A Recent Development in Image Analysis of Electrophoresis Gels

Xiangyun Ye[1,2,3] Ching Y. Suen[1] Mohamed Cheriet[1,2] Eugenia Wang[3]

1, Centre for Pattern Recognition and Machine Intelligence
Concordia University, Suite GM606, 1455 de Maisonneuve Blvd. West
Montréal, Québec H3G 1M8, Canada

2, Laboratorie d'imagerie, de vision et d'intelligence artificielle
École de Technologie Supérieure, de l'Université du Québec
1100, rue Notre-Dame Ouest, Montréal, Québec H3C 1K3, Canada

3, Bloomfield Centre for Research in Aging, Lady Davis Institute for Medical Research
Sir Mortimer B. Davis - Jewish General Hospital, Department of Medicine, McGill University
3755, Côte Ste-Catherine Road, Montréal, Québec H3T 1E2, Canada

email:{xyye,suen}@cenparmi.concordia.ca, cheriet@gpa.etsmtl.ca, cznu@musica.mcgill.ca

Abstract

Electrophoresis is an electrochemical separation process in which molecules, such as protein or RNA/DNA fragments, are made to migrate through a specific substrate, such as a polyacrylamide gel, under the influence of an electric current. The technique has a wide range of applications, in DNA sequencing and in studying variation in the identity and amount of proteins obtained from different sources. Techniques of image analysis and pattern recognition can be used to extract qualitative as well as quantitative information from the images, and spare human beings from voluminous, tedious image interpretation. More importantly, computerized data handling and interpretation provide accuracy and rapid speed without human errors. Here, we report the application of a newly developed system to the analysis of biological specimens that have undergone gel electrophoresis. The result of this application shows the capability to identify unique banding patterns of cDNA profiles, which paves the way for future full-scale investigation in the use of pattern recognition principles in biomedical information handling and interpretation.

I. Introduction

Electrophoresis is an electrochemical separation process in which biological molecules, such as protein or RNA/DNA fragments, are made to migrate through a specific substrate, usually a polyacrylamide gel, under the influence of an electric current. The technique can be used to separate mixtures of molecules on the basis of their molecular size, by making use of their electric charge differences. This difference, under the electric field charge, causes individual biological materials of the same size to migrate to discrete positions within the bed of polyacrylamide medium. Collection of these multiple positions in a linear fashion presents the separation of mixed biological materials into specific electrophoresis profiles. It has wide application in DNA sequencing, and in studying variation in the qualitative and quantitative separation of proteins or nucleic acids obtained from different sources. Scientists use electrophoresis to derive information about the substances under study, such as comparing the composition of samples, or quantifying the amount and properties of the different constituents present in a collection of samples. Electrophoresis has many variants, including one or two-dimensional electrophoresis, electrofocusing, isotachophoresis and several forms of immunoelectrophoresis[1]. It is almost 200 years since Ferdinand Frederic Reuss first observed through a microscope in 1807 the migration of colloid particles in an electric field, which might be regarded as the first electrophoretic separation. Although many clever developments and applications have appeared, it is only in the last 25 years that the technique has become widely used. There are a vast number of variations in electrophoretic techniques and in principle almost any problem that involves either mixtures of biological

macromolecules or changes in them can be studied. Since digital image processing and pattern recognition techniques have been introduced to analyze the electrophoresis gels, scientists are able to make more precise and reliable qualitative and quantitative studies[2].

