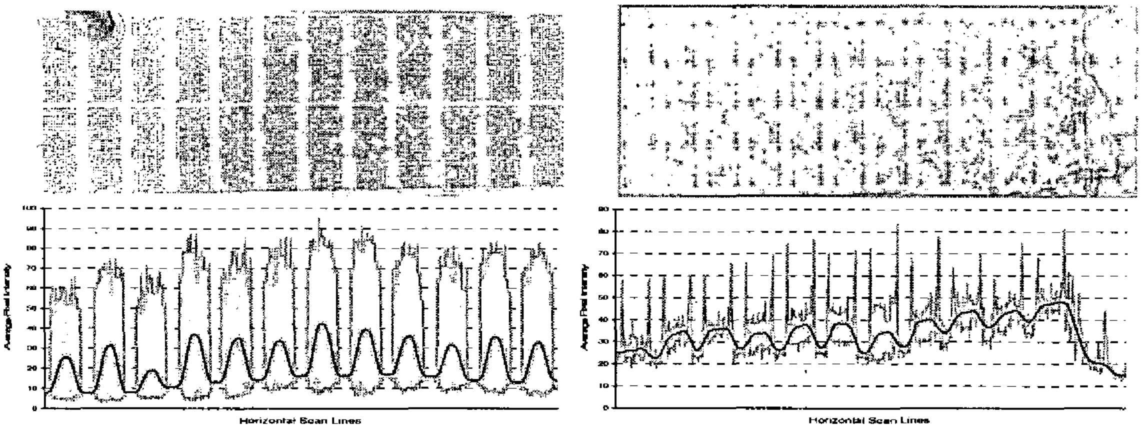


Figure 3: The horizontal profiles aligned to the raw imagery.

(ITE). An example of this would be the application of a low-pass filter to the data if a high background was detected as being detrimental to the analysis. This framework is designed to process images which have some form of regular structure like that found in microarray imagery, as such, the input for the process is always going to be the raw image data. The ITE is the only component which has direct access to this raw data; its function being to supply this data both unaltered and in various transformed and filtered views to the components as requested. For example, the view requested could be a simple summary such as providing the mean pixel intensity of the image, or a more complex image transformation and filtering technique. It is conceived that in this way components will not be restricted to one view of the data as is typical, but can benefit from a multitude of perspectives.

After the ITE has acquired the raw imagery the first components in the framework to be executed are those that make up the 'Structure Extrapolation' stage, which are designed to discover both the structure and composition of the image. The *Image Layout* component could use a variety of techniques to ascertain the general layout of the image surface; this constitutes the discovery of the gene blocks. One possible solution for this is to take an averaged cross-sectional view of the slide surface in the appropriate horizontal or vertical direction. Figure 3 shows two images with their corresponding horizontal profiles, where the light grey represents the image profile and the black line is generated using a combination of moving filters. This is a good example of how low pass filters can be applied in an attempt to improve the (subjectively measured) quality of the data for human or machine interpretability [15]. The left hand plot of figure 3 shows the profile for what would be classified as a well printed 'good' slide, and from this it is relatively easy to distinguish the 12 block rows that exist in the slide (the peaks) and the inter-block gaps in-between (the valleys).

With the major areas of interest defined, the *Image Structure* component then uses a similar process which is conducted with a finer granularity in order to discover detailed structures. With microarrays we know that the slide should have a regularly repeating structure in each gene block and therefore this information can be used to help guide the block structure discovery. This compositional stage can either utilise the raw data or more beneficially, one of the alternate views as provided by the ITE service.

The Copasetic Clustering (CC) method [14] facilitates the requirement of full-slide segmentation. Initially it arbitrarily divides up the image into spatially related areas (normally very small grid squares). Each of these areas are clustered using a standard technique such as *k*-means [11] or fuzzy *c*-means [16] and the result is stored. Then, representatives are calculated for each of the clusters that exist, and these are further clustered in the next generation, such a process is repeated until all of the sub-clustered groups have merged. The net result is every pixel will have been clustered into one of *n* groups, and this should be very close to clustering the whole image using a standard approach if such a process were possible. This technique gives the advantages of clustering while at the same time reducing the processing and memory requirements to those feasible with modern desktop computing technology.

