2D gel image processing and analysis for proteomics

Diao, Xiaoning

2005

Diao, X. (2005). 2D gel image processing and analysis for proteomics. Master's thesis, Nanyang Technological University, Singapore.

https://hdl.handle.net/10356/4216

https://doi.org/10.32657/10356/4216

Nanyang Technological University

2D Gel Image Processing and Analysis for Proteomics

Diao Xiaoning

School of Electrical & Electronic Engineering

A thesis submitted to the Nanyang Technological University

in fulfillment of the requirement for the degree of

Master of Engineering

2005

Acknowledgements

I would like to take this opportunity to express my gratitude to those who have provided me with help and encouragement during the development of the present works.

I want to express my sincere gratitude to my supervisor, Prof. Mao Kezhi, for being so encouraging and supportive in spite of his busy schedule. His willingness to help in time of difficulties has enabled me to go on and finish this research project successfully.

Appreciation should also be expressed to Prof. Soh Yeng Chai, Prof. Koh Tong San and Prof. Jagath Rajapakse, who had given much advice and help to me.

Table of Contents

List of Figures

List of Tables

Summary

In proteomics, two-dimensional gel electrophoresis (2DGE) is the most commonly used technique to separate the complex mixtures of proteins, and image processing and analysis plays an important role in 2DGE. We found that some spots which correspond to proteins might be missed when the watershed algorithm was used to detect the spots. Based on the properties of such spots, we proposed a clustering based method for spots detection. This method regards the pixels in the 2DGE image as cluster data and employs the subtractive clustering technique to detect the cluster centers, which can be used as the internal markers in watershed segmentation. With the new markers, more potential protein spots can be detected.

To model the saturated regions of protein spots, we proposed a new method which uses axis-parallel ellipses as covering models on the saturated region. Particle Swarm Optimization (PSO) and subtractive clustering are used to construct the model. By using the clustering method, we could obtain good estimation of the positions of the potential merging spots in **a** saturated spot. PSO is used to minimize the covering error and find the best covering ellipses for those protein spots. The Combination of all the detected ellipses makes up the model of the saturated spot region. Our simulations will show satisfying results of the new covering model.

1. Introduction

1.1 Motivation

Proteomics is the science of protein study. It has become increasingly important after the completion of the human genome sequencing. In proteomics, two-dimensional gel electrophoresis (2DGE) is the most commonly used technique to separate proteins in a biological sample on a gel. The advantage of two-dimensional gel electrophoresis is that it enables the investigation of the protein expression for thousands of proteins simultaneously from very small amounts of material. The result of 2D gel electrophoresis is an image where each black spot represents an accumulation of proteins. And analysis for the 2DGE image plays a significant role in the biological research.

Spots detection and modeling are important because the process results of them are the guarantee for the further analyses. They are in the first part of the analysis for images and data gathered from the protein samples, so the results of both processes should describe the presence and distribution of the protein samples accurately, which will ensure the accuracy of the further processes such as spots matching and comparison. Most of the existing commercial software with traditional techniques for two-dimensional gel electrophoresis image analysis requires some user intervention, which is tedious and time consuming. Therefore, the goal of this study is to develop better automatic spots detection and modeling methods.

1.2 Objectives

Most of the spots detection methods from traditional image segmentation techniques are very sensitive to noise in the 2D gel electrophoresis images. Currently, the most popular technique for spots segmentation and spots detection is the watershed transform. This method segments the 2DGE image into different regions, assuming that each region contains only one protein spot. But based on the observation on the 3D profile of some regions in the 2DGE image, this assumption is not always true. **As** a result, some protein spots might be missed if watershed algorithm is used. Therefore, we will explore some new methods to alleviate this problem.

Furthermore, after the spots detection, protein spots should be characterized and represented as a list of features for further analysis such as spots matching. Modeling methods can be used to get the respective parameters of the spots to the models. The methods of Gaussian and diffusion modeling assume perfect diffusion in the 2D gel medium. But in practice, the diffusion process is not perfect and spots may be formed with unpredictable and unusual shapes. Thus, the modeling result of large saturated spot might not be precision enough for further use. Therefore, we will try to find new modeling methods for saturated spots, based on the analysis for the shape characteristics of these spots.

1.3 Major Contribution of the Thesis

In this thesis, some traditional spots detection methods are reviewed. They are very sensitive to noise in the 2D gel electrophoresis images. Currently, the most popular technique for spots segmentation and spots detection is the watershed algorithm due to its robustness to noise. The problem of the watershed algorithm will be illustrated when it is used in the protein spots detection. From the 3D profile and the section image, it is obvious that some spots are missed with watershed detection. Based on analysis of the characters of such spots, we proposed a clustering based method for spots detection. If the pixels in the 2D gel electrophoresis images are regarded as cluster data, the subtractive clustering will be used to detect the cluster centers which can be looked as the internal markers in watershed segmentation. With the new markers, more potential protein spots can be detected by our method.

Furthermore, a new modeling method is proposed for modeling of the saturated region of the protein spots. It utilizes the axis-parallel ellipses to make covering on the saturated region. With clustering method, we will obtain reasonable estimation of the positions of the potential merging spots which construct a large saturated spot. Then particle swarm optimization will be used to minimize the covering error and find the best covering ellipses for those protein spots. The combination of all the detected ellipses builds up the model of the saturated spot region. Simulations show satisfying results of our modeling method.

The importance of our clustering based spots detection method and spots modeling methods in comparison with existing works is that our methods make the detection and modeling methods for protein spots image analysis more accurate and faster, so that can help researchers who do the job of protein spots detection and modeling to save their time and should be helpful to improve the performance of the further processes such as spots matching and comparison.

1.4 Organization of the Thesis

This thesis is organized as follows:

Chapter 2 gives a comprehensive literature. This includes the introduction of proteomics and the protein profiling methods, the details of biological and image processing of 2D gel electrophoresis, and spots detection methods that are currently used.

Chapter3 introduces the clustering methods and integrate the subtractive clustering to the gray scale images. Experiment results of some images are given to demonstrate the good performance of this method.

Chapter 4 presents a new modeling algorithm for saturated protein spots. The new algorithm integrates the particle swarm optimization algorithm with subtractive clustering to model the saturated protein spots. Simulation result has been given.

In Chapter 5, conclusions are presented and recommendations for further research are given.

2. Literature Review

2.1 Proteomics

2.1.1 Introduction of Proteomics

As genomics is about the study of genes in genomes, proteomics deals with the analysis and research of the proteome, which is the product of translation of the transcriptome [1].

Proteomics emerged as a discipline in the mid-1990s and has grown rapidly since its inception. This new approach to biology and biomedical research was enabled by the emerging availability of complete sequences of genomes and the development of high sensitivity mass spectrometry techniques and instruments for analyzing proteins and peptides [3].

Proteomics yields a new conception of the functional assignment issue in biology. ‘Prote-’ indicates that function is sustained by proteins, not by genes, and ‘-omics’ proposes that function defined in context [1].The function of a protein is not only an individual property but is defined as a combination of its biochemical interactions with its partners and the environment in which it exists. Information on the scale of the whole cell is needed to understand the function of protein comprehensively.

The whole of the genome sequence information is now available for many organisms of human. Sequence of this kind of information shows us many novel genes that no functions can be assign to. Even for many well studied model systems, the specific functions of many genes are unknown [2]. Understanding the function of each gene in the genome has led to development of large scale, high throughput experimental techniques that are collectively referred to as functional genomics. These studies include systematic disruption of predicted genes, mRNA expression profiling based on microarray or DNA chip technologies, protein expression profiling and mass spectrometry, and large scale mapping of protein-protein interaction.

In addition, the genome of a cell or organism is finite and static over its lifetime, but in contrast, proteomes are highly dynamic and constantly changing in response to external stimuli and during the development. Especially for some complex organisms, the potential number of unique proteomes is nearly infinite.

In contrast to conventional studies that focus on detailed characterization of one or at most a few genes or proteins, proteomics takes a broader, more comprehensive and systematic approach to understand biology. Also, proteomics is primarily discovery-based rather than hypothesis-driven [3]. Therefore, proteomics is not constrained by prior knowledge, and it can provide exciting new opportunities to advance knowledge by identifying new drug targets, correlate biological pathways and molecular mechanisms with disease and enable a systems view of biology.

Although the genome of human is now estimated to contain only nearly 35000 genes, the total number of unique protein components encoded by genome is of several millions. That is because of the alternative splicing of many mRNA and extensively variable posttranslational modifications of many proteins [3]. Therefore, as the next logical step after genome sequencing, research of proteomics is much more complicated and challenging than that of genomics.

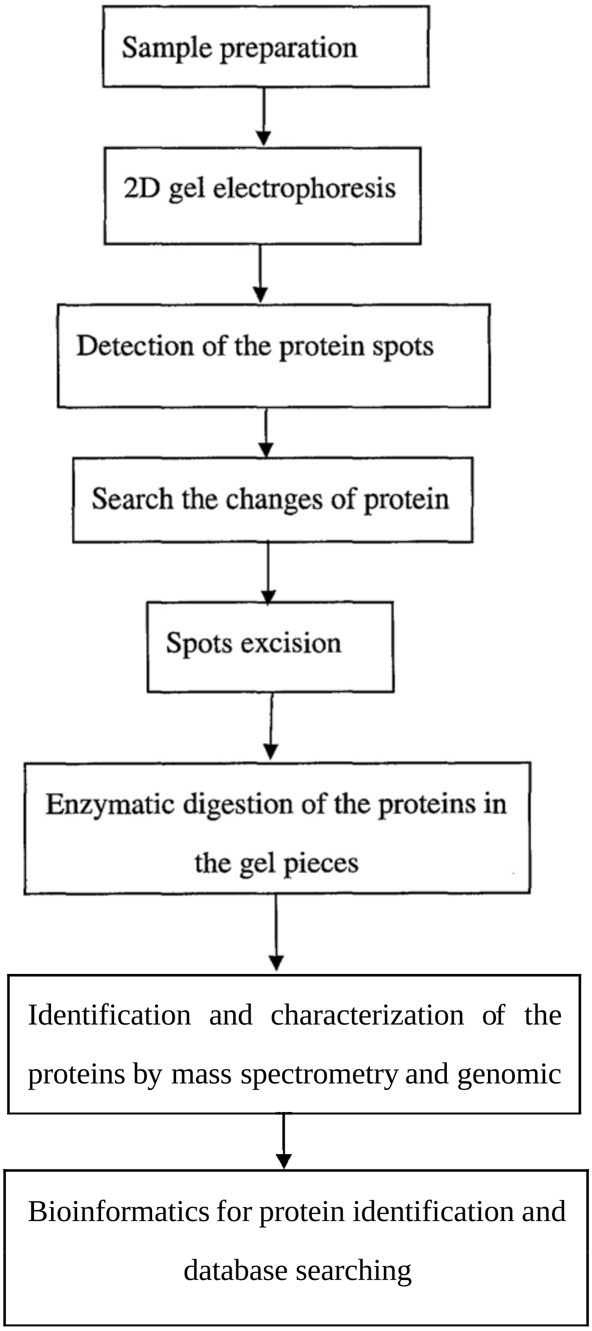
The types of proteomics studies include: (1) quantitative comparisons of protein levels in multiple samples (protein profiling); (2) analysis of protein-protein interactions including binary interactions and isolation/analysis of macromolecular complexes; and (3) protein compositional studies.

2.1.2 Protein Profiling Methods

Protein profiling, the most common type of proteome study, is the quantitative comparisons of protein in two or more samples. Analogous to most nucleotide array experiments, the purpose is to define the changes in levels of all proteins that define different physiological, experimental or disease states [3].

Basically, protein profiling methods can be classified as top down and bottom up approaches. With bottom up methods, two samples to be compared are differentially labelled with an appropriate stable isotope label and the mixture of the two unfractionated protein extracts is digested with trypsin prior to any separation. Protein identities are determined from the tandem mass spectrometry and relative abundances of proteins are determined by ratios of signals for the light and heavy forms of identified labelled peptides. Top down methods can separate intact proteins by separation modes. At the same time, the quantitative changes in levels of individual proteins are measured at the intact protein level. The proteins which are exhibiting the largest changes in abundance are identified by fragmenting the protein with trypsin followed by a mass spectrometry analysis.

With the top down method, the proteome analysis includes the following steps shown as flow chart [5]:


Sample preparation
2D gel electrophoresis
Detection of the protein spots
Search the changes of protein
Spots excision
Enzymatic digestion of the proteins in the gel pieces
Identification and characterization of the proteins by mass spectrometry and genomic
Bioinformatics for protein identification and database searching

2.2 2D Gel Electrophoresis

The important requirement for proteomics is the separation, visualization and then analysis of the complex mixtures that contain several thousands of proteins coming from the whole cells, tissues or organisms. Two-dimensional gel electrophoresis is the most frequently used and also the oldest top down protein profiling technology which remains the core technology of choice for separating complex protein mixtures in the majority of proteome projects. The basic method of 2D gel electrophoresis has been used for nearly 30 years and it has a number of important features [5]:

1. High resolving power
2. Tolerance to the crude protein mixtures to avoid the further modification
3. Tolerance to relatively high sample loads to detect the minor components
4. 2D gels are very efficient fraction collectors for several thousands of spots
5. Proteins are protected inside the gel matrix for analysis

2.2.1 Biological Processing

Two-dimensional gel electrophoresis enables separation of mixtures of proteins based on the differences of their isoelectric points in the first dimension, and subsequently by their molecular weight in the second dimension. The main biological processes of 2D gel electrophoresis are:

1. Sample Preparation

The first step of 2D protein gel electrophoresis is sample preparation. Adequate sample treatment is an important requirement for a successful experiment. The sample preparation procedure must stably solubilize all proteins, prevent protein aggregation and hydrophobic interaction, remove or thoroughly digest RNA or DNA and so on [6]. Every naturally polypeptide should be represented by only one spot in the gel. There is no universal protocol for the sample preparation. Different extraction and lysis techniques are used for different sample sources.

2. Loading of Proteins

Then the protein sample should be loaded in the gel. The amount of proteins from samples applied to a gel can vary from several micrograms to one gram. The protein capacity is dependent on the volume of the gel. Whether the sample is to be loaded on the anode or cathode end of the isoelectric focusing gel should be determined based on the experiment respectively for each new sample.

3. Separation in two dimensions

Two-dimensional gel separation first separates proteins based on their isoelectric point (pI) with isoelectric focusing (IEF).The isoelectric point is the pH value at which there is no net electric charge on a protein. IEF is an electrophoretic technique by which the proteins are separated in a pH gradient.

An electric field is applied to across the gel and charged proteins start to migrate into the gel. The proteins are differently charged and the electric field will pull them to the point where the pH cast into the IPG is the same as the pI of the protein. For instance, the pH value at which the numbers of positive and negative charges on the protein are the same, at this point no net electrical force pulls the protein. Eventually, the gradient and the proteins move to the positions in the pH gradient equivalent to the pI. Since the pI of a protein is based on its sequence of amino acid, this technique owns good power of resolving. The resolution of this technique can be changed further by adjusting the range of the pH gradient.

After the first dimensional focusing, the next step is to separate proteins based on molecular weight. It is like a molecular sieve so that the small molecules can pass more quickly than the large ones. An electrical field is applied in the perpendicular direction and the proteins migrate into the gel. Because all proteins have the same charge per unit length, the same electrical force will poll them. But the light proteins will meet less obstruction in the gel and will migrate with higher velocity through the gel. The large proteins meet more resistance and will migrate slower. Proteins which have the same p Imigrate in the column will be separated by the molecular weight.

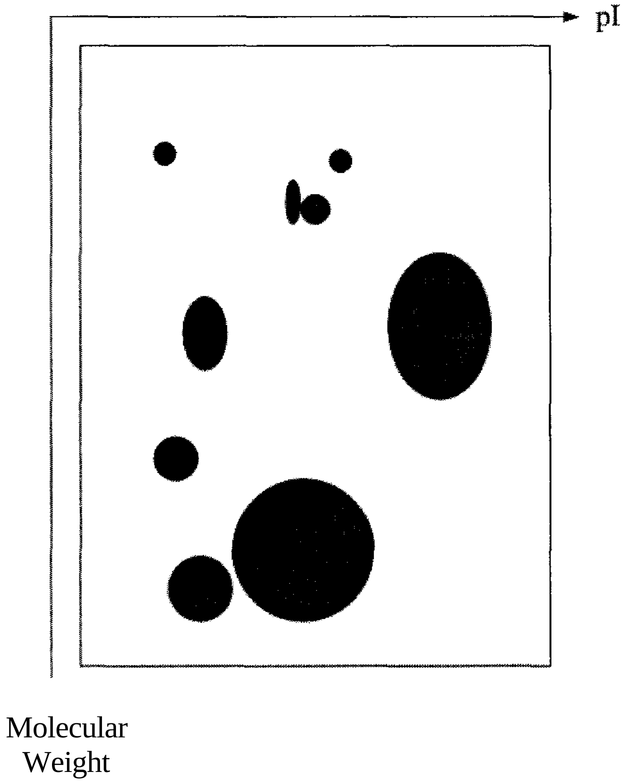


Figure 2-1 Proteins are separated in two dimensions

4. Protein detection techniques

The ideal protein detection techniques should have a wide linear dynamic range, be very sensitive, quantitative, compatible with the further analysis with mass spectrometry [5]. The detection techniques involve labeling and staining. Coomassie blue and silver are the most frequently used detection methods. Coomassie brilliant blue can stain nearly all proteins with good quantitative linearity, but it is not sensitive enough. Silver staining is the most sensitive technique for isoelectric focusing gel. Most of our experiment images are based on the silver staining technique.