In general, biological objects present great challenges to computer interpretation due to their irregular and variable shapes. The images to be analyzed in biometry may come from microscopy, medical scanning systems, electrophoresis, remote sensing, or simply from photographing illuminated objects. With the progress of image analysis techniques in recent years, the computer is becoming an important tool for scientists and professionals in data acquisition and interpretation[3,4]. The digitized analysis of electrophoresis gels emerges as one of the most important applications, to reduce human error and increase the speed of data evaluation. To analyze the electrophoresis gels accurately and quantitatively, images are produced containing bands or spots reflecting the amounts and characteristics of individual components. The digitization of gel images provides a way to quantify information objectively. The image processing and analysis techniques convert the information carried by electrophoresis gels into numerical, graphic, tabular and visual formats that are reliable, accurate, presentable, and consistent. Hence computers enable biologists to avoid subjective, tedious image interpretation. Meanwhile, statistical pattern recognition techniques also have many potential applications, including taxonomy, epidemiology, population studies and others. Due to the ever-growing demands in analyzing biological information, not only institutes such as *the National Human Genome Research Institute (NHGRI)*[5], *the National Institute of Mental Health (NIMH)*[6] and *Biomathematics and Statistics Scotland (BioSS)*[7] have committed themselves to investigating the methodology, but also many commercial companies are developing accurate, user-friendly software packages. There are many such software packages available to import and analyze a wide range of one- and/or two-dimensional gel electrophoresis images, including *GelPro* from *Media Cybernetics Inc.*[8]; *EDAS 120* (Electrophoresis Documentation and Analysis System 120) from *Eastman Kodak*[9]; *UVIdoc* from *UVItec Inc.*[10], *GelExpert* from *Nucleotech Inc.*[11]; *Phoretix 1D/2D* gel analysis package from *Phoretix International*[12]; *Biotech software* from *Advanced American Biotechnology & Imaging*[13], *AlphEase*™ from *Alpha Innotech Inc.*[14], etc.

In this paper, we present our work using image analysis principles in biometry to extract qualitative and quantitative information, which may consist of binary presence/absence categories or measures of object location, length or area, shape statistics, of gel-electrophoresis data. We use signal processing methods including histogram transformations[15], linear and non-linear filtering, and thresholding[16,17] to calibrate and enhance the raw images. Mathematical morphology methods[18,19] such as opening, closing, Top-Hat, and Hit-or-Miss transforms are used to measure morphological features of the objects; measurements such as lengths, areas, histograms, *etc.* are extracted[19,20] and interpreted using stereology, shape statistics or classification methods[21]. We present here a concise system capable of converting the raw electrophoretic pattern of cDNAs into digitized information, and the mathematical modelling allowing us to analyze the digitized images by pattern recognition principles.

II. Image Acquisition

Figure 1(a) shows an example of a one-dimensional (1D) electrophoresis gel, which is produced to identify differences in composition of different samples. Six mixtures of radioactively labeled fragments are positioned in distinct wells along one side of the gel. Each mixture migrates down the gel, and DNA fragments of different sizes produce separate bands. By placing a photographic plate over the gel, the radioactive emissions produce an autoradiograph in which each band corresponds to the presence and amount of each specific DNA fragment. Thus, digital images can be obtained *via* commercially available image capture devices, such as video cameras or scanners. The intensity of each pixel is then the time reflection of the emitted radiation after radioactive labeling, by the photographic imprinting onto the X-ray film.

Image acquisition is the key to subsequent analysis, which should have optimal resolution to capture enough data points within the smallest bands of interest to reflect the dynamic range of the biological materials confined to a particular region on the gel. The most commonly used measurement of the concentration of substances is Optical Density (*OD*) [17], which is a logarithmic function of brightness (Fig.2) and can be used to determine the concentration of substances in an image. Different devices can measure *OD*, which reflects relative concentrations. Since the response range of different instruments and media may limit the accuracy of quantification, it is very important to choose imaging devices with different linear *OD* ranges according to specific applications. While low light CCD (charge coupled device) cameras are ideal for sensitive detection of fluorescence and chemiluminescence banding images from gels, storage phosphor imagers can detect and quantify ionizing radiation produced from films exposed to gels and blots. While densitometers detecting transmitted light can be very effective with one- or two-dimensional gel images on film media, document scanners are a rapid and cost-effective solution to obtain digitized images from films derived from various types of protein and nucleic acid blots.

III. Image Enhancement

The background of most images varies significantly throughout the entire gel matrix, either because of the presence of noise, blurring, or a warping/distortion of the image frame due to imperfect experimental conditions for the actual running of electrophoresis through the polycrylamide gels. This means that background subtraction should account for differences between lanes, as well as along the length of each lane. For example, the intensity of background varies from one region to another region in Fig. 1(a), which makes it difficulty to compare similar features in different parts of the image. In addition, the bands are not horizontally aligned because of non-uniform migration, known as 'smile', on the gel. To make interpretation easier, it is necessary to remove the background intensity variation and distortion or 'smile' of the bands at the same position in different lanes. The goal of this method is to be able to obtain accurate quantification even from distorted gel images, when the different lanes do not show perfectly aligned gel bands.