From the structural information that has been determined, we can now start to identify objects of interest within the image; this constitutes grouping together all the pixels that form a gene spot. This is achieved by *Spatial Binding* which uses both the estimated gene centre position and the clustering results, to search and combine groups of pixels that fall within close proximity to each other. This process can be completed for the majority of the genes that were well defined and the information gained can be used to generate an average gene spot. This mask construct is required, as clustering will never guarantee that all pixels belonging to gene spots will be identified correctly. One of the main strengths of CC is its transparency into the intermediate layers as illustrated in figure 4. Using this knowledge (when coupled with the average gene spot characteristics) we can move back through the historical clustering results until the criteria for the average gene spot have been satisfied.

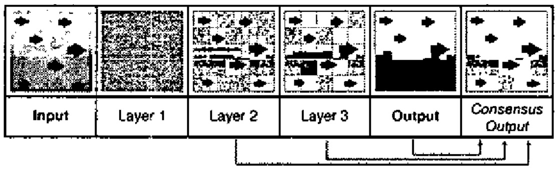


Figure 4: Example of CC Layers

The final stage consists of components that could be thought of as the more conventional stages of microarray analysis. Here, using the structural information that has been identified and the raw image data, the genes on the microarray are analysed. Generally this consists of two stages, one of *Post Processing* (such as background correction) and a second of *Final Analysis*, a data reduction stage (such as converting the values from all the pixels that make up a gene spot into one representative value)..

Experimental Results

In this section, we compare the overall performance of the framework with that as determined by a commercially viable process as seen with the GenePix® package. During the development of the CA process a variety of components were implemented in the slots. Ultimately the framework will determine the appropriate component on a per slot basis, however, for this evaluation we choose components which were known to perform well. First of all we will look at CA's success in discovering the structural composition of the slides, including overall block structure and gene spot locations. Then we will present a measure of accuracy for the entire process which will allow us to compare our automated system with that of an expert human operator.

It was initially envisioned that we would demonstrate the frameworks capability using two sets of disparately structured microarray images with varying quality (see figure 3). The left hand microarray image is typical of an industrially prepared slide where the genes are well defined and evenly printed, however there is minor meta-block drift and significant levels of background noise. This can be a challenge in itself when determining slide structure, but is made all the more interesting when larger anomalies are present (such as top left). The image on the right hand side of figure 3 is part of a calibration run on a bench top spotter device. In contrast to the previous image the gene spots and meta-blocks are poorly defined, with high levels of background noise throughout. These features mark the image as a good score card for algorithm development. These two images were chosen along with several contemporise giving a total of ten images.

Overall CA successfully processed the block layout of all slides, with no prior knowledge of their structure. Figure 5 shows a resultant structural analysis of an image where it can be seen that the master blocks have been clearly defined even though there is misalignment throughout.

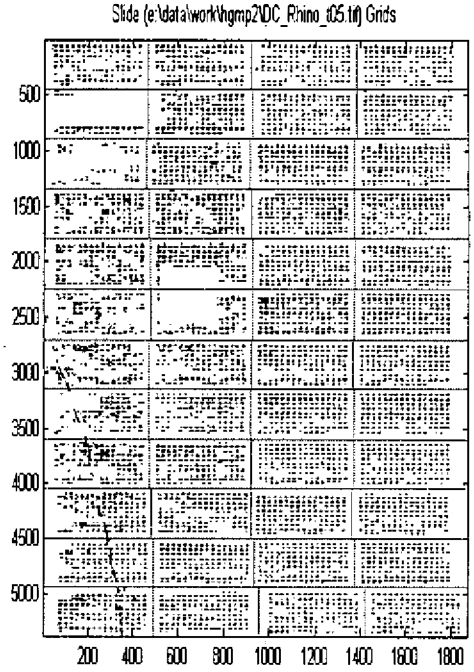
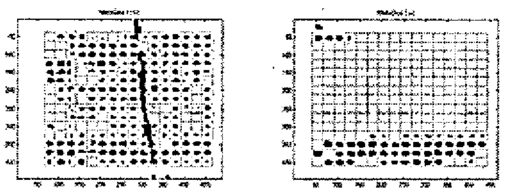


