INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH & DEVELOPMENT

ISSN: 2278 – 0211 (Online)

A Review On Segmentation & Intensity Extraction of cDNA Microarray Images

Anand Jatti
Associate Professor, Instrumentation Department, R.V.C.E., Bangalore, India

C. Amritha
M.Tech 2nd yr., BMSP&I, Instrumentation Department, R.V.C.E., Bangalore

Abstract:

DNA microarray analysis has evolved as the best promising technique for discovering diseases or for accessing mutations caused by exogenous inputs in a given biological organism. The DNA microarray technology allows compressing in a little microscope glass; hundreds of thousands of different DNA nucleotide sequences, and permits to have all of this information on a single image. The analysis of DNA microarray images allows the identification of gene expressions in order to draw biologically meaningful conclusions for applications that ranges from the genetic profiling to the diagnosis of oncology diseases. This paper briefly gives an overview of the basic concepts of DNA microarray based genomic research, methodology and to discuss the pros and cons of some algorithms for cDNA microarray image analysis. A new approach is proposed for DNA-chip image processing by using the combination of Adaptive circle algorithm with K-means algorithm for image segmentation. Intensity extraction can be measured using multiple thresholding technique.

Keywords: *DNA microarray, k-means algorithm, gene expression, adaptive circle algorithm, image segmentation, DNA chips, gridding*

1.Introduction

A DNA microarray (DNA chip or biochip) is a collection of microscopic DNA spots attached to a solid surface [1]. DNA microarrays are generated by either printing pre-synthesized cDNAs (500–2000 bases) or synthesizing short oligonucleotides (20–50 bases) onto glass microscope slides or membranes. cDNAs for microarrays may include fully sequenced genes of known function or collections of partially sequenced cDNA derived from expressed sequence tags (ESTs) corresponding to the messenger RNAs of unknown genes. DNA microarray image processing is used to measure and image the expression levels of large numbers of genes simultaneously or to genotype multiple regions of a genome. In general, thousands of gene-specific probes are arrayed on a small matrix, and this matrix is probed with labelled nucleic acid synthesized from a tissue type, development stage, or other condition of interest. The substrate of a microarray consists of a piece of glass, or a silicon chip, similar to a microscope slide. Onto this substrate, thousands of patches of single-stranded DNA are fixed which are called probes [2].

Figure 1: DNA Chip [20]

The discovery of DNA microarrays has fundamentally altered the way scientists monitor the expression levels of genes. Instead of analysing the expression level of one gene at a time, scientists can simultaneously analyse the expression levels of thousands of genes over different samples [3]. Due to this revolutionary feature, during the last decade cDNA microarrays have been broadly used in many biomedical application areas such as: i) cancer research, infectious disease diagnosis, and treatment (i.e., determination of molecular differences between normal and abnormal cells, classification of tumors, determination of risk factors, and monitoring of treatment during different disease stages); ii) pharmacology research (i.e., determination of correlations between the

genetic profiles of patients and their therapeutic responses to drugs) [3], [4]; iii) toxicology research (i.e., determination of correlations between toxic responses to toxicants and changes in the genetic profiles of objects exposed to such toxicants); and iv) agricultural development. In current-day bioinformatics, even if other techniques (such as proteomics) increasingly gain ground in the aforementioned biomedical applications, cDNA microarrays continue to be used.

DNA microarrays can be broadly classified into single-colour and two-colour arrays. In the former, the control and the experimental specimen are hybridized onto separate arrays, whereas, in the latter, they are hybridized on to the same arrays. The objective is to determine genes that are differentially expressed between two biological states, also known as control (e.g., normal) and experimental (e.g., cancer) samples. These samples are tagged with dyes (Cy3, control) and (Cy5, experimental), known as channels [5]. In particular, this technique can be used for studying the genetic basis of complex diseases. This provides a systematic way to survey the DNA and RNA variations, which could become a standard tool for both molecular biology research and genomic clinical diagnosis, such as cancer diagnosis and type 1 and 2 diabetes diagnoses [6]. A leading use of DNA microarrays is in determining which subset of a cell's genes are expressed, or are actively making proteins, under certain conditions, such as exposure to a drug, toxic material, or malignancy [2]. Microarray technology can facilitate accurate classifications of cancer. Historically tumours have been classified by pathologic examination supplemented by special stains and antibodies. Similar to the way in which DNA sequencing is revolutionizing the field of taxonomy, gene expression profiling is increasing the number of markers useful in tumor classification [14].