Image processing for purposes of enhancement can be performed in either the spatial domain (the array of pixels) or other domains. Background subtraction can be done using several different methods for complete control of variations across a gel. The background image caused by non-uniform illumination of the image can be acquired either by physically repeating the imaging procedure on a neutral standard, or by estimating directly from the acquired image. If an image has been spoiled by noise produced by unbounded radioactive isotopes permeating the whole gel, blurring forms that are either known or can be estimated, linear or nonlinear filters such as the *Wiener filter* or *maximum entropy restoration* can be utilized to optimally restore the original image[15,22]. To obtain an estimate of the background image, a number of points can be selected to perform least-square fitting of a function that approximates the background. Another approach used to remove the gradual variation in overall brightness is applied in frequency space. It assumes that the background variation in the image is a low-frequency signal, and can be separated in frequency space from the higher frequencies that define the features presented. The low-frequency components corresponding to background can thus be removed by a high-pass filter. Other techniques for removing the background include nonlinear filtering[23] and mathematical morphology[18], which have been widely used in solving image processing problems. which were difficult to solve by linear filters.

In our system, morphological closing of the gel image is used to estimate the background trend. By choosing an appropriate structuring element to describe the bands, which are relatively small compared with the background, regions larger than the structuring element are taken as background. We use one of the most widely used shapes for structuring element, a disc, to remove the slowly-varying background. The radius of the disc is dependent on the parameters used for spot and band detection. After the background is subtracted from the raw image, the bands become more distinct. For a raw image $F = \{f(x, y)\}$, the background removal procedure can be described as:

- Background estimation:
  $Background\ (x, y) = (f \bullet B)(x, y)$ for all $(x, y) \in F$
- Background subtraction:
  $f'(x, y) = f(x, y) - Background\ (x, y) = (f - (f \bullet B))(x, y)$ for all $(x, y) \in F$

This is known as a 'Top-Hat Transform'. An example of background removal by morphological filtering is shown in Fig. 1(b). A one-dimensional illustration, in Fig. 3, shows background (red line), band edges (pecked lines) and band profile (green). Fig. 4 shows the comparison between the raw and the enhanced images after background removal.

To make qualitative and quantitative comparison between profiles of lanes and positions of bands/spots on different gels, it is necessary to remove distortions. In general, human interpreters are good at compensating for distortions subjectively. However, human interpreters are less able to detect subtle differences between samples. By manipulating the digital images and removing the distortions, computers become more powerful and accurate than human eyes. As for the correction of 'smile's, Clasbey and Wright[24] proposed an algorithm assuming that the distortion may be described by a smooth continuous function, and that it can be estimated from the angles of the individual bands. Given a function $A(x, y)$ which specifies the angles that bands subtend with the x-axis in different part of a gel, in order to bring the bands into alignment with the x-axis, the transform $(x, B(x, v)) \rightarrow (x, v)$ is carried out, in which $B(u, v) = v + \int_0^u \tan A(x, B(x, v))dx$.

IV. Measurement Techniques & Image Analysis

The extraction of qualitative and quantitative information is one of the ultimate goals of gel image analysis. The analysis of a one-dimensional gel is a computational problem in three dimensions: width, thickness and optical density. The analysis of a two-dimension gel involves additional consideration for shape statistics including compactness, elongation, moments[25,26], perimeters[27], curvatures[28], principle axes[29], *etc.* Recent techniques such as mathematical morphology and stereology[30,31] can be used to study the size distribution, as well as estimate the volume of objects from gray-level information.