5. Image Acquisition

Image acquisition will be preferably done with scanners rather than with CCD cameras due to the requirement of high resolution and a homogeneous background from the center to the edge. Therefore, image acquisition significantly predetermines the final quality of the analysis.

The first two steps are used to get adequate biological sample and make it ready to start the separation process. With the third step, different kinds of proteins are physically separated to proteins spots in the different areas in the gel. The last two steps stain the protein spots with colorful substance firstly and then scan the gel to get an image that can totally express the presence of those spots in the gel. All the further analyses are based on this image.

2.2.2 Image Processing

Without the help of automatic image analysis tools, the tasks of image analysis that the scientists are required to perform are tremendous. First, they have to identify the spots on the image that are considered to be protein spots. Then they may landmark the spots that are identify by the previous step. After the landmark processing on the spots, comparison of the image is then performed with a match set (usually a synthetic gel image). The difference of protein spots in different gel images may be very subtle. If all of the procedures mentioned above will be performed manually, detecting thousands of protein spots in a gel image through human visual inspection is almost impossible to complete within acceptable time.

Depending on automatic image analysis, the processing speed and objectivity can be greatly improved. But we found that most existing commercial software for two-dimensional gel electrophoresis image analysis cannot correctly detect and segment the saturated protein spots because most of them are based on the general image analysis methods which can not suit the 2D gel protein image well (In section 3.1, we will show the detail of disadvantages of watershed method, which is the most popular method to detect protein spots). Therefore, there is a great need for the development of better image analysis methods for the processing of 2D gel electrophoresis image. Basically, image enhancement, spots detection and spots fitting are the main three image analysis processes for 2D gel electrophoresis image.

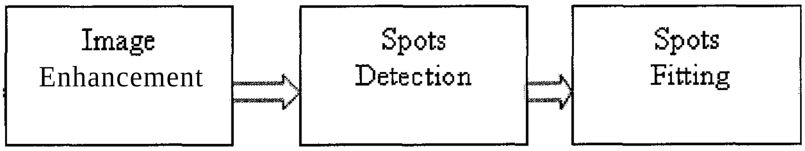


Figure 2-2 Steps of two-dimensional gel electrophoresis analysis

Figure above shows the sequence of image analysis of 2DGE. Image enhancement is used for smoothing, contrast enhancement, background subtraction and so on. It will help improve the optical quality of images or provide the exact data. Spots detection is the most important part of image analysis for 2D gel electrophoresis. It will help us to find the exact positions and shapes of the protein spots. Sometimes, the modeling methods of the proteins spots can also be used for detection. After the spots detection step, **spots** should be characterized and represented as a list of parameters for further analysis. These parameters may come from spots fitting based on the modeling of the spots figures.

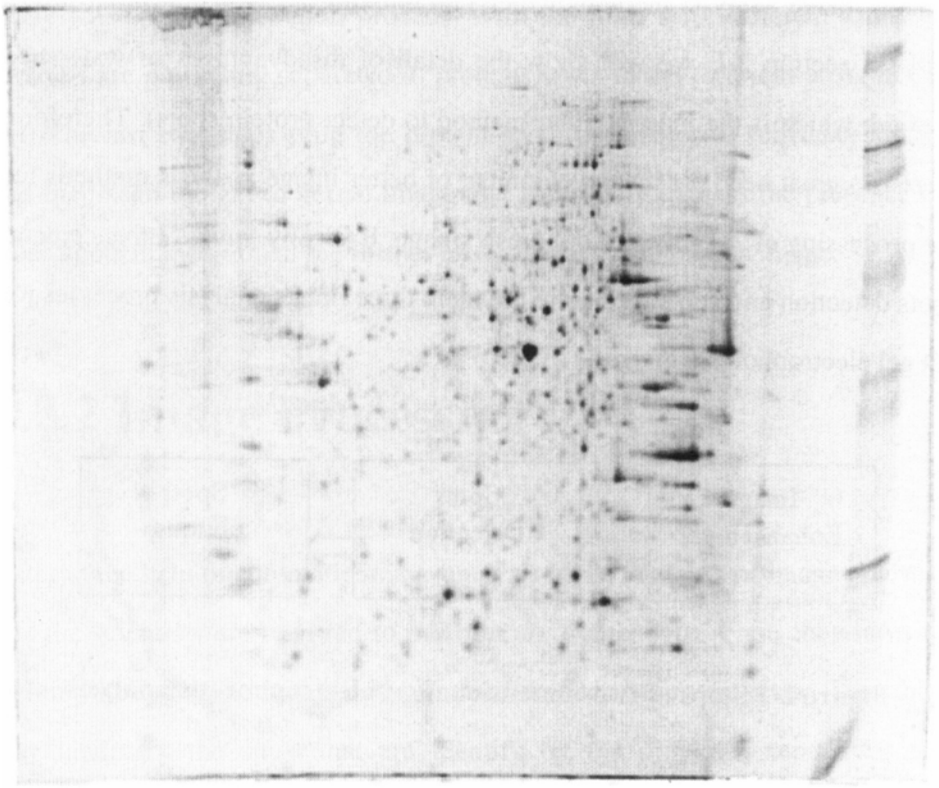


Figure 2-3 Original 2D gel electrophoresis image

2.3 Spots Detection Methods

In 2D gel electrophoresis image, protein spots vary greatly in intensity and size and are often difficult to distinguish from the noise and streaks. Furthermore, the heavily saturated spots are overlapped and have no observable boundaries between each other. In such situation, spot detection is very difficult. Currently, the most popular pipeline of spots detection is: (a) Detect the centres of spots as many as possible; (b) Segment the gel into different regions each of which containing one spot; (c) Model every region by a parametric spot model *[7]*.

Since 1980s, many spots detection methods had been proposed. As the techniques become more mature, the precisions of detection are much higher than before. Next, a brief review of spots detection methods is presented.

2.3.1 Laplasian of Gaussian Techniques

In Melanie Ⅱ [8], the protein spots are detected with a nonparametric method which is based on the Laplasian and second derivatives. Assuming there is a 2D gel electrophoresis image I(x, y) , a saturation threshold value T, a protein spot S_i and a point $\vec{p} = (x,y)$. To examine whether point $\vec{p}$ is a part of a protein spot, two decision rules are defined, based on the intensity $I(\vec{p})$ of the point $\vec{\boldsymbol{p}}$:

Rule 1: $Z(\vec{\boldsymbol{p}}) \le \boldsymbol{T}$ means it is an unsaturated value;

when $-\Delta I(\vec{p}) - L \ge 0$,

$$\vec{p} \in S_i \Leftrightarrow MIN(\frac{\partial^2}{\partial x^2} I(\vec{p}) - R, \frac{\partial^2}{\partial y^2} I(\vec{p}) - C) > 0 \qquad (2.1)$$

Rule2: $I(\vec{p}) > \boldsymbol{T}$ means it is a saturated value;

$$\vec{p} \in S_i \Leftrightarrow MIN(\frac{\partial^2}{\partial x^2} I(\vec{p}), \frac{\partial^2}{\partial y^2} I(\vec{p})) > 0 \qquad (2.2)$$

where $L, \boldsymbol{R}, C$ are three very small positive constants. Here, L represents the threshold for the Laplasian result of image I, R is the threshold value for the second derivative along the axis X and C is the threshold value for the second derivative along the axis y . Because of the staining technique and the image acquisition process, the 2D gel electrophoresis image may be saturated.

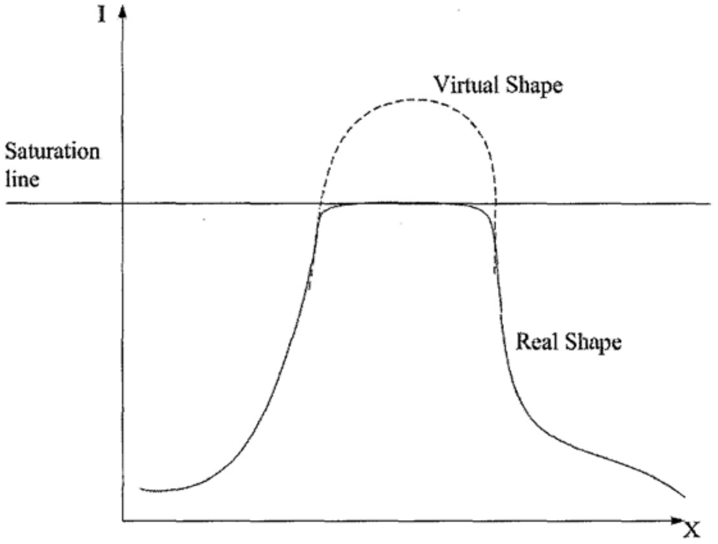


Figure 2-4 Profile of a saturated spot

Figure **2-4** illustrates a saturated spot, in which virtual shape is an ideal shape truncated due to the saturation effect, so the pixels above the threshold value T are considered to be saturated. The Laplasian ΔI and threshold T are defined as follows:

$$\Delta I(\vec{p}) = -(\frac{\partial^2}{\partial x^2} I(\vec{p}) + \frac{\partial^2}{\partial y^2} I(\vec{p})) \tag{2.3}$$

$$T = \max(I) - \frac{100 - saturation}{100} \times (\max(I) - \min(I)) \tag{2.4}$$

The saturation is a positive value between 0 and 100. If the saturation is equal to 100, it means that no pixels are saturated in the whole 2D gel electrophoresis image.

This method is easy to implement and has been used in the software Melanie II with combination of spot quantification.

2.3.2 Fast Spots Segmentation Algorithm

Wu ***et** al.* [9] proposed a fast spots segmentation algorithm for two dimensional gel electrophoresis analyses in 1993. It uses the second derivative of the original to generate the central image. In the spots extraction procedure, it encodes the entire image in single raster scanning pass to extract spots on the fly to avoid the much time consuming. This algorithm is described as follows:

Create the central core image and the second derivative magnitude image. The central core image is created also based on the second derivative result of the original image. In a location, if the second derivative results in both column and row directions are negative, the pixel in the central image is assigned to 1, or it is assigned to 0. The second derivative image is constructed at the same time by adding the absolute value of the second derivatives in both row and column directions. They use 16 bit pixels for the image buffer because the central core image will later be used to encode the spots regions. Thus it can hold up to a maximum of 65536 spots in 2D image.

Spots extraction is the core procedure that encodes the central core image with spots numbers in only one raster scanning pass through the image. After that, the central core image has been segmented with spot regions marked by their unique identification numbers. Therefore, in the encoded image, the pixels of a same spot have the same code value. In this algorithm, each image scanning line is treated in a similar way. To encode one row, they check every pixel with the following rules:

(a) If the value of pixel is 0, a background pixel but not a spot central core pixel, continue to the next pixel.

(b) If the value of pixel is 1 and its top neighbor and left neighbor are 0, encode the pixel with a new spot identification number. Then move to right for the next pixel (Figure 2-5).

(c) If the value of the pixel is 1 and its top neighbor has an identification number of spot and its left neighbor is 0, encode the pixel with the spot identification number as same as its top neighbor pixel. Then move to right for the next pixel (Figure 2-6).

(d) If the value of the pixel is 1 and its left neighbor has an identification number of spot and its top neighbor is 0, encode the pixel with the spot identification number **as** same as its left neighbor pixel. Then move to right for the next pixel (Figure 2-7).

(e) If the value of the pixel is 1 and both the top neighbor and left neighbor respectively have spot identification numbers, encode the pixel with the spot identification number as same as its top neighbor pixel and then set the spots identification number of the top neighbor and left neighbor to be equivalent for later use (Figure ***2-8).***

(f) Repeat from **(a)** to (e) for the pixels in the image row.

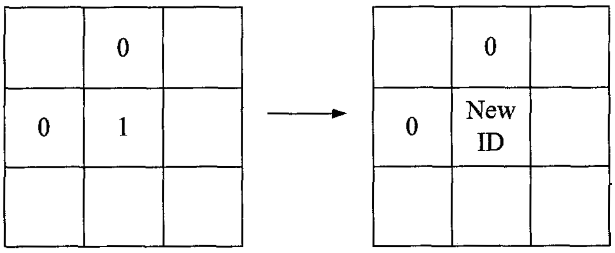


Figure 2-5 Transformation rules

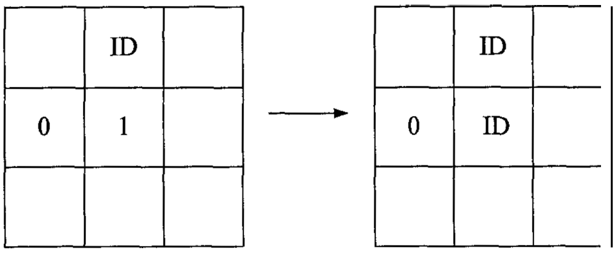


Figure 2-6 Transformation rules

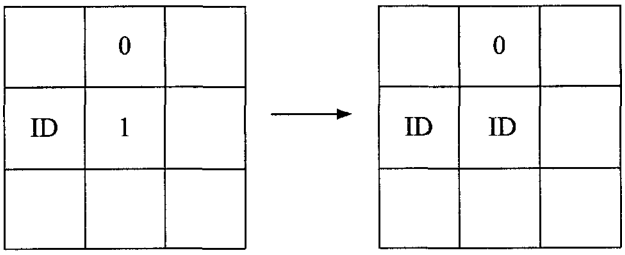


Figure 2-7 Transformation rules

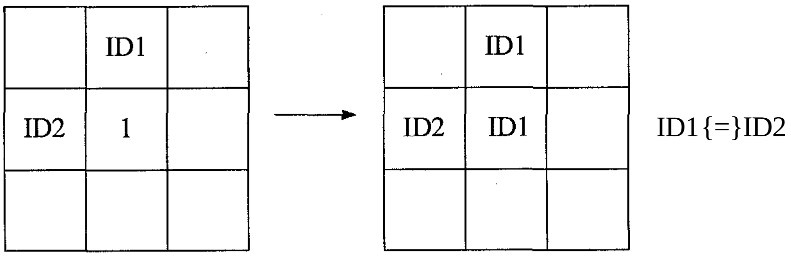


Figure 2-8 Transformation rules

Procedures above should be performed to all lines in the 2D gel electrophoresis image to encode the central core image with the spot identification numbers. Pixels in the same spots regions should be marked with the same identification numbers.

The next stage is spots propagation. In the segmented central core image, we find the maximum extent of all spots. The spots boundaries are finally generated at this stage.

(a) Find the minimum enclosing rectangle of the spot and get all the pixels on the boundary of this spot.

(b) For every pixel on the boundary, check all the eight connected neighbor pixels around it. If anyone of them meets some criteria, it is marked with the same identification number. It means that this pixel belongs to the spot. The criteria are: <1> The pixel is 0, a background pixel. <2> The image gray value decreases when moving from the spot boundary pixel to the background pixel. <3> The second derivative value increases when from the spot boundary pixel to the background pixel. The propagated pixel should be marked as a boundary pixel and be checked.

(c) Repeat (b) until all the boundary pixels for this spot are checked and record these pixels.

(d) Repeat from (a) to (c) for all the spots.

(e) Calculate all the spots features including spot area, optical density, spot centroid, spot boundary, minimum and maximum density, minimum enclosing rectangle and other features.

This algorithm is memory efficient to be able to process very large 2D gel electrophoresis images in a reasonable amount of computation time on a low cost computer, such as a personal computer or workstation.