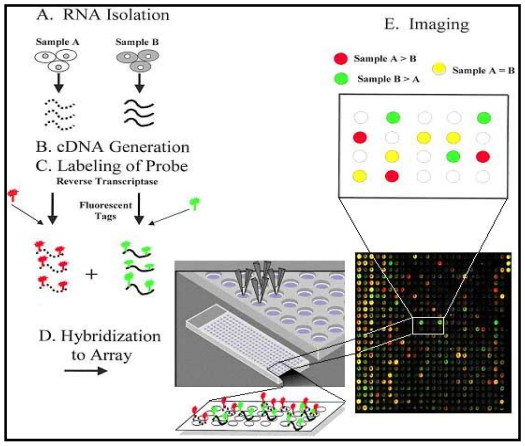

Figure 2: DNA Microarray procedure [20]

Some important aspects of DNA Microarray data analysis are:

- Quality Control and Estimation
- Finding differentially expressed genes
- Cluster Analysis of microarray information
- Gene Regulatory Network
- Data mining for promoter sequences [17]

The motivation for this paper is to review recent image analysis methods that have been applied to gene spot identification in DNA microarray images. Our contribution lies in comparing these algorithmic differences. This paper is expected to benefit for both researchers and academics of Bioinformatics alike by presenting a recent summary of DNA microarray image analysis development.

2.Stages Of Dna Microarray Image Processing

The basic stages in a microarray image analysis are:

2.1.Gridding

It is the process of assigning coordinates to each cell (spot). Basically it is the initial stage of microarray image analysis. The location of each spot is determined. Each spot on the image needs to be addressed unambiguously to a particular spot in the physical experiment slide; that is, to a certain gene. There are two approaches of gridding- a) manual and b) automatic [17]. The automatic gridding approach is widely used. Automatic methods aim at locating orthogonal rows and columns from the image instead of processing each spot individually. A common approach uses horizontal and vertical projection of image. This step creates a square ROI containing the pixels of both the spot and its background.

2.2.Spot Segmentation

It classifies cell-pixels as foreground (spot-pixels) or background [7]. For locating the spot there are different methods having different advantages as disadvantages and hence no single optimized method is available. Each spot in a microarray has both specifically bound DNA and non-specifically bound DNA. The latter is thought to generate a "background" signal that must be subtracted or otherwise removed from the primary

signal. So, background corrections are used to remove non-specific signal that arises from non-specific hybridization, the slide itself, or coatings or other materials on the slide. Many studies have demonstrated that careful removal of this signal can significantly increase the signal-to-noise ratio of a microarray experiment.

Antti Lehmussola et al. [15] evaluated performance of nine microarray segmentation algorithms- Fixed circle (FC), Adaptive circle (AC), Seeded region growing (SRG), Mann–Whitney (MW), k-means (KM), Hybrid k-means (HKM), Markov random field (MRF), Model-based segmentation (MBS) and Matarray (MA). The segmentation performance is derived using measures of probability of error and discrepancy distance. The results demonstrated how the segmentation performance depends on the image quality, which algorithms operate on significantly different performance levels, and how the selection of a segmentation algorithm affects the identification of differentially expressed genes. The k-means algorithm gave nearly error-free segmentation for the good quality images, whereas the Mann–Whitney algorithm produced clearly more erroneous segmentation for the same images.

	Probability of error		Discrepancy distance	
	Good	Low	Good	Low
FC	0.049	0.049	0.027	0.027
AC	0.019	0.192	0.017	0.074
SRG	0.099	0.114	0.037	0.048
MW	0.165	0.162	0.066	0.074
KM	0.000	0.025	0.000	0.041
HKM	0.017	0.020	0.016	0.029
MRF	0.154	0.053	0.063	0.039
MA	0.004	0.031	0.008	0.068
MBS	0.094	0.101	0.052	0.067

Figure 3: Summary of segmentation accuracy of different segmentation algorithms [15]

2.3.Intensity Extraction

It calculates ratios of red to green fluorescence intensities for the foreground and background respectively, which gives the expression levels of the genes. The intensity of each spot is measured in this stage. The intensity of each pixel is proportional to the level hybridization on the specific array location. The total fluorescence can be estimated by using all pixels inside the segmented spots [17]. The results obtained in the information extraction stage are highly dependent on the success of segmentation. Because the division between the background and foreground pixels was done in the segmentation

stage, corrupted segmentation results may unavoidably mislead the total Microarray experiment. Typically, the measured fluorescent intensity is considered as a combination of true signal arising from the hybridization of target molecules, and background component originating from unspecific hybridization, contamination and other artifacts.