Take a one-dimensional gel image as an example: each peak in the 1D profile of the individual lane defines a closed area, by dropping vertical lines at the two borders of the peak. Whatever is outside this sealed area is "noise" or background, and therefore is not measured. By calculating the area of the closed area, we may estimate the amount of substances in each band, and the molecular weight of each band can be estimated by considering the position relative to a predefined reference band. In order to analyze the presence and volume of bands in each lane, the bands with the same electrophoretic mobility property in different lanes must be aligned with each other according to their approximate molecular weight. Although distortion has been corrected in the image enhancement procedure, afterwards the corresponding bands may still have minor differences in their physical positions, which make direct comparison impossible. A matching method used in stereo vision can be used to make the final alignment[32,33]. The lanes and bands can be automatically identified by analyzing both the horizontal and vertical edge images (Fig. 5). The location of a lane is detected by selecting a pair of pixels with the highest edge in a local neighborhood, and the bands are detected by picking up the peaks of the one-dimensional profile (Fig. 6). The height of the curve for band detection, at any given point, is the mean of the intensities of a given row of pixels in the marked lane.

The measurement of bands or spots can be calibrated either in intensity or quantity. Since many photographic methods to reveal the bands are not linear, quantity calibration is the preferred method to allow all results to be calculated from a standardized curve, created from the measured values of standard bands or spots with known values. Frequently, lanes composed of samples prepared from serial dilution of biological materials are used as internal standards for this purpose.

To perform quantitative analysis of the volume and shape of a band, it is necessary to segment the images into objects *vs.* background. Basic segmentation methods include thresholding[34], edge-based and region-based segmentation, and Bayesian approaches[35]. Fig. 7 shows our result of lane and band detection based on Ostu's thresholding[36] method, which maximizes the between-class variance while minimizing the within-class variance. The bands in different lanes are aligned according to the correlation between profiles of intensities at each band. In Fig. 7, we are able to discern a unique band in the middle of the fourth lane, at the threshold calculated by Otsu's algorithm. Experiments on a large set of gel images have shown the ability of the system to identify bands at different levels of interest.

V. Clustering Analysis

The promise of computer application to electrophoresis gel image analysis is to provide not only accurate and reliable quantification, but also the ability to analyze statistically large quantities of samples. One of the important applications in population study is clustering analysis based on similarity measurements[21]. The clustering problem can be described as finding "natural groupings" in a set of data. This question actually involves two separate issues: how to measure the similarity between samples, and how to partition sets of samples into clusters among a large number of data entry points, such as bands and spots on electrophoresis gels.

The most obvious measure of similarity and dissimilarity between two samples is the distance between them. One way to begin a clustering investigation is to define a suitable distance function, and compute the matrix of distances among all pairs of samples. If the distance is a good measure of dissimilarity, then one would expect the distance between samples in the same cluster to be significantly less than the distance between samples of different clusters. Formal clustering procedures use a criterion function based on distance measurement to seek the grouping that maximizes the criterion function[21]. The K-means and ISODATA algorithms, and Vector Quantization are examples of clustering methods based on discriminant analysis. Other clustering methods include theoretical and practical achievements in neural networks[37], such as self-organizing feature mapping[38] and Adaptive Resonance Theory (ART map)[39], which can simulate a process of learning from task examples. In the context of electrophoresis gel image analysis, the samples can be clustered by either statistical clustering algorithms or neural network approaches, with the measures of similarity estimated by giving different emphasis to the presence or abundance of a given DNA fragment. In our system, a similarity measurement taking both the presence and abundance of bands is used:

$$Similarity(L_i, L_j) = \sum_{k=1}^{K} w_k \cdot Similarity_k(L_i, L_j)$$

Here, L_i and L_j are two feature vectors corresponding to the two lanes in comparison, $Similarity_k(L_i, L_j)$ refers to various measurements of similarity, including similarity for binary data representing the presence of the bands[40], and correlation-based similarity for analog data representing the abundance of the bands[41], and w_k is the corresponding weight of each method. By tuning the weights of different similarity measurements, the system is able to pay selective attention to qualitative or quantitative features of the bands.

VI. Interesting Aspects, Challenges and Future Directions

An increasing number of commercial software packages have become available in recent years. Biologists use them because they provide not only accurate quantification, but also the capability of retrieving useful information from large databases. Computers are able to analyze large numbers of samples in an accurate and objective manner, beyond the ability of detection and data processing by human operators. By introducing the newest developments of techniques in image analysis and pattern recognition, such as neural networks, fuzzy logic, expert systems, information retrieval, *etc.*, we shall be able to interpret and analyze results of biological activities that follow nonlinear and complex traits of regulation. However, it is important to mention that besides good image capturing devices and properly applied digital image analysis techniques, the quality of the initial raw data remains the key to precise and accurate analysis. This calls for the repeated evaluation of electrophoresis techniques, for example obtaining well-separated bands and lanes, and appropriate gel loading with minimal distortion during the actual running of the gel.