2.3.3 Line and Chain Analysis Algorithm

A scanning system is used for the quantitative analysis of complex protein mixture [10].With the detection of the spots, the densities are integrated after subtraction of local background. Then the algorithm screens out the streaks or other noise in the image and uses the curve-fitting techniques to resolve the partially overlapping spots. After film scanning, the image will be analyzed by three following processes:

1. Line analysis: Reduce the data in the scanning dimension

This analysis is implemented after the data has been got from scanner. Once the scanner traverses a protein spot, a density peak is received by the computer. Then it will integrate the peak related to the most recent reading background. If no peak integrated, background reading should be taken. It always makes sure that the background is always updated to a value corresponding to each location in the image. For each completed data line corresponding to one scan, the computer holds the integrated density and center position of each detected peak in memory.

In the film image, the scanning dimension of molecular weight is the horizontal axis. Across the film, every scan line has 381 reading of density, or one reading every 140 um. The curve is smooth, so the falling and rising slopes of the spot peaks are monotonic. Because the circuitry in the photometer average the signal received over time, the data is smooth. The derivative of the curves is also smooth so that very small unusual change of shoulder will indicate the overlap of protein spots.

Some parameters are contained which can be used to control the slopes for trigger integration, the change of the slope for detection of the overlapping spots, and the slopes for termination of the integration. These parameters are turned for optimum recognition of shoulders and the faint spots with minimal triggering due to the film noise. Because the parameters only work at the point where the integration begins and ends, they have no obvious effect to the integrated density for well resolved spots that rise above the triggering levels.

A peak rising from background is detected by two successive density increases. Peak detection requires typically a rise between two readings of 0.003 OD or more followed by 0.005 or more. If the second rise does not follow, the integration begins with the first rise will be aborted. The integration will also abort with density below a threshold value that typically 0.02 OD when the density stops to rise. The abort steps allow real peaks to be triggered at a value very near to background.

The inflection point is detected as the maximum rise rate on the rising side of every peak. From the negative inflection point slope, a falling side criterion slope is obtained. The criterion slope is equal to a fraction of the inflection point slope minus a constant slope. If the slope becomes less negative than the criterion slope, a subsequent decrease in slope will help to detect a shoulder. It is the falling side shoulder recognition point. With the detection of the shoulder, if the density of the peak exceeds the threshold for shoulder splitting, the curve will be treated as overlapping peaks. The integration of the second peak will be started immediately after the termination of the integration of the first peak. The shoulder splitting is not perfect due to its involving of a straight cut between two overlapping peaks. A peak will terminate with the return to near background, the stop of falling or a new rise.

2. Chain assembly: Match the peaks from successive scan lines

After the termination of the line analysis for all scan lines, the peak integrated densities and center position should be assembled into a form for analysis in the second dimension. For the chain analysis, in each line, the peaks are matched to peaks of the previous line based on the proximity of their centers. If distance of the two center positions is less than 0.8 mm, the two peaks of the adjacent scan lines will be matched. Every spot of the film will make a chain of peaks in adjacent scan lines.

For chain assembly, the data of peak integrated density and peak center position are integrated as a function of line number in the stepping dimension. Many chains have substantial displacements of peak centers in the scanning dimension. If a chain includes more than one spot, it will be accurately resolved during the chain analysis. When the spot is not resolved, gaps will be left in the chain. The program can recognize these gaps and fills them before the chain is stored. The value inserted into the gap is equal to the integrated density of the previous one in the chain.

3. Chain analysis: Reduce the data in the stepping dimension

With chain analysis each chain of peaks is analyzed to resolve overlapping spots, to find noise and streaks and to compute the full integrated density of each spot finally. After the completion of the chain analysis, the film image is reduced to a pattern of spots, that each of them is characterized by its two dimensional integrated density and two positional coordinates.

As in the case of line analysis, the chain analysis program also contains some adjustable parameters. For a chain which represents a simple and isolated spot, the chain of peak integrated density is a smooth Gaussian curve and the chain of peak center positions is a straight line. The chain of center positions and the chain of integrated densities are both used to determine the resolving of the overlapping spots. A change in the chain of center positions is a sensitive measure of spots overlapping.

The procedures of chain analysis are given as follows:

(a) Smooth the chain

The computer loads the chain of integrated densities and the chain of peak position into memory. After that, each of the chains will be smoothed to remove the systematic variations that may occur between rightward and leftward scans across a spot. This kind of variations in the integrated density is usually caused by an inhomogeneous background across the protein spots. Though the systematic variations of the peak center positions are small, they may occur because of the little differences in measuring position on rightward and leftward scans. This smoothing is performed by replacing each entry in the chain by the sum of that entry that immediately precedes it.

(b)Find the Gaussian mean position

Each chain of peak integrated densities can be resolved into one or more Gaussian curves. For curve fitting, it is necessary to find the accurate mean position about the fitted Gaussian. For the best mean, firstly they find the maximum integrated density in the chain and consider 5 points centered about the maximum. After subtract a large fraction of the maximum integrated density from each point, they leave just a small polygon representing the tip of the curve. The Gaussian mean position is taken to be weight-averaged center of the small polygon.

(c) Select points for curve fitting

A Gaussian curve will be fitted in both sides of the selected mean value. The fitted points are selected alternately, beginning from left and then from right of the mean value. The selection of point in one direction will stop at: 1) two successive points of the integrated densities have a value less than 40% of the maximum. 2) two successive points in the peak center positions are shifted by more than 0.5 mm from the average center of the selected points for fitting. 3) encounter a new rise of integrated density in the chain.

With the limitation of the right and left for Gaussian curve-fitting having been determined on the chain of integrated densities, the points on each side of the center are examined to find if one side falls more steeply than the other. The curve will be fit to points on the steep side.

(d) Fit the curve

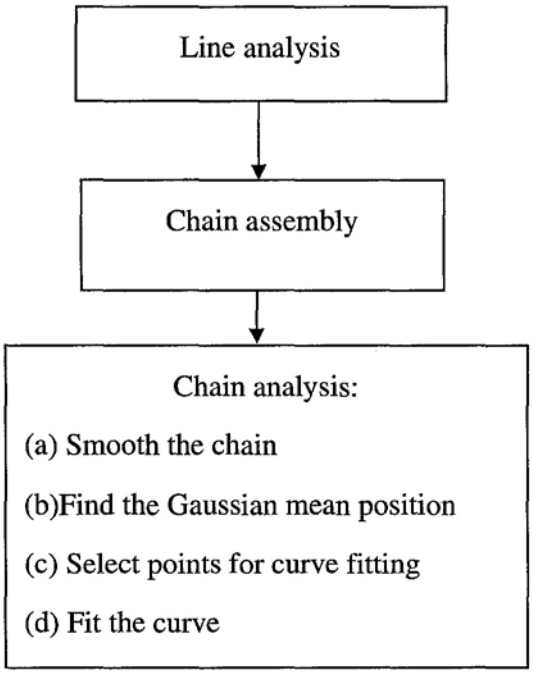
They use a linear regression to fit a Gaussian curve to a set of points. When the best straight line is fit through the transformed data, they use the slope and intercept to calculate the Gaussian width at half the maximum and height.

With the computed best Gaussian, they record the parameters and subtract it from the chain of integrated densities, then find the maximum Gaussian mean and analyze the remaining integrated densities in the chain. If no points in the chain of integrated densities are above a typically threshold, the curve fitting is completed.

After screens for noise and streaks and handing of the tails, the final result of the line and chain analysis will be obtained.

This algorithm allows the faintest spots on the gel to be quantified with greater accuracy. The Gaussian spot model allows curve fitting algorithms to be applied to resolve the clusters of some overlapping spots. But if three are many saturated spots in the gel image, this algorithm cannot get satisfying accuracy, because the Gaussian spot model is not suitable to fit to such spots.

The following block diagram will present the steps of line and chain analysis clearly:



2.3.4 Fully-Automated Algorithm Based on RLGS

This fully-automated spot recognition algorithm for 2D gel electrophoresis is based on RLGS (restriction landmark genomic scanning) [11].Without any human interaction, thousands of protein spots on a 2D gel electrophoresis image, including some hidden spots at the shoulder of the large spots, can be identified by this algorithm.

The fully-automated recognition algorithm of spot locations and intensities is introduced as follows:

1. Preprocessing for the image. Perform enhancement and smoothing. The preprocessed image is $\phi(x,y)$.

2. Apply the background normalization operation to $\phi(x,y)$. The resultant image is $\varphi(x,y)$.

3. With the conventional smoothed density histogram method, we can find the over all background level t of $\psi(x,y)$.

$$f(x,y)=\begin{cases}1, \psi(x,y)\geq t\\ 0, otherwise\end{cases} \quad (2.5)$$

The binary image $f(x,y)$ of $\psi(x,y)$ gives the possible spots domain.

4. In the area of $\{(X,y) \mid f(x,y)=1)$to$\psi(x,y)$, a ring operator is applied to detect the local maxima independently of background density.

$$C(x,y)=\{(u,v)\mid(u-x)^2+(v-y)^2/\alpha^2\leq r_M^2\} \quad (2.6)$$

$$R(x,y)=\{(u,v)\mid r_m^2\leq(u-x)^2+(v-y)^2/\alpha^2\leq r_M^2\} \quad (2.7)$$

a is the ratio of the short axis to long axis for an ellipse. The output of a ring operator is:

$$h(x,y) = \max_{C} \phi(x,y) - \max_{R} \phi(x,y) \qquad (2.8)$$

The ring operator was adopted to allow well detection of flat spots that have long tail in the first dimensional electrophoresis axis X . The recognized spots are labelled for further identification. The location of the recognized spot i is (s_i^x, s_i^y) .

5. To identify the spot domain, gray levels of $\phi(x,y)$ are analyzed and each pixel will be classified into one spot domain by:

(a) Label the pixels that have the highest gray level in the image $\phi(x,y)$. Therefore, the labeled pixels are local maxima of normal spots or in the regions of the flat spots. Then the labeled pixel is classified into the respective spot domain D_i . The pixels on the domain of $\{(x,y) | \psi(x,y) \leq 1)$are ignored.

(b) Another label marks those pixels whose gray levels are the highest among all the unlabeled pixels, These second labeled pixels are classified into the respective spot domain that appears in the previous steps. This procedure will stop when no more second labeled pixels adjoin any domain. After that, the residual pixels are used to define new spot domains.

(c) Repeat above steps until no pixel can be labeled in the image.

6. The spot domain corresponding to spot i is D_i . Gaussian function centred

at (s_i^x, s_i^y),

$$g_i(x,y) = a_i \exp\{-(x-s_i^x)^2/2S_i^x - (y-s_i^y)^2/2S_i^y\} \qquad (2.9)$$

is fitted to the normalized $\psi(x,y)$. These parameters an adjusted to minimize the function:

$$E_i = \sum_{(x,y)\in D_i} [\log\{\psi(x,y)\} - \log\{g_i(x,y)\}]^2 \qquad (2.10)$$

7. Subtract the fitted Gaussian functions from the normalized image $\psi(x,y)$. The residual is:

$$r(x,y) = \psi(x,y) - \sum_i g_i(x,y) \qquad (2.11)$$

8. Apply the ring operator to the residual image $r(x,y)$ to detect additional hidden spots.

9. Use Gaussian function to fit to the residual image $r(x,\mathrm{y})$ and find the fitting parameters.

10. Refine all the fitting parameters to minimize the error function,

$$E_i = \sum_{(x,y)\in D_i} \{\psi(x,y) - \sum_i g_i(x,y)\}^2 \qquad (2.12)$$

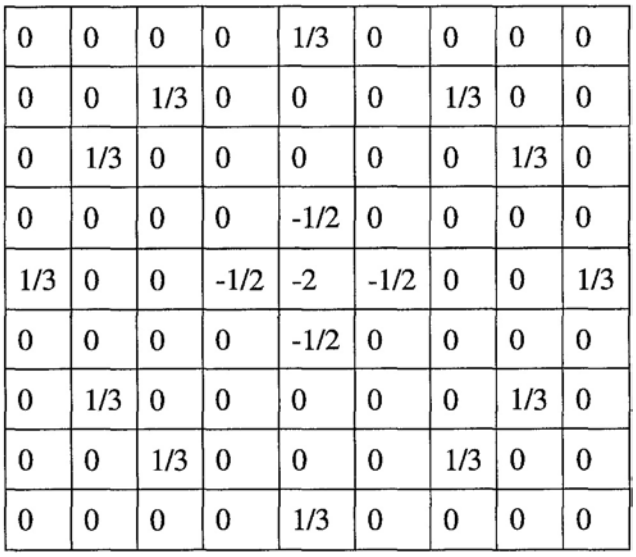
After this process, calculate the integrated density of spot i,

$$I_i = \sum_{(x,y)} a_i \exp\{-(x-s_i^x)^2/2S_i^x - (y-s_i^y)^2/2S_i^y\} \qquad (2.13)$$

It uses Gaussian modeling of the landmark spots. If the possible location of the spot is detected by the elliptic ring operator, each spot on a RLGS profile will be fitted with Gaussian-type function. Comparing with the previous works, this algorithm can remove the ill-recognized spots and detect the hidden spots automatically.

2.3.5 Detect with Coefficient Template

Prehm ***et** al.* [12] assumed the spots belonged to areas displaying **a** positive curvature of the gray level surface. The first approximations of position, size and shape of protein spots are the spot kernels. They are determined by this way: the gray values of the pixels in the neighbourhood of ***n x n*** pixels are multiplied by the coefficients of a segmentation template:



0	0	0	0	1/3	0	0	0	0
0	0	1/3	0	0	0	1/3	0	0
0	1/3	0	0	0	0	0	1/3	0
0	0	0	0	-1/2	0	0	0	0
1/3	0	0	-1/2	-2	-1/2	0	0	1/3
0	0	0	0	-1/2	0	0	0	0
0	1/3	0	0	0	0	0	1/3	0
0	0	1/3	0	0	0	1/3	0	0
0	0	0	0	1/3	0	0	0	0

Figure 2-9 9×9 coefficient template

The resulting products are summarized to get the degree of curvature in the central pixel of the neighbourhood. If the sum is bigger than zero, the central pixel belongs to a spot kernel, or it belongs to the background. But sometimes many spots are located very close to each other and show asymmetric shapes. These spots will often not be detected as two spot kernels because the area of the gray value surfaced between the spots does not show a negative curvature. However, the degree of curvature in these areas is lower than the degree of curvature in the area of the spots, so for separating the close spots, these may be found from the spot kernels and attributed to the background. For this purpose, they use a surrounding area with other ***m x m*** pixels and calculate the differences in the degree of curvature between the central pixel and the pixels of the rim of the neighbourhood.

This method is easy to implement and try to detect all the visible spots, as many as possible of the invisible but true spots, and to avoid all the invisible artifactual spots.

2.3.6 H-basin Transformation

Horgen and Glasbey [13] applied the h-basin transformation to 2D gel electrophoresis image analysis.

H-basin method is reversed from h-dorm which is a morphology based method and a technique to determine the maximal structures in gray scale image. The regional maxima not exceeding h in size are found by first subtraction h from all pixels in the image. Iterative geodesic dilations are then performed on the resultant image until stability. When this image is subtracted from the original images only the maxima domes remain. The 2D gel electrophoresis images are traditionally viewed as dark protein spots on a bright background. In such image, the extraction of minimal structures (the spots) can be done by the h-basin (reverse of h-dome) method.

This algorithm is based on the mathematical morphology and is not sensitive to the noise. But for all pixels in the image, it is necessary to select a threshold h, which cannot be adaptive in the whole image, so it will inevitably encounter problems in some regions of the image when performs segmentation.

2.3.7 Watershed Algorithm

Watershed is a terminology from geoscience [14][15]. It is the most popular technique for gel segmentation in nowadays due to its robustness to noise. For 2D gel electrophoresis image, it can segment the spots regions by using some neighbourhood properties.

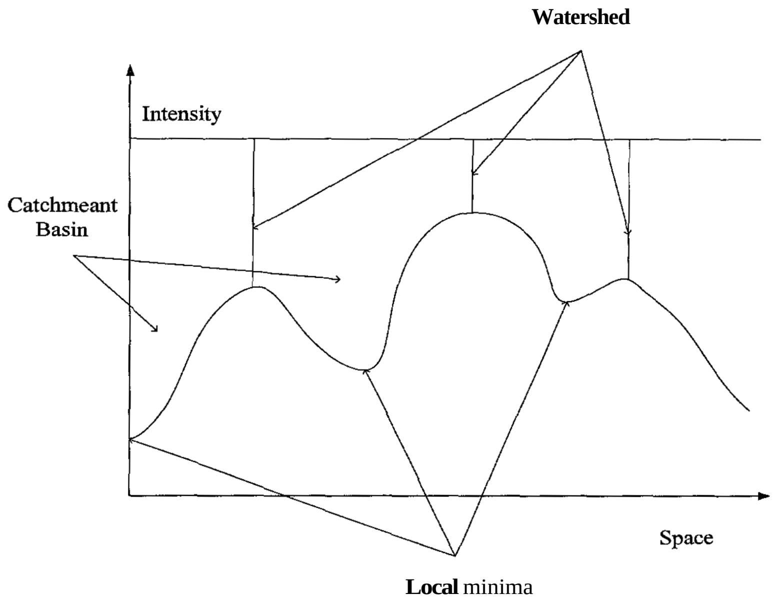


Figure 2-10 Principle of the watershed transformation as an immersion process

On the topographic representation, a watershed is the boundary of a region in a landscape and each region is a catchment basin. The lowest points in the catchment basins are called local minima. Because the 2D gel electrophoresis image is a gray scale image, it can be regarded as a landscape. If the image is separated to meaningful regions by watershed segmentation, each region containing a real spot is regarded as a catchment basin. The points with the lowest gray value of a spot are corresponded to the local minima in the catchment basin.