The goal of microarray image analysis steps is to extract intensity descriptors from each spot that represent gene expression levels and input features for further analysis. Biological conclusions are then drawn based on the results from data mining and statistical analysis of all extracted features. Using state-of-the-art techniques and algorithms of image processing, we can easily develop efficient algorithms and remove the current pitfalls. Some of the commercial software packages include ScanAlyze [2], [4], GenePix [9], Dapple [4], ImaGene [4], [9], QuantArray [8] etc. for microarray image analysis.

2.Methodology

DNA microarray technology has a high variation of data quality. Therefore, in order to obtain reliable results, complex and extensive image segmentation methods should be applied before actual DNA microarray information can be used for biomedical purpose. Thus, the aim is to solve the main issues in spot validation that arises during an automatic cDNA microarray analysis procedure.

The objective of this work is to use cDNA microarray image processing as a tool for future prospects to diagnose a complex disease (e.g. cancer) at its genetic level by expression level of the genes. The work is mainly concentrated on the image segmentation followed by intensity extraction of cDNA microarray images.

A new methodology for segmentation of cDNA microarray images based on combination of adaptive circle algorithm with k-means algorithm is proposed. Next to the segmentation stage is intensity extraction to calculate the intensities of foreground and background which gives the expression levels of genes. Intensity extraction step can be implemented using multiple thresholding technique.

Microarray image grid and spot position determination is a very important step in the analysis of microarray image because it is the first part we need to do for the analysis and making this part automated and fast is also important for further analysis [16], [18].

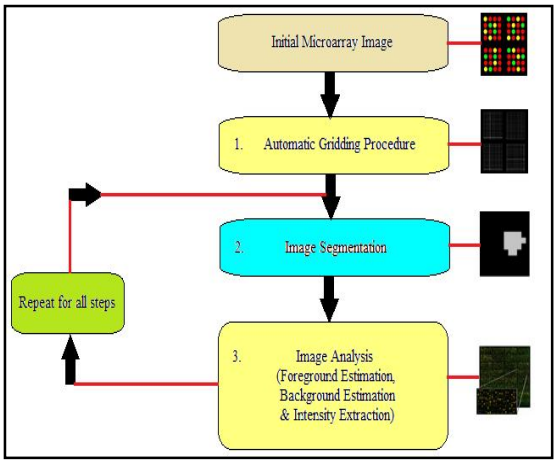

Figure 4: Block Diagram of the methodology

Matlab software is used to implement the algorithm because Matlab is a high performance language which integrates computation, visualization and programming in an easy-to-use environment where problems and solutions are expressed in familiar mathematical notations. Matlab Image Processing Toolbox is a collection of functions that extend the capability of the Matlab numeric computing environment. The toolbox supports a wide range of image processing operations, such as image analysis and enhancement, region of interest operations, linear filtering and filter design [16].

2.1.Algorithm For Gridding

- Read image file
- Crop specified region of interest from the input image
- Display red & green layers to make visualization more intuitive
- Convert RGB image to grayscale for spot finding as it allows us to focus first on spot locations.
- Increase the contrast to adjust the image intensity values
- Create horizontal profile to identify where the centres of the spots are and where the gaps between the spots can be found.
- Estimate spot spacing by autocorrelation to enhance the self similarity of the profile.
- Remove background morphologically from the intensity profile
- Segment peaks and number each peak region

- Locate centers by extracting the centroids of the peaks
- Determine divisions between spots as it provides grid point locations
- Transpose and repeat the analysis for horizontal spacing
- Put bounding boxes around each spot to address each spot individually