Meanwhile, with the consistently emerging biological technologies, different types of images will need new methodologies of data interpretation. Genomics, the study of the location and function of the 100,000 genes in every human cell, is increasingly becoming the focus of medical and pharmaceutical research. Pharmaceutical and biotechnology companies are turning to genomics in their tasks of gene discovery for effective new diagnostic tests, and new drug development. As a new powerful tool, microarray technology is currently being widely adopted in genomics analysis[42,43]. Data from a single microarray experiment provides researchers with the ability to accurately measure gene expression levels in thousands of samples. Microarray technology necessitates the application of digital image analysis and recognition techniques, so that the different gene expression in samples can be identified qualitatively and quantitatively in a rapid and accurate fashion. While image analysis techniques are continuously growing in power and utility, we look forward to the emergence of an advanced systematic approach, *i.e.*, the convergence in the intelligent recognizing abilities of computers with those of human operators, which will enable the computers to assist scientists in handling large database of biological information with ease.

Acknowledgement:

The authors would like to thank Mr. Gregory Proestou for preparing the differential display gels.

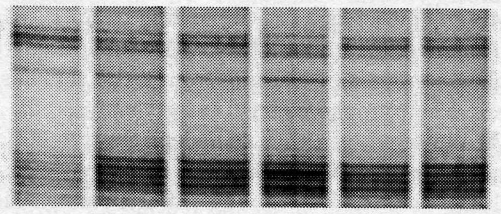

Figure 1 (a) Raw image from electrophoresis gel

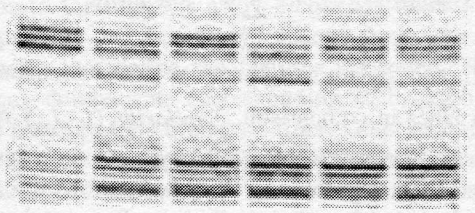

Figure 1 (b) Enhanced image after background removal and enhancement

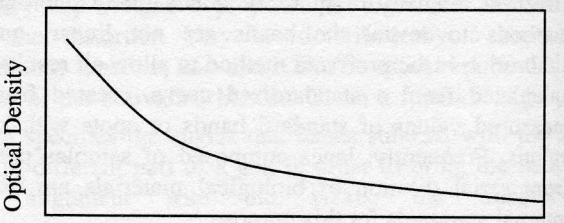

Figure 2 Optical Density related to Brightness[17]

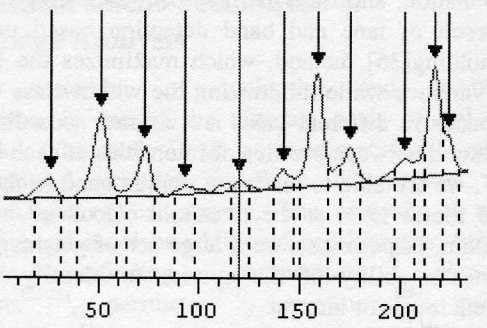

Figure 3 One-dimensional illustration of background removal and band detection

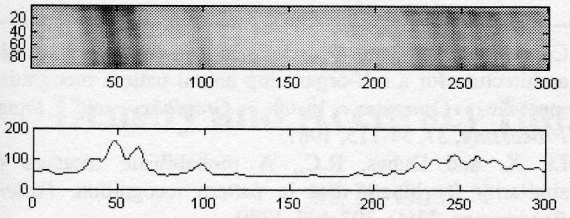

Figure 4 (a) One-dimensional profile of raw image

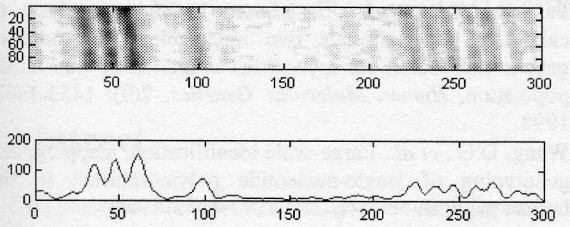

Figure 4 (b) One-dimensional profile of enhanced image after background removal and contrast manipulation