The input image is transformed to output an image whose minima mark relevant image objects (catchment basins) and whose crest lines correspond to image object boundaries (watersheds). This is illustrated in Figure 2-10.

To get the watershed line, we can regard the gray scale image as a topographic landscape at first and punch a hole in each local minima of the catchment basin. Then the landscape is slowly submerged in water. Starting from the minima at the lowest altitude, the water will progressively flood in the catchment basins of the image. In addition, dams are built at the place where the water comes from two different minima will merge. At the end of the flooding procedure, every catchment basin is surrounded by dams. The whole set of dams correspond to the watersheds.

In the result of basic watershed segmentation, the gel electrophoresis image is successfully separated into meaningful regions. Based on the basic methods of watershed segmentation, there is only one set of local minima in one region because the local minima depend on the lowest gray value inside the regions. If conventional detection methods are employed, in one region we will get one boundary of the spot inside, which contains the local minima.

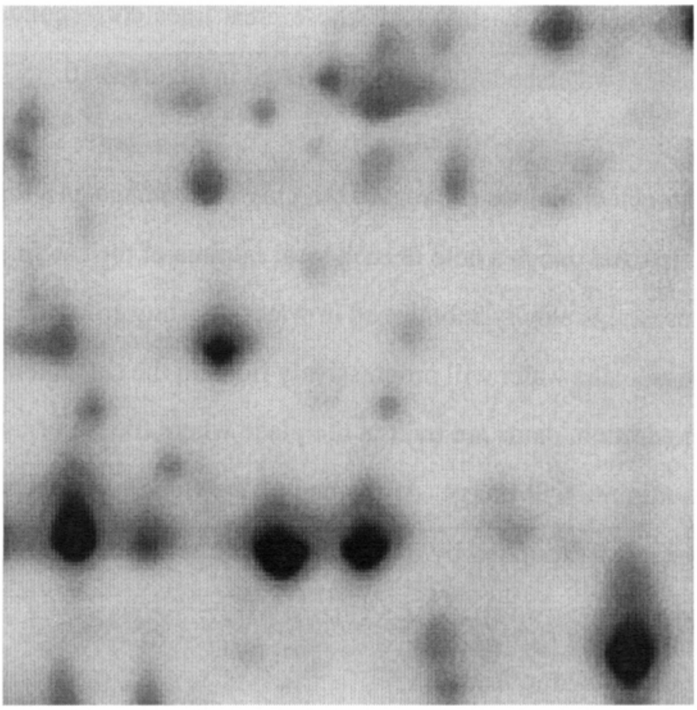


Figure 2-11 Original 2D gel electrophoresisimage

A marker is a connected component belonging to an image. From the result of the watershed processing to the original 2D gel electrophoresis image, we have the internal markers associated with spots, and the external markers associated with the background. The points along the watershed lines are good candidates for the background because they pass along the highest points between neighbouring markers.

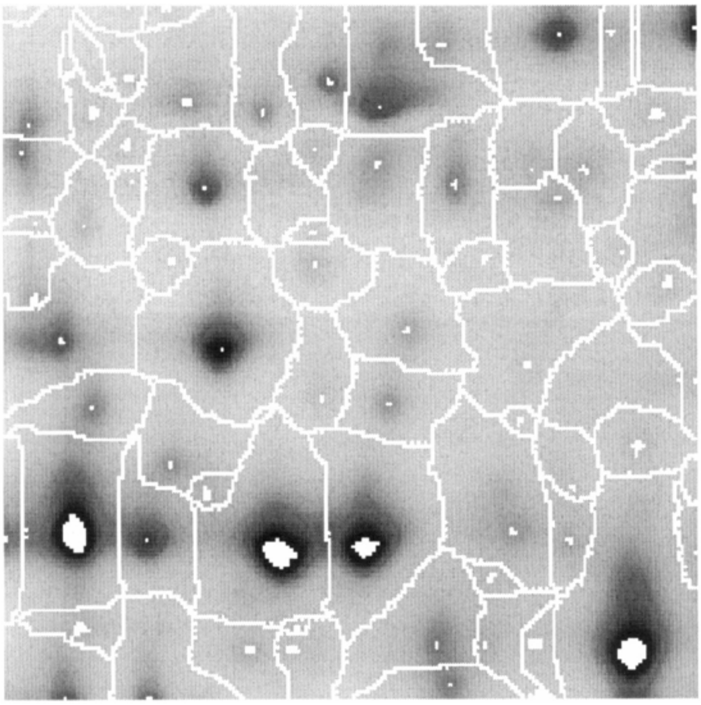


Figure 2-12 Processed image with internal and external markers

From the observation of original image Figure 2-11 and processed image Figure. 2-12 with internal and external markers, the local minima are signed as internal markers and the watershed lines are signed **as** external markers.

The external markers effectively partition the image into regions with each region containing a single internal marker and part of the background. The problem is thus reduced to partition each of these regions in two parts: the spots and their background.

We continue to apply the watershed segmentation algorithm to each individual region in the gradient of the 2D gel electrophoresis image shown as Figure 2-13. And the algorithm is then restricted to operate on a single watershed, which contains the marker in the particular region.

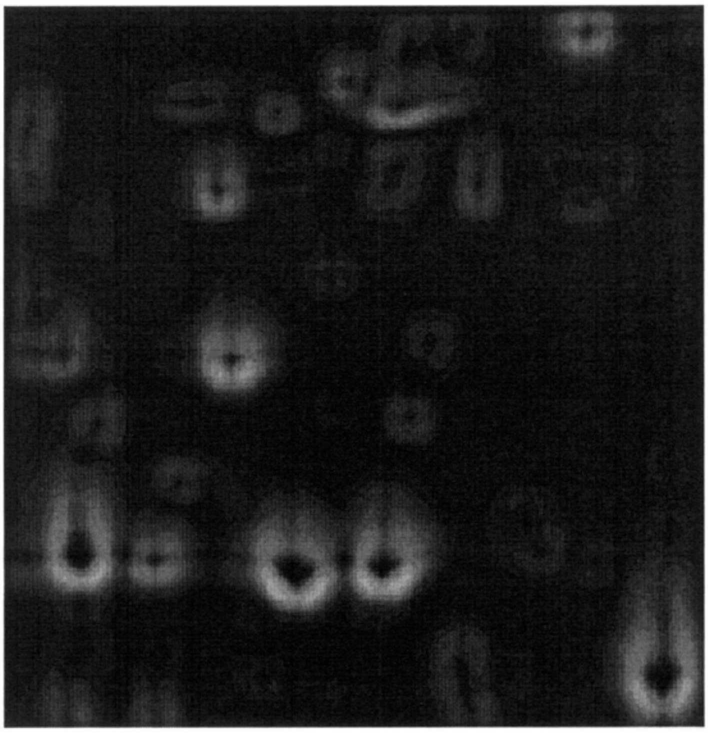


Figure 2-13 Gradient image

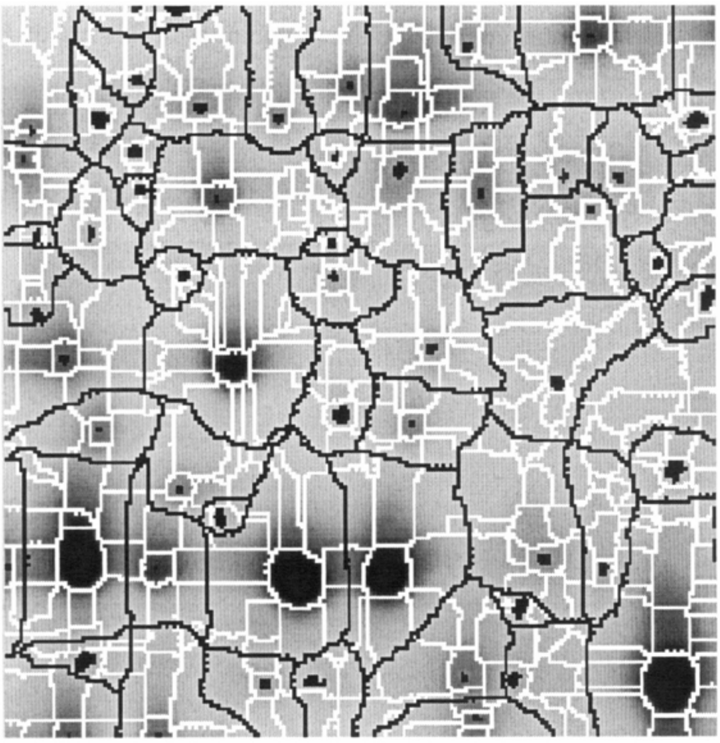


Figure 2-14 Image with gradient watersheds and markers

In Figure 2-14, the black lines and small spots (not original protein spots) are respectively external and internal markers. In every region which is surrounded by the external markers, the weight lines inside the region are the watershed segmentation results of the gradient image corresponding to the region. Watershed lines separate the regions of gradient image into different sub-regions as if they separate the original image into different regions.

To detect the boundary of the spots, considering the sub-regions in every region, one of the sub-regions must contain the internal marker of this region (shown in Figure 2-14). Therefore, the internal maker helps to note a sub-region that can represent a spot inside the region.

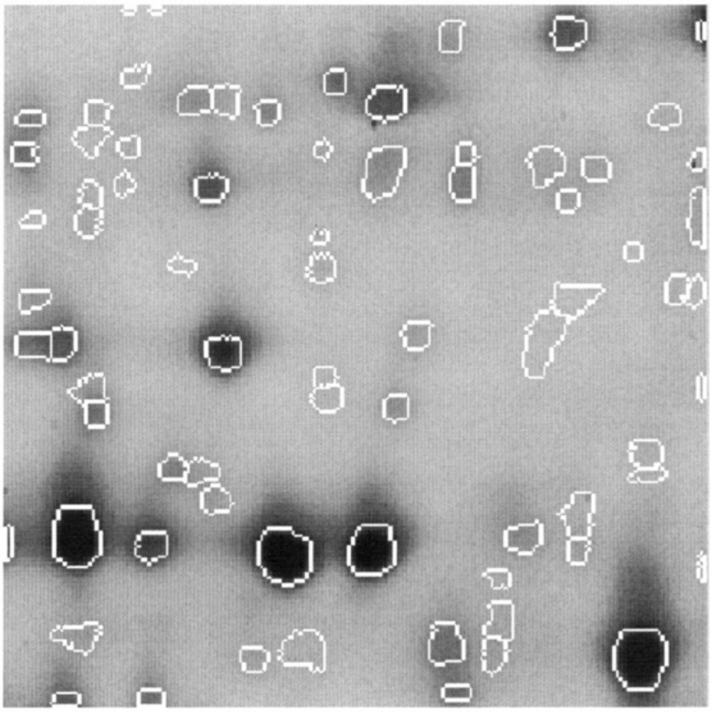


Figure 2-15 Result of spots detection

In Figure 2-15, the boundaries of the detected spots are obtained from the boundaries of sub-regions in the respectively regions.

For the basic watershed algorithm, one disadvantage is the tendency for over segmentation due to noise creating false minima. The two main solutions are:

(1) Marker controlled watershed. (2) Region merging. Another disadvantage **is** the limitation of one internal marker in each sub-region in the gradient image, which is resulted in the missing of some potential spots. We will give the detail of this issue in the following parts.

2.4 Summary

Most of the spots detection methods stated above are very sensitive to noise in the image. Currently, the most popular technique for spots segmentation and spots detection is the watershed transform due to its robustness to noise. Based on the principle of watershed algorithm, there **is** only one internal marker in every region which is segmented by watershed lines, because in gray value image, the marker is always the lowest part in every region.

However, by careful observation of the 3D image and the section image of some spots, we find that sometimes there are more than one spots in every region and some of the spots may be missed due to the only one marker inside. One major disadvantage of watershed transform used in spots detection is the limitation of internal marker. Several protein spots maybe merge together with only one peak remained and the peak is considered as the only internal marker. Therefore, we try to introduce some clustering based methods to overcome the disadvantage of watershed transform for the application to the analysis of two-dimensional gel electrophoresis image.

Some appropriate graphical techniques will be selected as pre-processing methods for the 2D gel electrophoresis images.

3. Clustering Based Spots Detection

3.1 Background

As stated above, the use of watershed to detect the spots in 2D gel electrophoresis image relies on local minima, also called internal markers, in different regions. The internal marker is the lowest part of the regions, so one region contains only one marker.

In the processing of making 2D electrophoresis gel, the protein spots move according to their iso-electric point and molecular weight. After moving, the spots will stay at the certain positions respectively. Because of the viscidity of the protein spots, if two protein spots are close enough, they will stick to each other. Therefore, if we make the gray values of protein spots in 2D gel electrophoresis image to be the third dimension, the two spots are considered to be two hills and they should have two peaks in normal situation. With the adherence of the two spots, sometimes there will be only one peak remained. Usually the remained peak corresponds to the darker spot of the two. As a result, only one spot can be detected in such regions by using watershed segmentation algorithm.

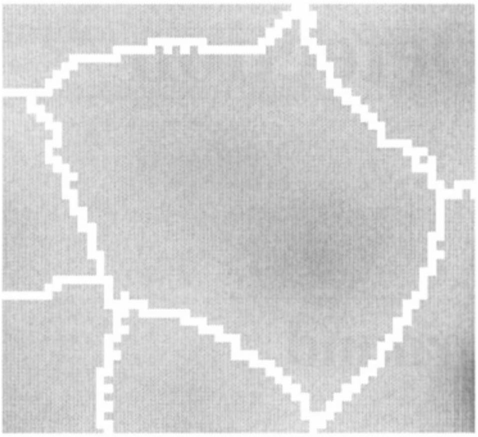


Figure 3-1 Result of watershed segmentation

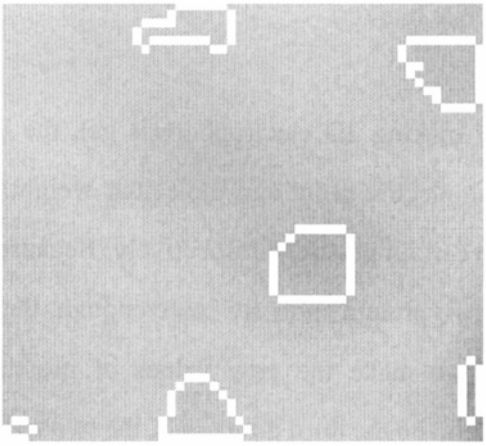


Figure 3-2 Result of spots detection

In Figure 3-2, at the center of the 2D gel electrophoresis image, only one spot has been detected. But from the result of watershed segmentation Figure 3-1, another potential spot can be found on the top and left of the detected spot.

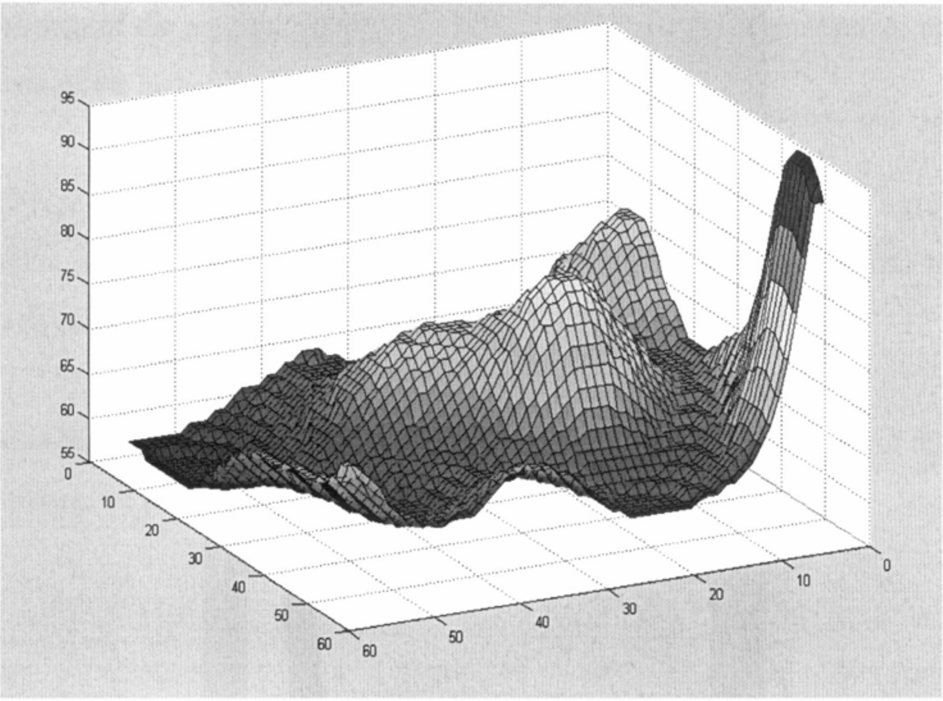


Figure 3-3 3D profile of 2D gel electrophoresis image

Figure 3-3 is the 3D profile of original 2D gray level image. At the center of the image, two hills adhere to each other. The higher hill corresponds to the darker spot.