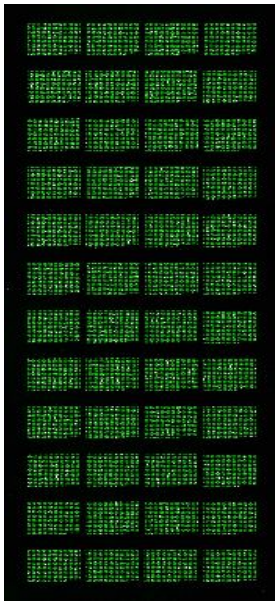

Figure 5: Input Microarray image

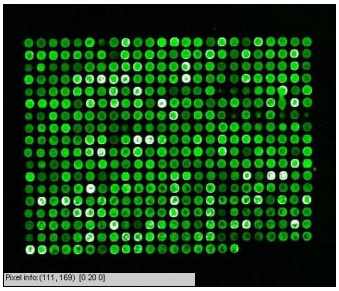

Figure 6: Cropped Image

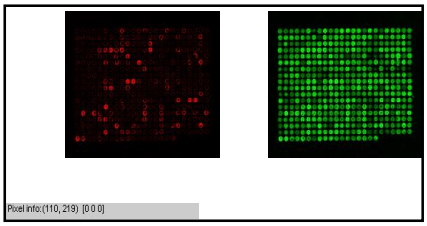

Figure 7: Red & Green layers

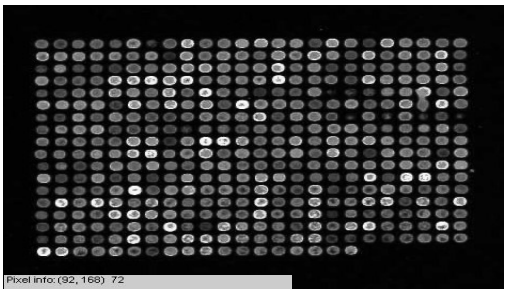

Figure 8: Gray Scale Image

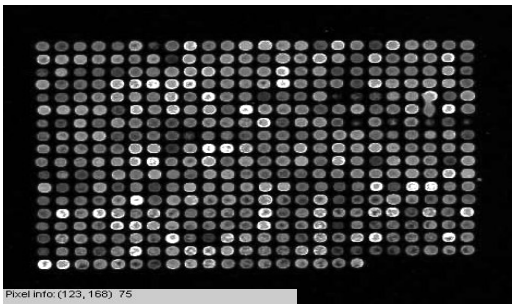

Figure 9:Adjusted Image

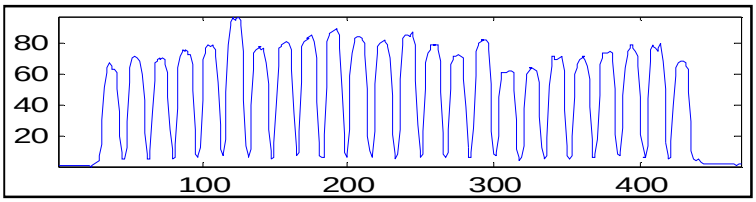

Figure10: Horizontal profile

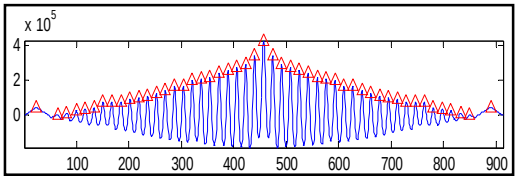

Figure11: Autocorrelation of profile

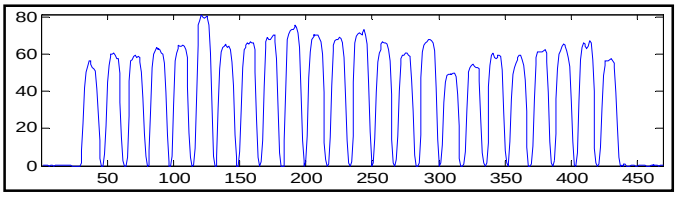

Figure12: Enhanced horizontal profile

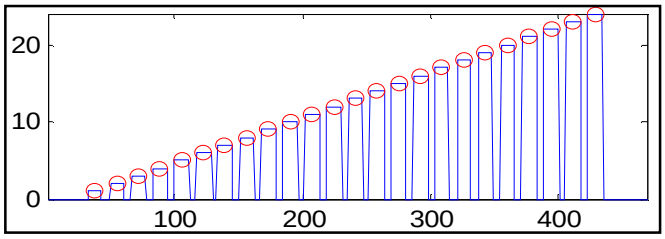

Figure13: Region centers

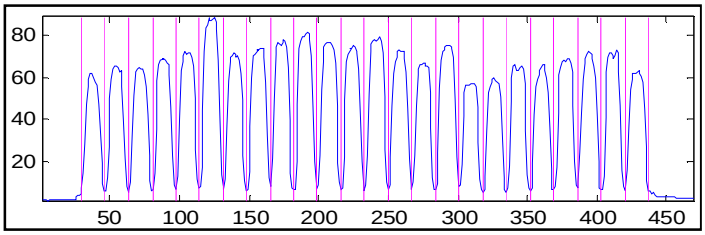

Figure14: Vertical separators

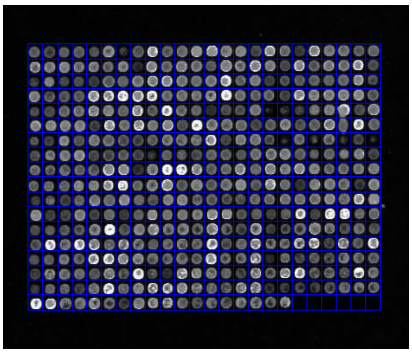

Figure15: Gridding

Next step is intended to develop a new algorithm for spot segmentation for which combination of adaptive circle algorithm with k-means algorithm is proposed. Next to the segmentation stage is intensity extraction for each spot which can be implemented using multiple thresholding technique.

The k-means segmentation algorithm is based on the traditional k-means clustering, and was first used in connection with microarray images. The segmentation result is derived using simultaneous information from both channels. That is, for each spatial location the intensities from both channels are combined as one feature vector. Since the segmentation is used for dividing the image into the regions of foreground and background, the number of cluster centers k is assigned to two. As the initial cluster centers, the pixels with minimum and maximum intensities are selected. All data points

are then assigned into nearest cluster centers according to Euclidean distance. Thereafter, new cluster centers are calculated. Finally, the algorithm is iteratively repeated until the cluster centers stay unaltered.

Adaptive circle (AC) algorithm is also known as variable area segmentation method. This algorithm provides flexibility for the traditional fixed circle. It assumes all spots as circular. However, the radius for each spot is estimated separately. Some software allows users to manually adjust the radius for each spot. Considering the large amount of microarray spots, such approach is extremely laborious