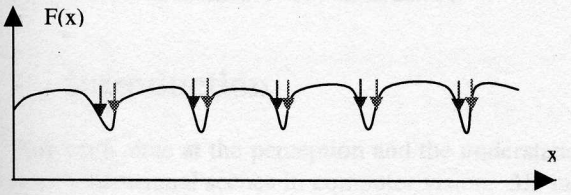

Figure 5 (a) Lane detection by analyzing the horizontal edge

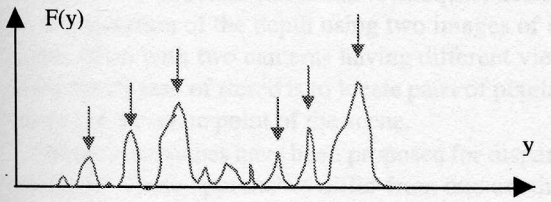

Figure 5 (b) Band detection by picking out the peaks in one-dimensional profile

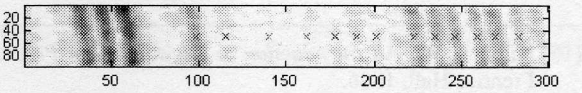

Figure 6(a) Raw image of lane **1** and detected bands

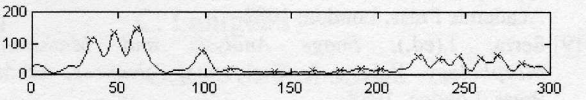

Figure 6(b) Band detection in one-dimensional profile

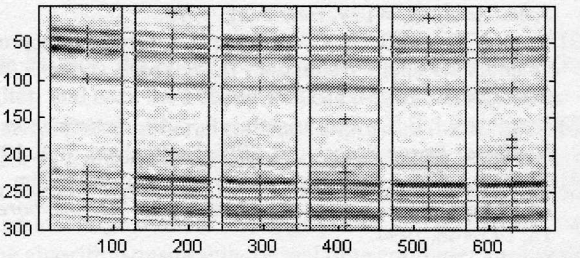

Figure 7 Band alignment for distorted gels

References

[1] Hames, B.D. and Rickwood, D.(eds.), *Gel Electrophoresis of Proteins: A Practical Approach*, IRL Press Ltd, London, 1981.

[2] Horgan, G.W. and Glasbey, C.A., Uses of digital image analysis in electrophoresis, *Electrophoresis*, **16**, 298-305, 1995.

[3] Glasbey, C.A. and Horgan, G.W., *Image Analysis for the Biological Sciences*, Wiley, Chichester, 1995.

[4] Glasbey, C.A. and, Berman, M., A review of image analysis in biometry, *in Advances in Biometry: 50 Years of the International Biometric Society*, (P. Armitage and H. A. David, eds), Wiley:New York, 385-403, 1996.

[5] http://www.nhgri.nih.gov : National Human Genome Research Institute, National Institutes of Health

[6] http://www.nimh.nih.gov :National Institute of Mental Health, National Institutes of Health

[7] http://www.bioss.sari.ac.uk : Biomathematics and Statistics Scotland

[8] http://www.mediacy.com : Media Cybernetics Inc.

[9] http://www.kodak.com : Eastman Kodak Inc.

[10] http://www.uvitec.demon.co.uk : UVItech Inc.

[11] http://www.nucleotech.com : Nucleotech Inc.

[12] http://www.phoretix.com : Phoretix International

[13] http://www.aabi.com : Advanced American Biotechnology & Imaging

[14] http://www.alphainnotech.com/software.htm : Alpha Innotech Inc.

[15] Rosenfield, A. and Kak, A.C., *Digital Picture Processing* (2nd edition), Academic Press, San Diego, 1982.