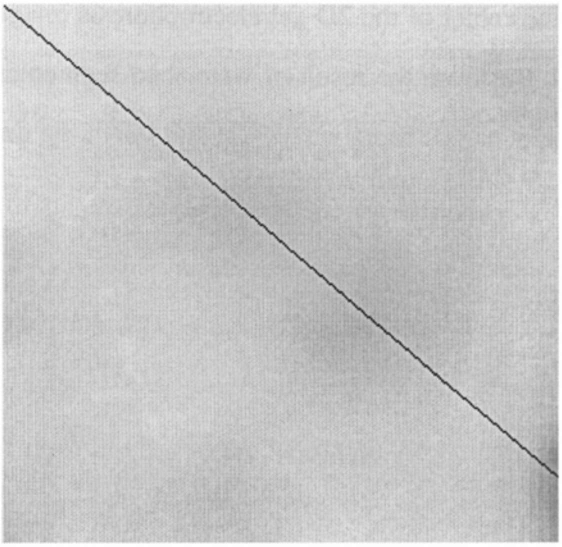


Figure 3-4 Original image with transect line

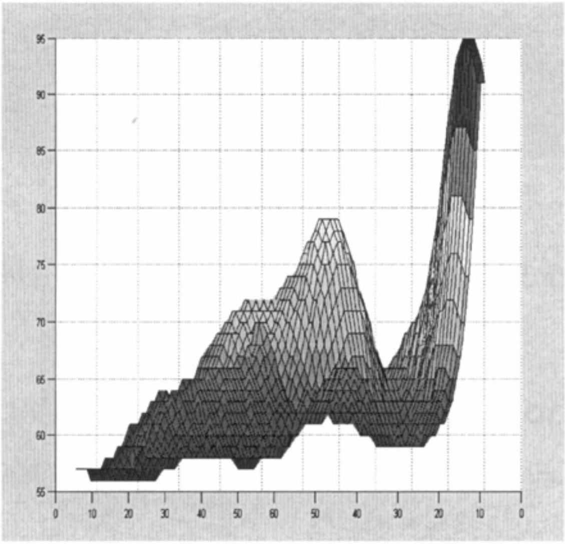


Figure 3-5 Transect image of the spots

The image of Figure 3-5 is the transection of the original 2D gel electrophoresis image, which is cut by the black line in the Figure 3-4. The two hills are integrated together with only one peak remained. Therefore, referring to the result of watershed segmentation, only the peak is considered to be the local minima of the region. With the use of gradient watershed segmentation, the edge of the higher hill is determined by the internal marker.

In Figure 3-5, at the center of the two images, the lighter spot also has a gray value that is evidently different from the background, so it should also be confirmed as a valid spot.

How to detect the missed spot of the two in one region? In the next part of this chapter, we propose new methods to solve this problem.

3.2 Clustering Methods

Based on the properties of 2D gel electrophoresis image, the black protein spots may be considered as hills in a 3D image, with the gray value of the protein spots corresponding to the height of the respective hills. Also, the markers of the spots correspond to the peaks of these hills. Therefore, we consider that the clustering methods can be used to find the peaks of integrated hills and then apply it to the shape detection of protein spots. In this section, we will show the detail of the clustering techniques when they are used in the protein spots detection.

3.2.1 Mountain Clustering

Mountain clustering method [16] is an effective method to estimate the cluster prototypes. This method can provide satisfactory results when it is used in applications that need to approximate cluster centers. In addition, mountain cluster method is less sensitive to noise than some other clustering methods [39].

Assume there are n data points $\{x_1, x_2, ...x_n\}$ in the w dimensional space R^w. We denote x_{kj} as the $j-th$ coordinate of the $k-th$ point, where $k=1,2,...,n$ and $j=1,2,...,w$. Without loss of generality, we restrict the w dimensional space R'' to a w dimensional hypercube $I, \times I, \times ... \times I,$ of R'' where the interval $I_j = 1,2,...,w$ is defined by the range of the coordinate x_{kj}.

$$I_j = [\min_k(x_{kj}), \max_k(x_{kj})] \qquad (3.1)$$

Therefore, the hypercube contains all the points of the data set $\{x_1, x_2, ...x_n\}$. Then we will discretize each of the intervals I_j into r_j equidistant points.

Such a discretization forms a w dimensional grid in the hypercube. The significance of the discretization is that the grid nodes are potential cluster centers. In fact, the process of obtaining the set of potential cluster centers, the equidistant nodes, need not be based on a uniform griding of the space. We can use variable griding of the space, which is resulted in $N = r_1 \times r_2 \times ... \times r_w$ grid points N_i , $i = 1,2,...,N$ in R^w . We construct a mountain function which is defined over the set grid nodes by adding an amount to each node proportional to its distance from the data point. Then we will use the distance measure $d(\mathrm{x}_k, N_i)$ to score the contribution of the data point x_k to the value of the mountain function at the grid note N_i . We assign higher scores to the nodes that are closer to a data point. The following mountain function provides the value at the vertex point N_i ,

$$M(N_i) = \sum_{k=1}^{n} e^{-\alpha d(\mathrm{x}_k, N_i)} \tag{3.2}$$

where a is a positive constant. It is evident from Equation 3.2 that the grid points which have more data points nearer to them will get high values from the mountain function. The values of the function $M(N_i)$ can be looked as heights of a mountain range. Hence, we call the function $M(N_i)$ mountain function. The value of the mountain function is closely related to the density of data points in the neighbourhood of the grid point. It can be seen to represent the potential ability of any of the grid points to be a cluster center. Therefore, the grid points with a high value of mountain function are more suitable to be the potential cluster centers of the given data set [17]. The mountain function value can be used as indicator of the cluster centers.

3.2.2 Estimation of Cluster Centers

If the mountain function for each grid point is calculated, the grid point with the maximum mountain function value, the mountain range peak, is selected. If there are more than one maximum, we will select randomly one of them. We denote the maximal value of the mountain function as

$$M^* = Max\ [M(N_i)] \tag{3.3}$$

$$N^* = Arg \max_{1 \le i \le N} [M(N_i)] \tag{3.4}$$

M^* is the global maximum of the mountain function. Let the grid point N^* indicate the maximal value M^* of the mountain function. This peak note N^* is selected as the first cluster center.

The peak N^* is usually surrounded by many grid points that also have high values of mountain function. This is due to the process of constructing the mountain function and the inherent continuity of the hypercube.

In order to find the other cluster centers, the previously built mountain function is discounted to eliminate the effects of the detected centers. To accomplish this task, we subtract a value at each node from the current mountain function. The amount subtracted at each grid point is proportional to the distance of the point from the maximal. Yager [16] suggested the following update equation for the mountain function

$$M_{new}(N_i) = M_{old}(N_i) - M^*_{old} e^{-\beta d(N^*_{old}, N_i)} \tag{3.5}$$

In the above equation, $M_{new}(N_i)$ is the new mountain function, $M_{old}(N_i)$ is the previous mountain function, N^*_{old} is the previous detected center and β is a positive constant.

The process of equation 3.5 is used to destroy the mountain function. It will guarantee that those grid points closer to the old identified cluster center will have their mountain value more strongly reduced. By using equation 3.5, new cluster centers are found and the mountain function is reduced until the level of the current maximum $M*$, compared with the first maximum M^*_1 become too low. The process of destroying the mountain function will be stopped when

$$\frac{M^*}{M_1^*} < \delta \tag{3.6}$$

where δ is a positive constant less than 1.

If the value of the parameters α , β , δ are selected appropriately, the algorithm will perform well. But it is very difficult to obtain the proper value of these parameters that can work for all kinds of data sets. The accuracy of this algorithm depends on the fineness of the grid, but with increase in the fineness of the grid, the algorithm becomes computationally more expensive. In addition, with the increase in the dimensionality of the data the computational will also increase.

In the next part, a variation of the mountain clustering method called subtractive clustering will be described. It is computational less expensive than the mountain clustering.

3.2.3 Subtractive Clustering

The mountain clustering method is computationally expensive and the amount of computation will grow rapidly if the dimensions of the data increase. To solve this problem, Chiu [18] suggested using each data point itself, unlike the mountain clustering that uses the new grid points, as the potential cluster centers.

For the subtractive clustering method, the mountain function is calculated for each data point by

$$M(x_i) = \sum_{k=1}^{n} e^{-\alpha d(x_k, x_i)} \tag{3.7}$$

and the discounting function of the mountain is

$$M_{new}(x_i) = M_{old}(x_i) - M^*_{old} e^{-\beta d(x^*_{old}, x_i)} \tag{3.8}$$

In the equation **3.8,** x^*_{old} is the previouslydetected cluster center. This method **is** somewhat similar to the method of mountain function, but because the grid points are not used, the prospective cluster centers only depend on the data points themselves. The amount of original data is always much less than that of the constructing grid points of the mountain clustering, so the time used for computation of subtractive clustering is also much reduced.

3.3 Clustering Based Spots Detection

The disadvantage of the basic watershed algorithm applied to spots detection is that it depends on the local minima when performing of marker controlled segmentation. Based on the principal of the algorithm, there is only one internal marker in each region. In the gradient image, only one of the sub-regions in every region can be selected as the boundary of the spot. The problem is that sometimes there are more than one spots in one region as we have seen in the 3D profile images of previous section. Due to the phenomenon of combination, two or more spots incorporate together with only one peak remained in the 3D profile image. When we use watershed segmentation, the peak is used as a marker for spots detection. If the marker of respective spots is missed, the protein spots cannot be detected with watershed segmentation method. To solve this problem, subtractive clustering method will be used with watershed segmentation algorithm, to avoid the loss of the potential markers and get more accurate result.

3.3.1 Subtractive Clustering for Spots Detection

In two-dimensional image space, the original mountain function is

$$M(x_i, y_i) = \sum_{k=1}^{n} \exp(-\frac{(x_k - x_i)^2 - (y_k - y_i)^2}{2a^2}) \qquad (3.9)$$

In equation 3.9, x_i and y_i denote the position of the objective pixel i in the image. $M(x_i, y_i)$ is the mountain value at pixel i (i =1,2,..., n) and n is the number of pixels in the region. The constant a is effectively a radius defining a neighbourhood and data points outside this radius have much less influence on the potential [19].

From equation 3.9 we can find that it evidently uses the form of subtractive clustering method. Because the positions of pixels in 2D gel electrophoresis image also construct a whole grid, all the data points themselves can also be looked as the grid points for this data space. Therefore, this function has the similar form with mountain clustering method.

Because the pixels of 2D gel electrophoresis image also have intensities that represent as gray value, we will incorporate them to build the mountain function as:

$$M(x_i, y_i) = \sum_{k=1}^{n} I_k \exp(-\frac{(x_k - x_i)^2 + (y_k - y_i)^2}{2a^2}) \qquad (3.10)$$

In equation 3.10, I_k is the gray value of pixel k and $\boldsymbol{a}$ is a positive constant.

The gray value is a way to introduce a respective weight to every pixel. By the form of this function, the biggest mountain value will appear at the collection center of many pixels with higher gray value but not the background.

The form of update equation of the mountain function **is** unchanged like this

$$M_{new}(x_i, y_i) = M_{old}(x_i, y_i) - M^*_{old} \exp\left(-\frac{(x^*_{old} - x_i)^2 + (y^*_{old} - y_i)^2}{2b^2}\right) \quad (3.11)$$

Where M^*_{old} is the mountain value of the cluster center before update and x^*_{old} and y^*_{old} are the coordinates of this center and b is a positive constant.

3.3.2 Preprocessing Methods

Image enhancement is widely used in the field of computer graphics with smoothing, contrast enhancement and so on. It is one of the most important tools for image analysis [20]. Due to the noise and faintness of the 2D gel electrophoresis image, image enhancement is necessary for preprocessing. We will introduce the smoothing and contrast enhancement methods used before spots detection. In addition, a threshold method for the pre-estimation of the protein spots regions will also be introduced.

1. Smoothing:

Due to the acquisition and susceptibility to dust, 2D gel electrophoresis images need to be smoothed to suppress the noise inherent in them. The two most frequently used smoothing techniques are median and Gaussian filters.

To suppress the impulse noise of 2D gel electrophoresis images that is introduced by the image capture devices such as CCD cameras [21], median filtering is a quite effective tool. The median filter is a kind of nonlinear filter, which was firstly used by J.W.Jukey in one dimensional signal processing and was then introduced into two dimensional image processing [22].

For mean filter, the output of each pixel is set to the mean of the pixel values in the neighbourhood of the corresponding input pixel. However, with median filter, the value of an output pixel is determined by the median of the neighbourhood pixels, rather than the mean. The median is much less sensitive than the mean to impulse noise (called outliers) in the image. Therefore, median filtering is able to remove these outliers without reducing the sharpness of the image.

For example, if the window of median filter has 5 pixels, the gray values are 80, 90, 220, 110 and 130. Then the gray value of the center of the window is replaced by the median 110. Finally, the outlier 220 is removed by this filter.

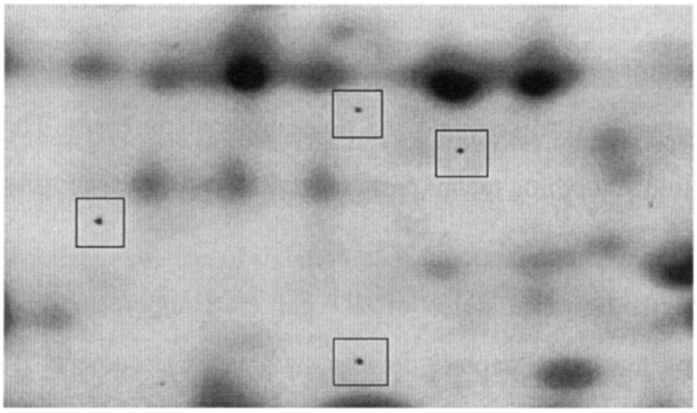


Figure 3-6 Original 2D gel electrophoresis image

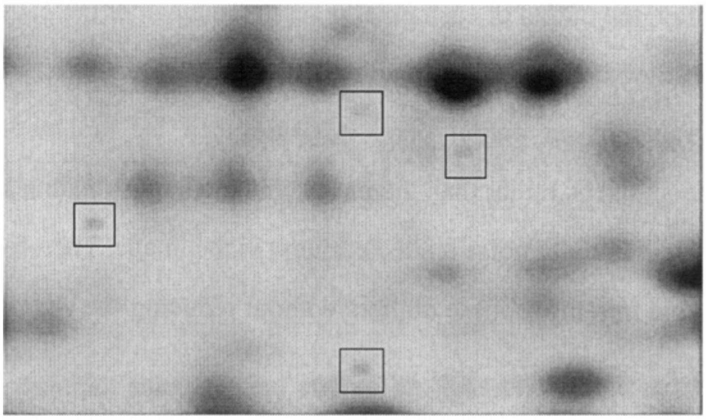


Figure 3-7 Result of mean filtering

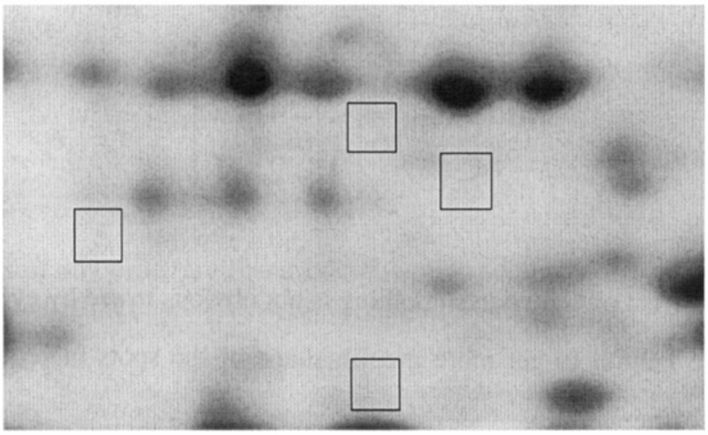


Figure 3-8 Result of median filtering

Figure *6* is an original 2D gel electrophoresis image. From the filtering results of the 5 by 5 medianx filter (Figure 3-7) and results of 5 by 5 mean filters (Figure 3-8), we can find that the impulsive noises shown as evident small black point in the segment of an original 2D gel electrophoresis image are successfully removed by the median filtering. But as the result of mean filtering, the impulse noise is not completely removed and the whole image is more obscure than before.