Figure 5: Examples of master blocks

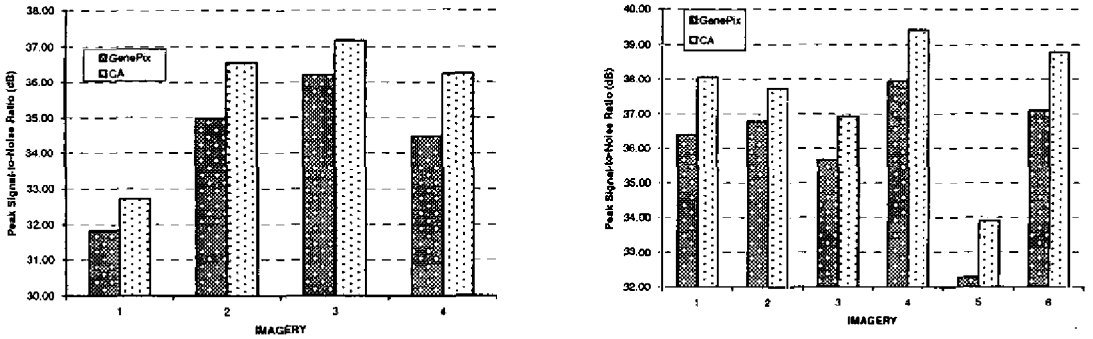
The CA process was able to successfully determine the underlying structure of the previously determined blocks, even when these blocks contained large artefacts (figure 6a) or partial gene spot information (figure 6b). In this case the image contained enough structural information for the CA process to discover the image's layout in a single pass. At present if this was not the case, the gene blocks could be rebuilt by the use of either successfully discovered gene blocks or by a similar experiments image (i.e. the second channel).



a) Example block 1

b) Example block 2

Figure 6: Examples of sub-block discovery



a) Image Set 1 b) Image Set 2

Figure 7: PSNR comparison between GenePix® and CA results.

In order to quantify the performance capabilities of the automated CA framework against that of the human operator, a quality measure is also required which will allow the judgment of how well the calculated templates fit the gene spot position. Both techniques produce a mask that classifies the pixels as belonging to either foreground (the gene spots) or background. By overlaying the mask with the original image a metric can be utilised to quantify the disparity that exists between the two groups of pixels. If the masks fit the genes closely there will be high separation between these groups, any misalignment between them will lead to a diminished separation value.

There are many alternative metrics that can be used here; typically a preferred algorithm is the mean square error (MSE) which is defined as: -

$$MSE = \frac{\sum[f(i,j) - F(i,j)]^2}{N^2} \qquad (1)$$

where $f(i,j)$ represents the source or original imagery that contains NxN pixels and a mask image $F(i,j)$. Error metrics are computed on the luminance signal such that pixel values $f(i,j)$ range between black (0) and white (1). There are, for [0, 255] grey scale images, two disadvantages of the MSE percentage as defined in equation 1. Firstly the denominator is usually very large compared to the numerator, meaning that the reconstruction process's improvement reduces this numerator value, but this might not be observable. Second, the MSE metric is sensitive to the brightness of the original image. Therefore a more objective image quality measurement is known as the peak signal-to-noise ratio (PSNR) [17]. This metric is defined for NxN images with a [0, 1] or [0, 255] grey-scale range, in dB as: -

$$PSNR = 20\log_{10}\left(\frac{1}{RMSE}\right) \qquad (2)$$

where the RMSE (root mean squared error) represents the norm of the difference between the original signal and the mask. The PSNR is the ratio of the mean squared difference between two images and the maximum mean squared difference that can exist between these. Therefore the higher the PSNR value, the more accurately the mask fits the raw imagery. For all images present the proposed framework gave more accurate results.

From figure 7, we directly compare PSNR values determined by GenePix® and CA for the individual images and on average CA has shown a marked 1 – 3dB improvement. Essentially the CA process has consistently outperformed the human expert using GenePix® in terms of gene spot identification.