[16] Niblack, W., *An Introduction to Digital Image Processing*, Prentice Hall, 1986.
[17] Russ, John C., *The Image Processing Handbook*, Boca Rato: CRC press, 1992.
[18] Serra, J., *Image Analysis and Mathematical Morphology*, Academic Press, London, 1982.
[19] Serra, J.(ed.), *Image Analysis and Mathematical Morphology. Volume 2: Theoretical Advances*, Academic Press, London, 1988
[20] Haralick, R., Sternberg, S., and Zhuang, X., Image analysis using mathematical morphology. *IEEE Trans. PAMI*, 9(4):532-550, 1987.
[21] Duda, R.O., and Hart, P.E., *Pattern Classification and Scene Analysis*, Wiley, New York, 1973.
[22] Tang, Y.T. and Suen, C.Y., Image transformation approach to nonlinear shape restoration, *IEEE Trans. Systems, Man and Cybernetics*, **23**, 155-171, 1993.
[23] Dougherty, E.R. and Astola, J., *An Introduction to Nonlinear Image Processing*, SPIE, Bellingham, WA, 1994.
[24] Glasbey, C.A. and Horgan, G.W., An algorithm for unwarping multi-track electrophretic gels, *Electrophoresis*, **15**, 143-148, 1994.
[25] Hu, M., Visual pattern recognition by moment invariants, *IEEE Trans. Information Theory*, 179-187, 1962.
[26] Prokop, R.J. and Reeves, A.P., A survey of moment based techniques for unoccluded object representation and recognition, *Computer Vision, Graphics, and Image Processing: Graphical Models and Image Processing*, **54**(5), 438-460, 1992.
[27] Koplowitz, J. and Bruckstein, A.M., Design of perimeter estimators for digitized planar shapes, *IEEE Trans. PAMI*, **11**, 611-622, 1989.
[28] Nishida, H., Curve description based on directional features and quasi-convexity/concavity, *Pattern Recognition*, **28**(7), 1045-1051, 1995.
[29] Lin, J., Universal principal axes: an easy-to-construct tools useful in defining shape orientations for almost every kind of shape, *Pattern Recognition*, **26**(3), 485-493, 1993.
[30] Stoyan, D., Kendall, W.S., and Mecke, J., *Stochastic Geometry and its Applications*, Wiley, Chichester, 1987.
[31] Stoyan, D., Stereology and Stochastic Geometry, *International Statistical Review*, 58, 227-242, 1990.
[32] Haralick, R.M. and Shapiro, L.G., *Computer and Robot Vision, Volume 1*, Addison-Wesley, Reading, Massachusetts, 1992.
[33] Birchfield, S. and Tomasi, C., A pixel dissimilarity measure that is insensitive to image sampling, *IEEE Trans. PAMI*, **20**(4), 1998.
[34] Sahoo, P.K., Soltani, S., Wong, A.K.C., and Chen, Y.C., A survey of thresholding techniques, *Computer Vision, Graphics, and Image Processing*, **41**, 233-260, 1988.
[35] Pal, N.R., Pal, S.K., A review on image segmentation techniques, *Pattern Recognition*, **26**(9), 1277-1294, 1993.
[36] Ostu, N., A thresholding selection method from gray-level histograms, *IEEE Trans. Systems, Man & Cybernetics*, **9**(1), 62-66, 1979.
[37] Zurada, J.M., *Introduction to Artificial Neural Systems*, West Publ., New York, 1992.
[38] Kohonen, T., Automatic formation of topological maps in self-organizing systems, *Proc. of the 2nd Scandinavian Conf. on Image Analysis*, Helsinki, Finland, 214-220, 1981.
[39] Carpenter, G.A., and Grossberg, S., A massively parallel architecture for a self-organizing neural pattern recognition machine, *Computer Vision, Graphics, and Image Processing*, **37**, 54-115, 1987.
[40] Li, X. and Dubes, R.C., A probabilistic measure of similarity for binary data in pattern recognition, *Pattern Recognition*, **22**(4), 397-409, 1989.
[41] Pratt, W.K., Correlation techniques of image registration, *IEEE Trans. Aerospace and Electronic Systems*, **10**, 353-358, 1974.
[42] Pastinen, T., *et al.*, Array-based multiplex analysis of candidate genes reveals two independent and additive genetic risk factors for myocardial infarction in the Finnish population, *Human Molecular Genetics*, **7**(9), 1453-1462, 1998.
[43] Wang, D.G., *et al.*, Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome, *Science*, **280**, 1077-1082, 1998.