In addition, to suppress the statistical Gaussian noise inherent in the 2D gel electrophoresis images [23], Gaussian smoothing is used in this project. Gaussian smoothing of a 2D gel electrophoresis image is performed by convolving it with a 3 by 3 mask operator [24]:

$$mask = \frac{1}{16}\begin{vmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 1 \end{vmatrix}$$

Though the effect of Gaussian smoothing is not obvious to see by eyes, the use of it will be helpful to get more precise shape of the spots in processing of watershed segmentation.

2. Contrast Enhancement:

The principal objective of contrast enhancement is to highlight the valuable details in an image or to enhance the details that have been blurred, either in error or as a natural effect of a particular method of image acquisition. In 2D gel electrophoresis image, spots are the valuable detail, including the intensity and edge of them. The processing of contrast enhancement will make the details of spots sharper, thus the result of spots detection will be more accurate.

Depending on the features of the 2D gel electrophoresis image, Laplacian operator is usually used for enhancement. The Laplacian approach is basically consists of defining a discrete formulation of the second-order derivative and then contrasting a filter mask based on that formulation. The Laplacian filter is an isotropic filter whose response does not depend on the direction of the discontinuities in the image [25].

For a function f(x, y) of two variables, the Laplacian is defined as

$$\nabla^2 f = \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} \tag{3.12}$$

Derivatives of any order are linear operations, so the Laplacian is also a linear operator.

To make it useful for digital image processing, the equation needs to be expressed in discrete form. Taking into account that we have two variables in the digital image, the first-order derivative is

$$\begin{cases} \dfrac{\partial f(x,y)}{\partial x} = f(x,y) - f(x-1,y) \\ \dfrac{\partial f(x,y)}{\partial y} = f(x,y) - f(x,y-1) \end{cases} \tag{3.13}$$

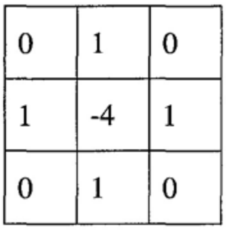
so the second-order derivative is

$$\begin{cases} \dfrac{\partial^2 f(x,y)}{\partial x^2} = f(x+1,y) + f(x-1,y) - 2f(x,y) \\ \dfrac{\partial^2 f(x,y)}{\partial y^2} = f(x,y+1) + f(x,y-1) - 2f(x,y) \end{cases} \tag{3.14}$$

The digital implementation of the Laplacian is obtained by summing the two components:

$$\nabla^2 f = f(x+1,y)+f(x-1,y)+f(x,y+1)+f(x,y-1)-4f(x,y) \tag{3.15}$$

The equation above is implemented using the mask shown in Figure 3-9, which will give an isotropic result for filtering of 2D gel electrophoresis image.



0	1	0
1	-4	1
0	1	0

Figure 3-9 Filter mask of Laplacian

Because the Laplacian is a derivative operator, it highlights gray value discontinuities in an image and emphasizes the regions with quickly varying gray values. Laplacian operation simply adds the original and Laplacian images. But if the center coefficient is negative, it subtracts, rather than adds the Laplacian image to get a sharpened result.

$$g(x,y) = f(x,y) + k\,|\nabla^2 f(x,y)| \tag{3.16}$$

In this equation, f and g are the image before and after sharpening, and k

is the sharpening coefficient. The value of k depends on further processing of spots detection. Because the valuable details of 2D gel electrophoresis image is black spots and the gray value of the spots is lower than the background, ***k*** should be minus for enhancing the black spots.

In the figures below, the black spots in the 2D gel electrophoresis image after contrast enhancement (Figure 3-11) are more evidently than the original image (Figure 3-10) after the contrast enhancement processing. After many experiments, we find that -3 is an appropriate value for ***k*** to get a satisfying result for spots detection.

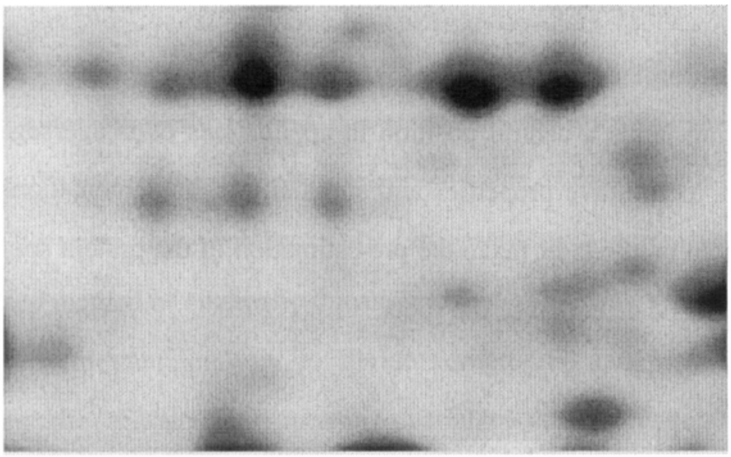


Figure 3-10 Original image after smoothing

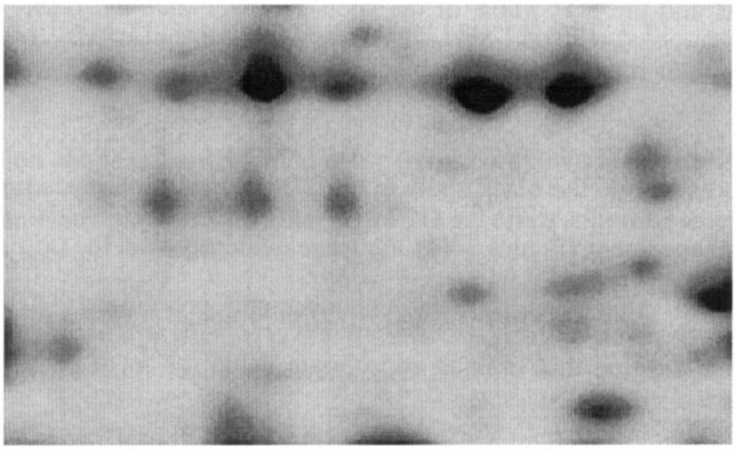


Figure 3-11 Result of contrast enhancement

3. Thresholding:

Thresholding holds a central position in applications of image segmentation because of its intuitive properties and simplicity of implementation. We will use threshold methods to make the pre-estimation of the protein spots regions before the shape detection of the spots.

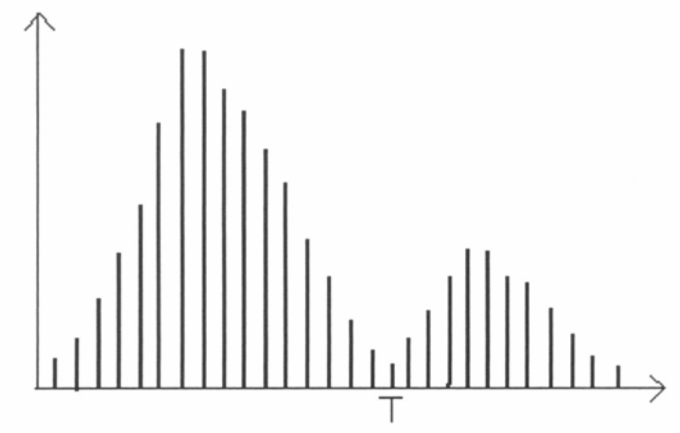


Figure 3-12 Gray level histogram

Suppose that the gray level histogram shown in figure 3-12 corresponds to an image $f(x,\text{y})$ which is composed of light objects on a dark background, in such a way that object and background pixels have gray levels grouped into two dominant modes. The thresholding method for extracting the objects from the background is to select a threshold $\boldsymbol{T}$ to separate the two different modes. The point (X,y) for which f(x, y)$>T$ is classified as an object point and otherwise it is classified as a background point.

For more general case, it is not that easy as shown in figure 3-12 and the threshold is not evidently to find due to the noise, unequal peaks or flat histograms. Therefore, the method to select the value of threshold is also more complex. To solve this problem, Ostu [26] proposed a nonparametric and unsupervised method of automatic threshold selection for image segmentation. He used discriminant analysis to divide foreground and background by maximizing the discriminant measure variable that is the measure of separability of the resultant classes in gray levels.

Suppose that the gray levels of the original image is $\boldsymbol{L}$ and there are n_i pixels whose gray value are i, so the amount of all the pixels in the image is

$$N=n_0+n_1+...+n_{L-1} \qquad (3.17)$$

Normalize the histogram, then $p_i = \frac{n_i}{N}$ and $\sum_{i=0}^{L-1} p_i = 1$.

It is divided to two classes depending on the gray value:

$$C_0 = (0,1,2,...t) \quad \text{and} \quad C_1 = (t+1, t+2,...L-1).$$

The probabilities of class C_0 and class C_1 are given by:

$$\omega_0 = P_r(C_0) = \sum_{i=0}^{t} p_i = \omega(t) \tag{3.18}$$

$$\omega_1 = P_r(C_1) = \sum_{i=t+1}^{L-1} p_i = 1 - \omega(t) \tag{3.19}$$

$$\mu_0 = \sum_{i=0}^{t} ip_i / \omega_0 = \mu(t) / \omega(t) \tag{3.20}$$

$$\mu_1 = \sum_{i=t+1}^{L-1} ip_i / \omega_1 = \frac{\mu_T(t) - \mu(t)}{1 - \omega(t)} \tag{3.21}$$

In the equations: $\mu(t) = \sum_{i=0}^{t} ip_i$ and $\mu_T = \mu(L-1) = \sum_{i=0}^{L-1} ip_i$.

For any t, we can get: $\omega_0 \mu_0 + \omega_1 \mu_1 = \mu_T$ and $\omega_0 + \omega_1 = 1$.

The variances of class C_0 and class C_1 are given by:

$$\sigma_0^2 = \sum_{i=0}^{t} (i - \mu_0)^2 p_i / \omega_0 \qquad (3.22)$$

$$\sigma_1^2 = \sum_{i=t+1}^{L-1} (i - \mu_1)^2 p_i / \omega_1 \qquad (3.23)$$

The variance inside the two classes is

$$\sigma_\omega^2 = \omega_0 \sigma_0^2 + \omega_1 \sigma_1^2 \qquad (3.24)$$

The variance between the two classes is

$$\sigma_B^2 = \omega_0 (\mu_0 - \mu_T)^2 + \omega_1 (\mu_1 - \mu_T)^2 = \omega_0 \omega_1 (\mu_0 - \mu_1)^2 \qquad (3.25)$$

The total variance is

$$\sigma_T^2 = \sigma_B^2 + \sigma_\omega^2 \qquad (3.26)$$

The discriminant criterion about t is

$$\eta(t) = \sigma_B^2 / \sigma_T^2 \qquad (3.27)$$

The threshold will make the best separation of class C_0 and class C_1. It can also maximize the discriminant criterion $\eta(t)$, so the best threshold is equal to:

$$t^* = Arg \max_{0 \le t \le L-1} \eta(t) \qquad (3.28)$$

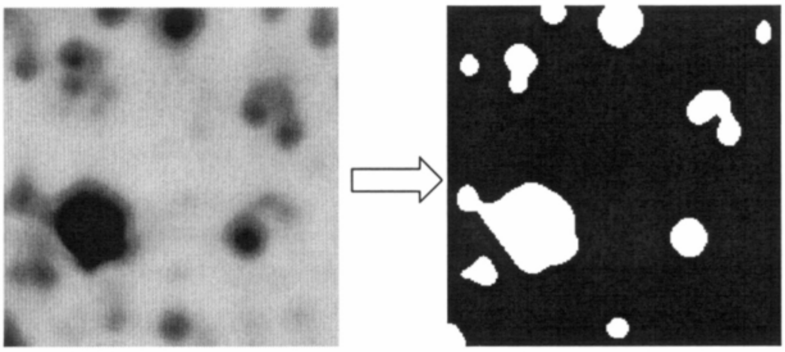


Figure 3-13 2D gel electrophoresis image to Ostu thresholding

But if Ostu's threshold method is directly used in the whole 2D gel electrophoresis image, the effect is not satisfying enough, because the intensities of the spots and background across the image are very different. We can find from figure 3-13 that many spots are missed because of their obscure intensity. The intensity of spots in some regions is less than the intensity of background in other regions. The threshold method only considers the gray value of the whole image, but not the distribution of the gray value and shape or other features of the target objects. Therefore, threshold method usually fails in the image regions that are not uniform enough.

If the relation of the target objects and the background are not complicated, as in the segmented regions of the watershed result for the 2D gel image, each of which contains only a few similar spots, Ostu's threshold method can successfully separate all the potential spots regions from the background.

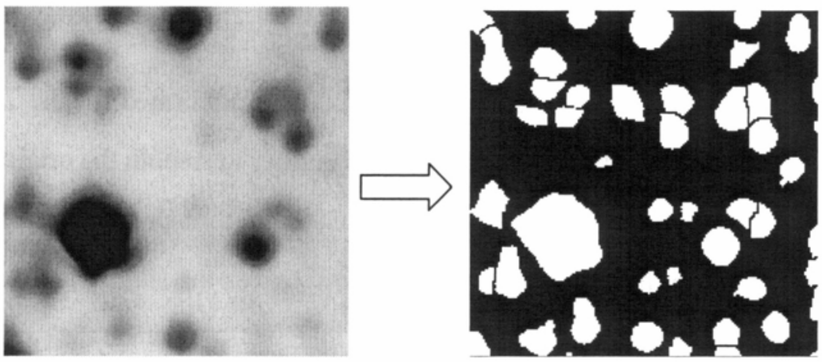


Figure 3-14 Ostu's thresholding result from all the regions of watershed segmentation results

With the watershed segmentation result from the original image, Ostu's thresholding method can be performed in all the segmented regions. The result of figure 3-14 is shown that all the potential spots regions are detected. Those potential spots regions will be very helpful for our further spots detection method.

3.3.3 Implementation of Clustering Based Spots Detection

In the implementation of clustering based spots detection in this thesis, our testing images are all silver stained 2D protein gel electrophoresis images from biological sample.

1. Preprocessing

Before the spots detection, preprocessing should be performed for the 2D gel electrophoresisimage. First we use 5x5 median filter to smooth the original gel image and remove the impulse noises, which are introduced by the capture devices. Then use the Gaussian smoothing to suppress the statistical Gaussian noises inherent in the 2D gel electrophoresis images. To enhance the smoothed image, the Laplacian operator is performed to get a contrast enhancement image, which is more appropriate for gel segmentation processing.

2. Separate the image With Watershed operation

After preprocessing, watershed operation is performed on the 2D gel electrophoresis image and its gradient transform image. To illustrate the experiment clearly, we will focus on the typical regions which will not get good result from the traditional detection methods.

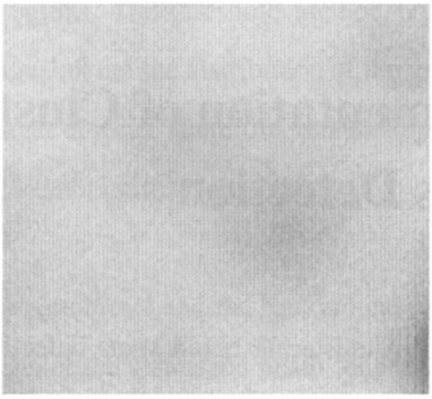


Figure 3-15 Original 2D gel electrophoresisimage

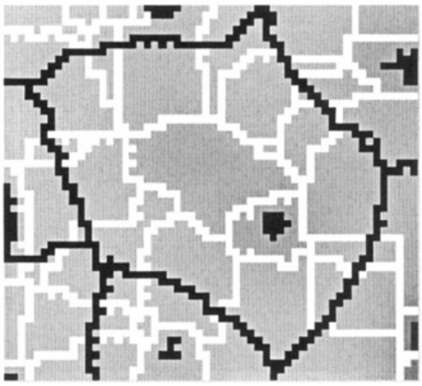


Figure 3-16 Image with watershed lines and markers

In Figure 3-16, the black lines come from the watershed transform to the original 2D gel electrophoresis image (Figure 3-15) and white lines come from the watershed transform to the gradient image of the original in all the regions. The black marks also have the meaning of internal markers and external markers. Our experiments will focus on the biggest region at the center of the image.