Discussion & Future Work

We have presented a novel data-driven framework that attempts to improve the full workflow process of microarray image analysis. Specifically, the framework consists of several components that process a microarray image from its raw 16bit scanned representation to the final $\log_2$ ratios and related statistics without human intervention. *Copasetic Analysis* as detailed in figure 2 offers the following advantages over current implementations: *Copasetic Clustering* not only generates historical information allowing accurate image prediction, but also has the computational benefits of processing previously infeasible datasets; *Image Layout* and *Image Structure* perform blind grid alignment on the imagery; *Spatial Binding* reconstructs the determined grid cell positions with accurate spot profiles; *Post Processing* corrects for the background noise and *Final Analysis* computes final microarray statistics. In the experimental part of the paper we demonstrated the potential of *Copasetic Analysis* using direct comparisons between our proposed approach and a commercially accepted process (GenePix®) over the dataset.

In future, we would like to focus on enhancing the current implementations of the framework's component parts. For example the *Image Transformation Engine's* multi-view approach has proved to be beneficial in this initial testing; we are interested in exploring this component's potential in greater detail. Along with this, we intend to develop more sophisticated methods of slide structure reconstruction to further enhance the speed and reliability when processing particularly noisy slides. Finally an important step will be the biological validation of these results; to this end we plan to analyse images containing control spots and a high number of biological repeats.

Acknowledgements

For kindly providing the datasets used, the authors would like to thank both: Paul Kellam from the Dept. of Immunology and Molecular Pathology, University College London, and Dr Su Ling Li from the Institute for Cancer Genetics and Pharmacogenomics, Brunel University, London.

References

[1] M.B. Eisen, P.T. Spellman, P.O. Brown and D. Botstein, Cluster analysis and display of genome-wide expression patterns, in Proceedings of the National Academy of Sciences, 1998, pp. 14863-14868.

[2] P. Kellam, X. Liu, N. Martin, C.A. Orengo, S. Swift and A. Tucker, A framework for modelling virus gene expression data, *Journal of Intelligent Data Analysis*, vol. 6, pp. 265-280, 2002.

[3] S.K. Moore, Making chips, *IEEE Spectrum*, vol. 38, pp. 54-60, 2001.

[4] C.A. Orengo, D.T. Jones and J.M. Thorton, "Bioinformatics: Genes, Proteins & Computers: Mining gene expression data," in First Edition ed., BIOS Scientific Publishers, 2003, pp. 229-244.

[5] Anonymous The Chipping Forcast II, *Nat.Genet.*, vol. 32, pp. 461-552, 2002.

[6] N. Otsu, A threshold selection method from grey level histograms, *IEEE Transactions on Systems, Man and Cybernetics*, vol. 8, pp. 62-66, 1978.

[7] M. Cheriet, J.N. Said and C.Y. Suen, A recursive thresholding technique for image segmentation, *IEEE Transactions on Image Processing*, vol. 7, pp. 918-920, 1998.

[8] Y.H. Yang, M.J. Buckley, S. Dudoit 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.

[9] D. Bozinov and J. Rahnenfuhrer, Unsupervised technique for robust target separation and analysis of DNA microarray spots through adaptive pixel clustering, *Bioinformatics*, vol. 18, 2002.

[10] D. Bozinov, Autonomous system for web based microarray image analysis, *IEEE Transactions on Nanobioscience*, vol. 2, pp. 215-220, 2003.

[11] J. McQueen, Some Methods for Classification and Analysis of Multivariate Observations, in Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability, 1967, pp. 281-297.

[12] X.H. Wang, Istepanian, R. S. H. and Y.H. Song, Application of wavelet modulus maxima in microarray spots recognition, *IEEE Transactions on Nanobioscience*, vol. 2, pp. 190-192, 2003.

[13] M. Katzer, F. Kummert and G. Sagerer, A markov random field model of microarray gridding, in Proceedings of the 18th ACM Symposium on Applied Computing, 2003, pp. 72-77.

[14] K. Fraser, P. O'Neill, Z. Wang and X. Liu, "Copasetic Clustering": Making Sense of Large-Scale Images, in Proceeding of Chinese Academy of Sciences Symposium on Data Mining and Knowledge Management, 2004.

[15] W. Niblack, *An introduction to digital image processing*, London, England: Prentice-Hall international (UK) limited, 1986.

[16] C.J. Dunn, A fuzzy relative of ISODATA process and its use in detecting compact well-separated clusters, *Journal of Cybernetics*, vol. 3, pp. 32-57, 1974.

[17] A.N. Netravali and B.G. Haskell, *Digital pictures: Representation, compression and standards (2nd Ed)*, New York: Plenum Press, 1995.