and time consuming. An automated version of the adaptive circle is available in the Dapple software, where the radius for each spot is estimated using edge detection. First, the outline of each spot is enhanced using the second-difference approximation of Laplacian. Thereafter, the radius of a circle matching the given enhanced edges is identified with matched filtering.

For intensity extraction step using the thresholding process, individual pixels in the image are marked as "object" pixels if their value is greater than some threshold value (assuming an object to be brighter than the background) and as "background" pixels otherwise. This convention is known as threshold above. The steps are as follows:

- An initial threshold (T) is chosen. This can be done randomly or according to any other method desired.
- The image is segmented into object and background pixels as described above, creating two sets
- The average of each set is computed.
- A new threshold is created that is the average calculated above.
- Now using the new threshold computed above, keep repeating until the new threshold matches the one before it (i.e. until convergence has been reached).

This iterative algorithm is a special one-dimensional case of the k-means clustering algorithm, which has been proven to converge at a local minimum—meaning that a different initial threshold may give a different final result.

3.Conclusion

The popularity of DNA microarrays is derived from the promise that this technology will rapidly advance our understanding fundamental biologic questions. Microarrays have the potential to markedly increase our understanding of not only the process of disease but

also the interactions between biologic organisms and their environment. Several commercial software packages are available to perform computer analysis of the Microarray images. But the methods available at the moment for analyzing the results of microarray experiments are often far from being satisfactory. Many improvements are needed for the already existing algorithms in order to make them more accurate and reliable.

Image analysis of a scanned array consists of calculations designed to identify the location of each feature in the image and to detect and correct flaws and variations in the scan quality.