3. Find the potential spots regions based on Ostu's threshold method

To estimate all the pixels which are the potential spots pixels inside the region, Ostu's threshold method will be performed on the image. The principal is that the threshold operation is applied inside every region, so the gray value histogram of each region will obtain from the pixels inside the region. By this operation, the region is separated into two parts, which are the objects containing all the potential spots, and the background.

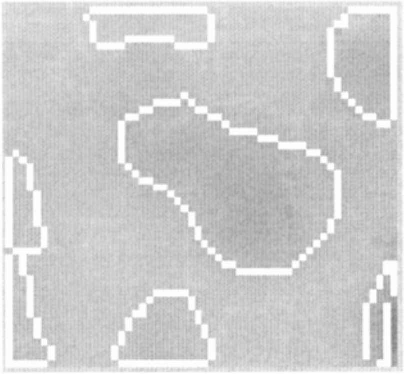


Figure 3-17 Threshold operation result

In Figure 3-17, the object at the center of the image is the potential spots region inside the center area of the watershed image.

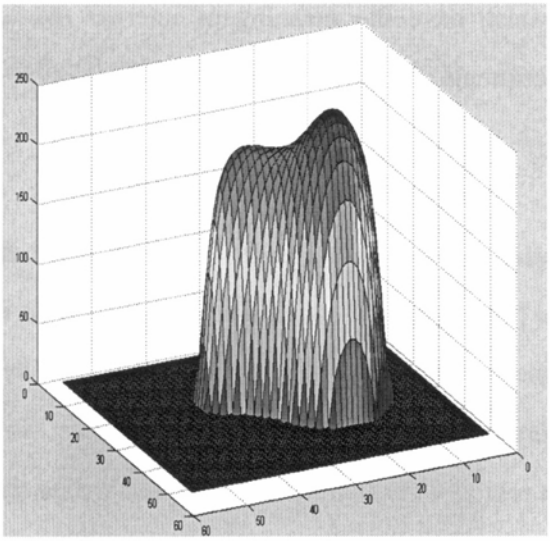


Figure 3-18 3D profile of mountain function result

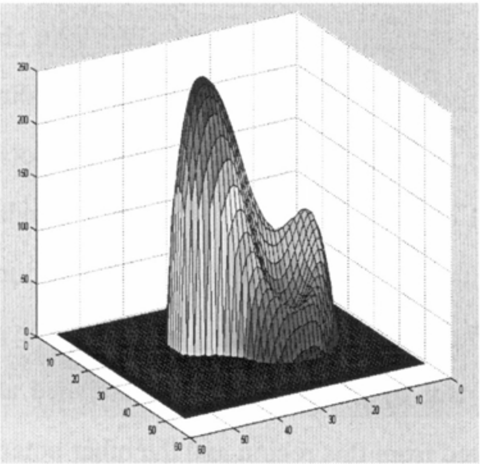


Figure 3-19 3D profile of the updated mountain function result

4. Use subtractive clustering to find the spots centers

From Figure 3-18 and 3-19, we can see that by using the subtractive clustering method on the threshold result at the center area, two center points will be obtained. The highest peaks in Figure 3-18 is subtracted, with the left peak remained in Figure 3-19. The two center black points are shown as the highest peaks respectively in the two 3D profile images.

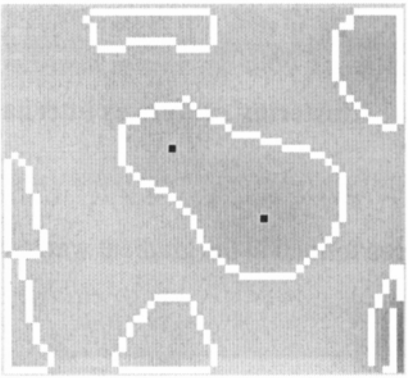


Figure 3-20 Subtractive clustering result

5. Detect the spots in watershed operation results with clustering centers

In Figure 3-20, the clustering centers are shown as black points inside the object. They will be used to mark the potential boundary of spots from the gradient watershed of 2D gel electrophoresis image. If one boundary of the gradient watershed image contains a center point, it should be a boundary of a potential spot in the region. Therefore, the clustering center has the same function with the internal marker. But by using traditional watershed algorithm, only one marker can be obtained from this region and the other potential spot is missed. In our spots detection method, the other potential spots will be marked by the second clustering center.

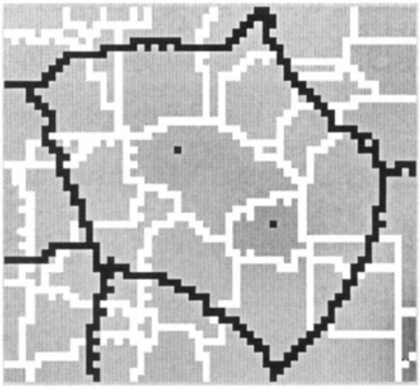


Figure 3-21 Clustering centers as internal marker

In Figure 3-21, two shapes formed in the gradient watershed lines are marked by the black clustering center point.

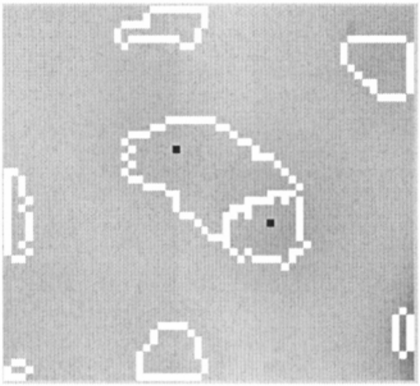


Figure 3-22 Spots detection result

Figure 3-22 is the spots detection result with two potential spots in the central area. This image is also an example shown in the basic watershed segmentation that was introduced in the beginning of this chapter (figure 3-2). But with the basic watershed segmentation, only one spot can be detected in the central area. With our detection method, all the two potential spots in the central area are successfully detected. When our detection method is used in other different typical regions, the experiment results are shown as follows:

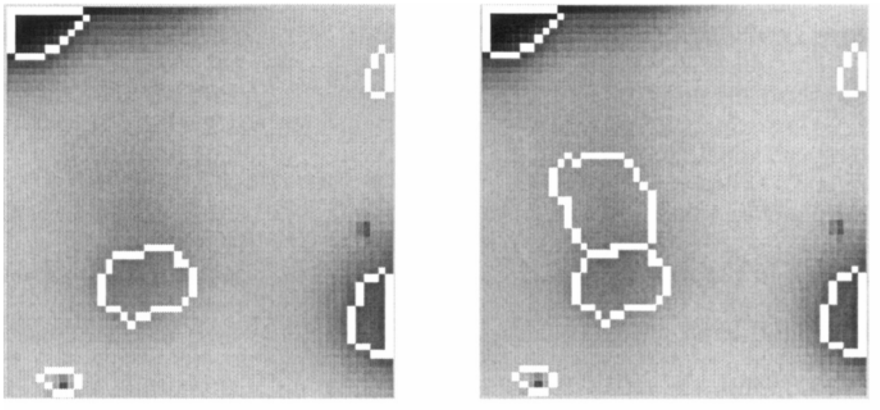


Figure 3-23 (a) Basic watershed detection result with one potential spots missed. (b)Detection result of our method.

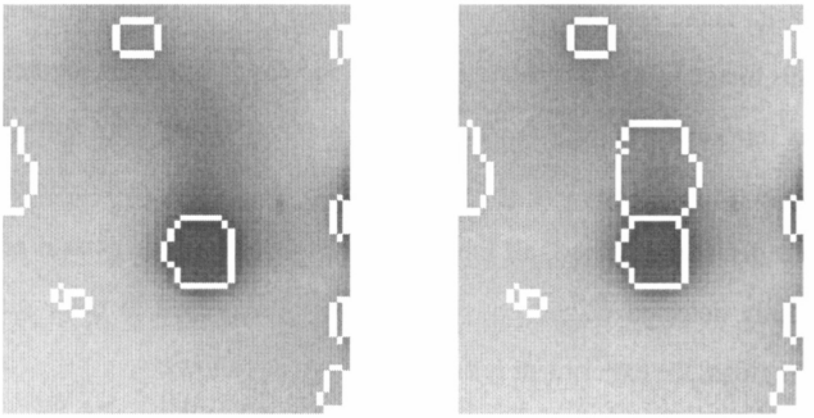


Figure 3-24 (a) Basic watershed detection result with one potential spots missed. (b)Detection result of our method.

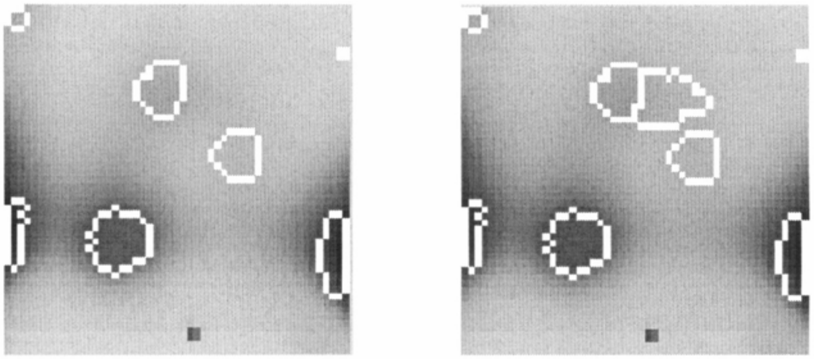
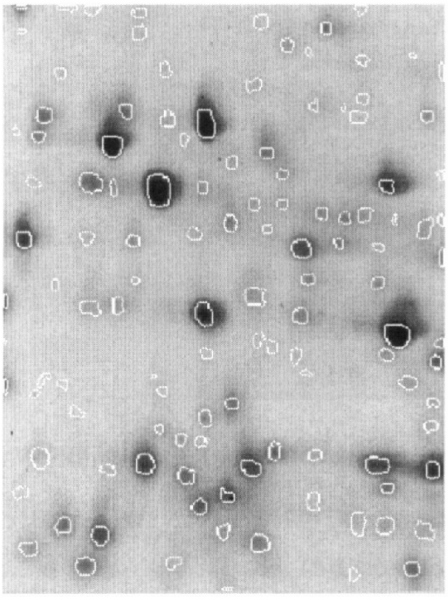
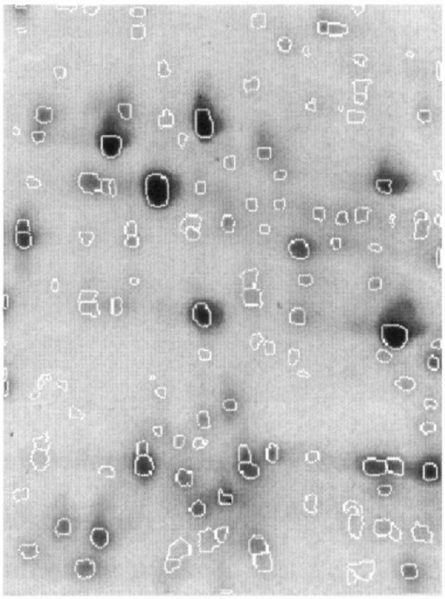


Figure 3-25 (a) Basic watershed detection result with one potential spots missed. (b)Detectionresult of our method.

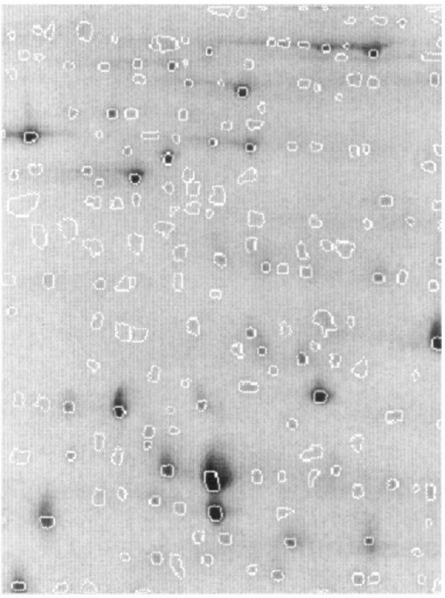
In the cutting figures above, for every figure, out detection method can always find the spot near the central area that is missed by the basic watershed detection method. When it is performed in the different large 2D gel electrophoresis images, the detection results are shown as follows:



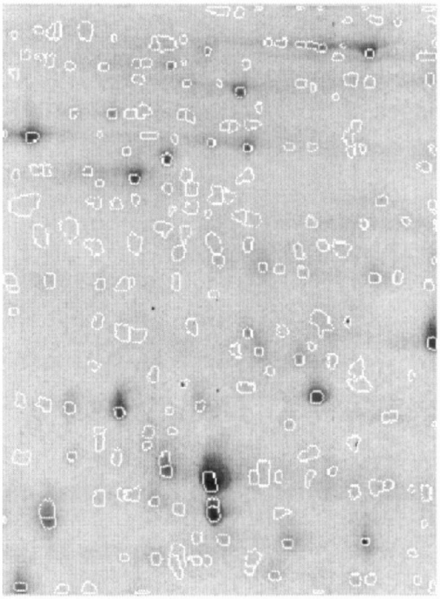
(a1)



(a2)



(b1)



(b2)

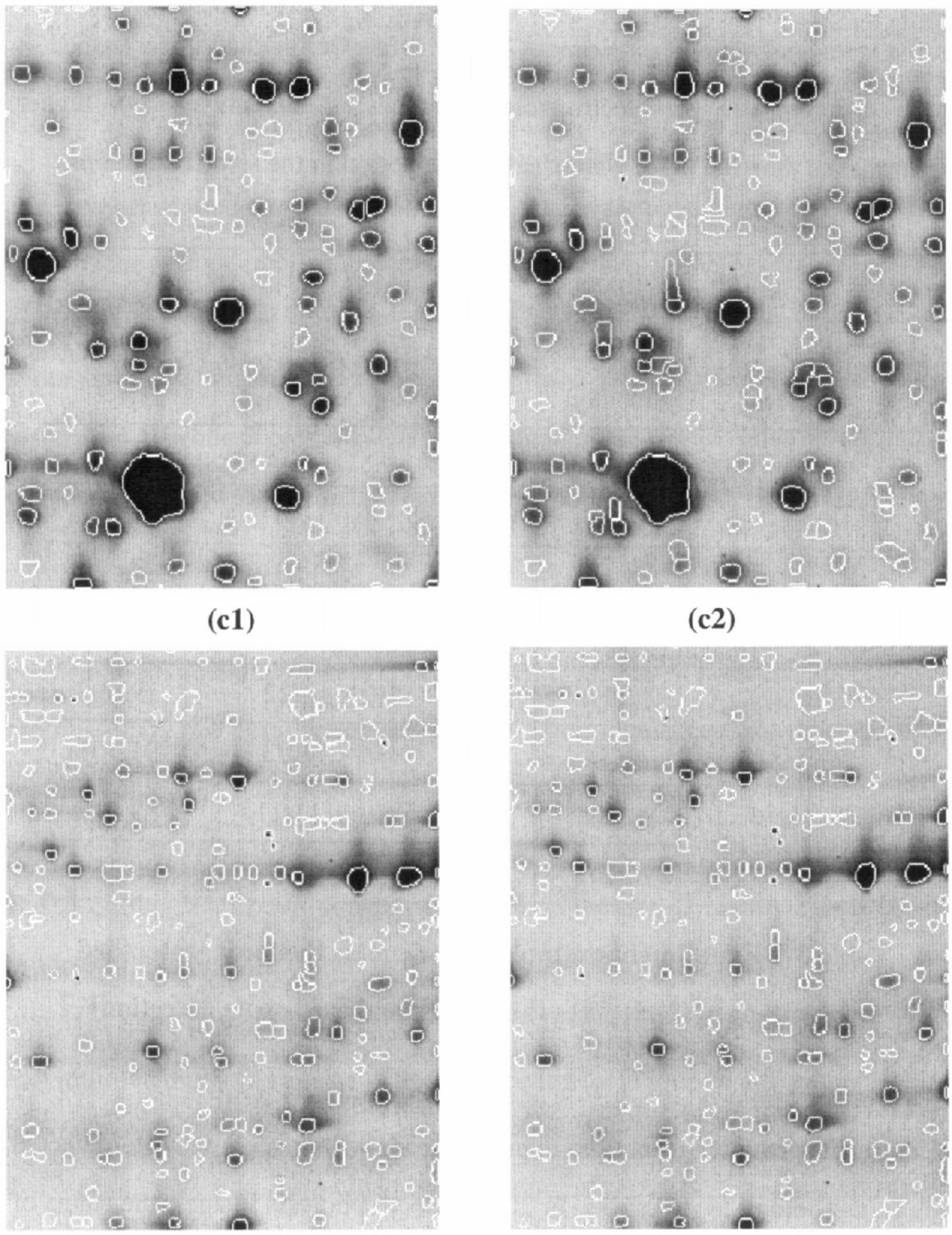
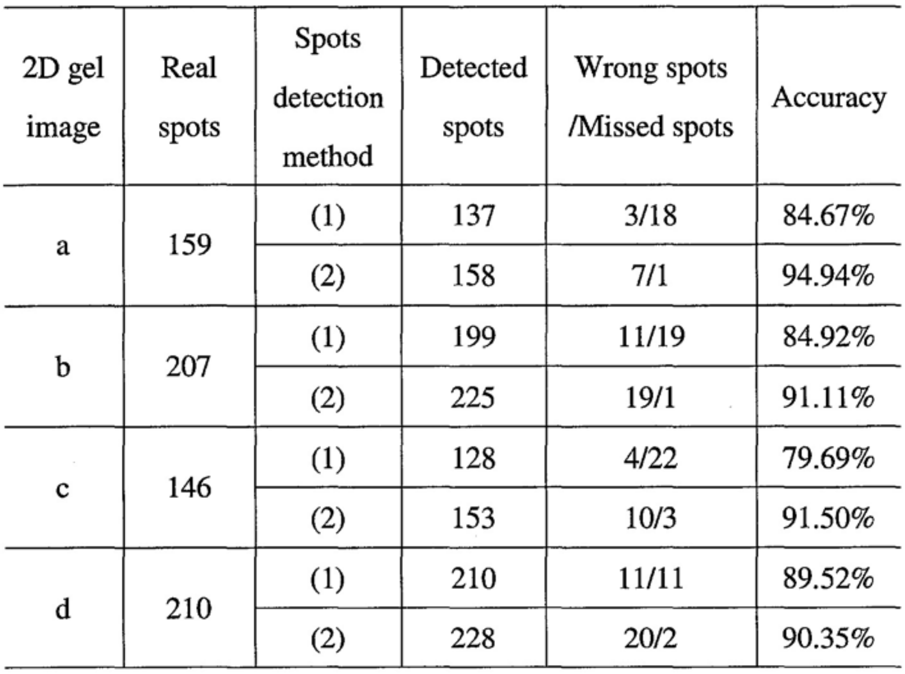