4.Reference

1. Luca Sterpone and Massimo Violante Politecnico di Torino (2007). A new FPGA-based edge detection system for the gridding of DNA microarray images, Instrumentation and Measurement Technology Conference - IMTC 2007
2. Shadrokh Samavi, Shahram Shirani and Nader Karimi (2006). Real-Time Processing and Compression of DNA Microarray Images, IEEE Transactions on Image Processing, Vol. 15, No. 3, pp. 754-766, March
3. Eleni Zacharia and Dimitris Maroulis (2010). 3-D Spot Modeling for Automatic Segmentation of cDNA Microarray Images, IEEE transactions on Nanobioscience, Vol. 9, No. 3, pp.181-192, September
4. Eleni Zacharia and Dimitris Maroulis (2008). An Original Genetic Approach to the Fully Automatic Gridding of Microarray Images, IEEE transactions on medical imaging, vol. 27, No. 6, pp. 805-813, June
5. Radhakrishnan Nagarajan and Meenakshi Upreti (2006). Correlation Statistics for cDNA Microarray Image Analysis, vol. 3, no. 3, pp. 232-238, July-September
6. Robert S. H. Istepanian (2003). Microarray Image Processing: Current Status and Future Directions, IEEE Transactions on Nanobioscience, Vol. 2, No. 4, pp. 173-175, December
7. Antonis Daskalakis, Dionisis Cavouras, Panagiotis Bougioukos, Spiros Kostopoulos, Christos Argyropoulos and George Nikiforidis (2006). Improving Microarray Spots Segmentation by K-Means driven Adaptive Image Restoration", June 30
8. Peter Bajcsy, Lei Liu and Mark Band (2007). DNA Microarray Image Processing, University of Illinois at Urbana-Champaign (UIUC), DNA Press
9. Emmanouil Athanasiadis , Dionisis Cavouras , Panagiota Spyridonos, Dimitris Glotsos, Ioannis Kalatzis and George Nikoforidis (2007). Segmentation of microarray images using Gradient Vector Flow active contours boosted by Gaussian Mixture Models, 2nd IC-EpsMsO Athens, 4-7 July
10. D. Huang, Tommy W. S. Chow (2005). Efficient Selection of Discriminative Genes From Microarray Gene Expression Data for Cancer Diagnosis, Vol. 52, No. 9, pp. 1909-1918, September
11. Angulo J. and J. Serra (2003). Automatic Analysis of DNA Microarray Images Using Mathematical Morphology, Bioinformatics, vol. 19. No. 5, pp. 553-562

12. Paolo Arena, Luigi Fortuna, and Luigi Occhipinti (2002). A CNN Algorithm for Real Time Analysis of DNA Microarrays, IEEE Transactions on Circuits and Systems—I: Fundamental Theory and Applications, Vol. 49, No. 3, pp. 335-340, March
13. M´onica G. Larese and Juan C. G´omez (2008). Automatic Spot Addressing in cDNA Microarray Images, JCS&T, Vol. 8, No. 2, pp. 64-70, July
14. Edward K. Lobenhofer, Pierre R. Bushel, Cynthia A. Afshari, and Hisham K. Hamadeh (2001). Progress in the Application of DNA Microarrays, Environmental Health Perspectives, Vol. 109, No. 9,pp. 881-891, September
15. Antti Lehmussola, Pekka Ruusuvuori and Olli Yli-Harja (2006). Evaluating the performance of microarray segmentation algorithms, Bioinformatics Original Paper, Vol. 22, No. 23, pp. 2910–2917
16. Basim Alhadidi, Hussam Nawwaf Fakhouri and Omar S. AlMousa (2006). cDNA Microarray Genome Image Processing Using Fixed Spot Position, American Journal of Applied Sciences 3 (2): pp. 1730-1734
17. Rezaul Karim, Md. Khaliluzzaman and Sohel Mahmud (2011). A Review of Image Analysis Techniques for Gene Spot Identification in cDNA Microarray Images, ICNIT, 2nd International Conference, pp. 36-41
18. Zhang Yao and Wu Shunxiang (2012). Statistsics-adaptive method for cDNA Microarray images gridding, 2012 Fourth International Conference on Digital Home, IEEE Computer Society, pp. 380-383
19. A.K. Jain (1989). Fundamentals of Digital Image Processing, Prentice-Hall
20. S.C. Rastogi, N. Mendiratta and P. Rastogi (2010). Bioinformatics: Methods and Applications- Genomics, Proteomics and Drug Discovery, 3rd ed., PHI Learning Pvt. Ltd.
21. R. Gonzalez and R. Woods (2002). Digital Image Processing, 2nd ed. Upper Saddle River, NJ: Prentice-Hall