Figure 3-26 Experiment examples [1 is the basic watershed segmentation, 2 is the clustering based spots detection.]

In the statistical table of experiment results, when our method is performed in these different 2D gel electrophoresis images, it has always got higher performance than the basic watershed segmentation in the same conditions.



2D gel image	Real spots	Spots detection method	Detected spots	Wrong spots /Missed spots	Accuracy
a	159	(1)	137	3/18	84.67%
		(2)	158	7/1	94.94%
b	207	(1)	199	11/19	84.92%
		(2)	225	19/1	91.11%
c	146	(1)	128	4/22	79.69%
		(2)	153	10/3	91.50%
d	210	(1)	210	11/11	89.52%
		(2)	228	20/2	90.35%

Table 3-1 Experiment examples [(1) denotes the basic watershed segmentation. (2) denotes the clustering based spots detection]

In these images after detection, there are some wrong spots that are detected but do not belong to real spots. In 2D gel electrophoresis images, the shape of the protein spots should be nearly an ellipse and the ratio of the long axis to the short axis of the ellipse should not very big. Therefore, most of the wrong spots can be deleted because of the abnormal shapes, when they are fitted to the regular ellipse.

3.4 Conclusion

We have illustrated the issues of the traditional watershed algorithm when it is used in the spots detection process. By observation of the 3D image and the section image of some spots, we find that sometimes there are more than one spots in every region, but some of them are always missed by the basic watershed detection method. To solve this problem, we have proposed a clustering based watershed segmentation method for spots detection in 2D gel images. Subtractive clustering method is used to find cluster centers of spots as internal markers in the regions. Then we can use these markers to find and confirm more potential protein spots in each region. Most potential spots which are missed by the basic watershed segmentation method can be detected with our method.

Based on the description above, our detection method is mainly based on the process of existing watershed detection method but use clustering techniques, which is more powerful and accurate, to detect the center of the protein spots and improve the performance of spots detection technique as a result.

4. Clustering and PSO for Spots Modeling

4.1 Background

After the spots detection step, protein spots should be characterized and represented as a list of features for further analysis such as spots matching. Spots characterization can be divided into two categories: parametric and nonparametric [27]. Nonparametric method carries out various heuristic post-processing routines on the raw segmentation boundaries to delineate the spots. No explicit constraints on the shape of the boundaries or the appearance of the spot are imposed. Parametric methods use models to parameterize the protein spots. If the model is established, all the detected spots will be fitted to the certain model and the parameters of each spot to the model will be obtained. The parameters of the model will be used in further process such as spots matching. Next, the Gaussian model and the diffusion model will be introduced.

4.1.1 Gaussian Model

In 2D gel electrophoresis image analysis, the most frequently used spot model is the Gaussian model. It is based on the Gaussian function:

$$S(x,y) = B + I\exp(-\frac{x-x_0}{2\sigma_x^2})\exp(-\frac{y-y_0}{2\sigma_y^2}) \qquad (4.1)$$

where B is background intensity and I is spot intensity, x_0 and y_0 control the location of spot and σ_x and σ_y control the spread of the Gaussian independently in x and y directions. The formation of protein spots is a diffusion process. Under ideal conditions, the bivariate Gaussian model represents diffusion of an initial concentration focused at a single point in two independent directions.

4.1.2 Diffusion Model

In practical situation, the assumptions that must hold for a Gaussian model to accurately represent protein spots are often not correct. Bettens [28] [29] introduced a more detail theoretical mode of anisotropic perfect diffusion characteristics. The model is more accurate than the Gaussian model when the initial single point assumption is invalid.

It assumes that the medium in which the diffusion takes place is anisotropic: there are two main directions of diffusion, with different diffusion properties for each direction. The initial distribution should occupy a finite area but not only in a point. It assumes that the diffusing substance is initially distributed uniformly through a circle of radius a . The solution of the diffusion equation is:

$$C(x,y,t)=\frac{1}{2}C_0(erf(\frac{a+r}{2\sqrt{Dt}})+erf(\frac{a-r}{2\sqrt{Dt}}))$$

$$+\frac{C_0}{r}\sqrt{\frac{Dt}{\Pi}}(\exp(-\frac{(a+r)^2}{4Dt})-\exp(-\frac{(a-r)^2}{4Dt})) \qquad (4.2)$$

$$\text{with } r=\sqrt{D(\frac{(x-x_0)^2}{D_x}+\frac{(y-y_0)^2}{D_y})} \qquad (4.3)$$

with C_0 the initial concentration in the circle, x_0 and y_0 the place coordinates and Dx,y diffusion constants for the x and y directions.

To combine equations 4.2 and 4.3 as a new model in a fitting procedure, the symmetric parameters are eliminated and an extra parameter to compensate for the background is added. The new model is:

$$C(x,y) = B + \frac{1}{2}C_0(erf(\frac{a'+r'}{2}) + erf(\frac{a'-r'}{2})) + \frac{C_0}{r'}\sqrt{\frac{1}{\Pi}}(\exp(-\frac{(a'+r')^2}{4}) - \exp(-\frac{(a'-r')^2}{4})) \quad (4.4)$$

$$\text{with } r = \sqrt{D(\frac{(x-x_0)^2}{D'_x} + \frac{(y-y_0)^2}{D'_y})} \quad (4.5)$$

The **7** parameters to be fitted are:

$$B, C_0, a' = \sqrt{\frac{D}{t}}a, D'_x = D_x t, D'_y = D_y t, x_0, y_0$$

The best parameters for spots are determined by minimizing a χ^2 function.

4.1.3 Summary

Gaussian models produce an inadequate fitting to some protein spots that are most notably large volume saturated spots. The Gaussian formulation is not flexible enough to represent the true variation in spot appearance. Bettens's model is based on the physics of the spot formation. Protein spots are formed by a diffusion process. Gaussian model is adequate to represent it only when the initial concentration distribution occupied by the sample has a very small area. Diffusion model more adequately represents spots in the gel when Gaussian assumption is not met.

But both the Gaussian and diffusion models assume perfect diffusion in the 2D gel medium. In practice, the diffusion process is not perfect and spots may be formed with unpredictable and unusual shapes. The most difficult task is the modeling of the large saturated spots. With Gaussian or diffusion model, it is impossible to get satisfying results for such kind of spots.

To give better solution for this problem, we will propose a new method based on clustering and PSO for the modeling of large saturated spots.

4.2 Particle Swarm Optimization

Swarm intelligence attracts many researchers' attention nowadays and there are many evolutionary algorithms based on swarm intelligence. The Particle Swarm Optimization (PSO) is one of them. It is a relatively new evolutionary computation technique that has been empirically shown to perform well on many optimization problems. PSO is a population-based search algorithm and is initialized with a population of random solutions, called particles. Unlike the other evolutionary computation techniques, each particle in PSO is also associated with a velocity. Particles fly through the search space with velocities, which are dynamically adjusted according to their experience. Hence the particles have a tendency to fly towards better and better search area.

4.2.1 PSO Algorithm

The original PSO algorithm is developed based on simplified social model simulation. It is related to the bird flocking, fishing schooling and swarm theory. The PSO was first designed to simulate birds seeking food. The bird would find food through social cooperation with other birds around it (within its neighbourhood). It was then expanded to multidimensional search. The topological rather than Euclidean neighbourhood was utilized [30][31]. The original PSO algorithm is described as below:

$$\begin{aligned} V_i(d) = V_i(d) + c_1 \times rand1() \times (Pbest_i(d) - X_i(d)) \\ + c_2 \times rand2() \times (Gbest(d) - X_i(d)) \end{aligned} \tag{4.6}$$

$$X_i(d) = X_i(d) + V_i(d) \tag{4.7}$$

where c_1 and c_2 in the equation are the acceleration constants, which represent the weighting of stochastic acceleration terms that pull each particle toward *pbest* and *gbest* positions. And $rand1()$ and $rand2()$ are two random functions in the range $[0, 1]$; $X_i = (x_{i1}, x_{i2}, ..., x_{iD})$ represents the position of the

ith particle; $Pbest_i = (pbest,,pbest,, ,...,pbest_{iD})$ represents the best previous position (the position giving the best fitness value) of the ith particle; $Gbest = (gbest,,gbest,, ...,gbest,)$ represents the best previous position of the population; $V_i = (v_{i1}, v_{i2}, ..., v_{iD})$ represents the rate of the position change (velocity) for particle i.

Equation (4.6) and (4.7) is to describe the flying trajectory of a population of particles. Equation (4.6) describes how the velocity is dynamically updated and Equation (4.7) the position update of the “flying” particles. Equation (4.6) consists of three parts: The first part is the momentum part. The velocity cannot be changed abruptly. It is changed from current velocity. The second part is the “cognitive” part, which represents private thinking of itself, learning from its own flying experience. The third part is the “social” part, which represents the collaboration among particles – learning from group flying experience [32].

In equation (4.6), if the sum of the three parts on the right side exceeds a constant value specified by user, then the velocity on that dimension is assigned to be $\pm V\max$, that is, particles’ velocity on each dimension is clamped to a maximum velocity $V\max$, which is an important parameter, and originally is the only parameter required to be adjusted by users. Big Vmax makes particles have the potential to fly far past good solution areas while a small Vmax makes particles have the potential to be trapped into local optima, therefore unable to fly into better solution areas. Usually a fixed constant value is used as the Vmax ,but a well designed dynamically changing V max might improve the PSO performance [33].

4.2.2 Application

The original procedure for implementing PSO is as follows:

```
PSO:
Initialize the swarm:
Initialize positions and associated velocities of all particles in the population randomly in the D-dimension space. Evaluate the fitness values of all particles. Set the current position as pbest and the current particle with the best fitness value in the whole population as gbest.
For k=l to max iteration
For i=l to ps, ps is the population size
For d=l to D
 $V_i(d) = V_i(d) + c_1 \times rand1() \times (Pbest_i(d) - X_i(d)) + c_2 \times rand2() \times (Gbest(d) - X_i(d))$
 Limit the velocity $V_i(d) = \min(V_{\max}(d), \max(-V_{\max}(d), V_i(d)))$
   $X_i(d) = X_i(d) + V_i(d)$
 End
 Calculate the fitness value for $X_i$
 Update $Pbest_i$, $Gbest$ if needed.
```

```
End
Stop if a stop criterion is satisfied
End
$c_1 = c_2 = 2$
```

Like the other evolutionary algorithms, PSO algorithms is a population based search algorithm with random initialization, and there is an interaction among population members. In PSO, each particle flies through the solution space has the ability to remember its previous best position and survives from generation to generation [34]. Further more, compared with the other algorithm, PSO is faster in initial convergence while slower in the refining [35][36]. Also, the PSO algorithm is simple in concept, easy to implement and computational efficient.

In the next part of this chapter, PSO will be used with subtractive clustering method for the shape construction of saturated spots.

4.3 Saturated Spots Modeling

4.3.1 Introduction

In the 2D gel electrophoresis image, sometimes the large spot region are badly saturated, therefore, the gray level information is not helpful for the modeling of saturated part. Basically, this kind of modeling is a covering issue for the shape of the saturated part of the protein spots. To model the shape for further use, we should give the enough precision description of the shape. That means we use a modeling method to cover the saturated part as much as possible, avoiding the pixels of background.

For this issue, a statistical modeling method is proposed by Rogers et al[37] which uses the point distribution models (PDM) to model the protein spots shape. A PDM represents the statistics of the observed variation in a training set of the spots shapes. The PDM is constructed in three steps: 1.parameterize the spots shapes by placing landmark points on object boundaries in training set of images, 2.align the landmarks, 3.analyze the remaining variation amongst the aligned training data. But with such methods, the speed of complex computation is very slow and data amount of the spots parameters for storage is very large. In addition, this is only a description of the shape. Actually, a large saturated protein spot is constructed by the merging of several small protein spots at the processing of 2D gel electrophoresis. Therefore, the results of statistical method will lose all the information about those merging spots which may be used by further analysis.

Some others [38] use axis-parallel ellipse to cover the saturated spots region as much as possible. Each of the axis-parallel ellipses is possible to correspond to a merged protein spot from the large saturated spot, so such methods give a good affiliation of the real spots and the models. But all such methods use nearly complete random searching to generate the best covering ellipses for the saturated shapes. The complete random method will use much more time and the covering results are usually not satisfying enough.

Basically, a large saturated protein spot in the 2D gel electrophoresis is constructed with a main spot in the center and several spots around it. We will use subtractive clustering to estimate the center of all the protein spots and perform PSO to make a fast and precise covering search with ellipses for the saturated shapes. The details will be given in the next part.

4.3.2 Proposed Modeling Method

In this method, to model a large saturated spot that is merged by several small spots, subtractive clustering will be used to find the probable centers of all the small spots and PSO will be used to detect the ellipses fitting to the shape of all the small spots. In the covering search process, PSO will minimize the covering error of the ellipses on the saturated region.

As we described before, a large saturated spot is constructed with a main spot in the center and several spots around it. The first task is to estimate the main spot. The clustering method will be used to estimate the center of the main spot, which is just the first cluster center. Then we perform particle swarm optimization to search the best ellipse fitting of the main spot. We choose PSO based on its simplicity of implementation and high speed of computation, which are very suitable for fitting of the large amount saturated protein spots.

The first fitting only cares about the shape of the main spot. After that, from the difference between original saturated spot region and covering of the first fitting ellipse, the left saturated regions can be obtained. For these regions, subtractive clustering will be performed again to find all the potentia1 spots. In this step, if the mountain value of a cluster center is evidently bigger than others, the center of a protein spot is located around the cluster center. To make sure which cluster center can be a potential spot center, Ostu's thresholding method is used to find a suitable threshold for these cluster centers.

All the cluster centers picked are considered as the estimated center of the protein spots. Based on these centers, we use PSO again to search the fitting ellipses for all the left spots. In the fitting of each left spot, to make the processing faster, the ellipse should be based on its original center and do not contain the centers of other spots. That's because two spots do not intersect to each other too much. Then we make PSO search processing to find the fitting ellipse shape for each potential spot. Each of the fitting ellipses corresponds to a potential protein spot. With the combined covering of all the fitting ellipses searched, we can obtain the model for the saturated protein spots.

4.3.3 Simulation and Results

We will take the saturated spots from 2D gel electrophoresis image (Figure 4-1) for example as follows:

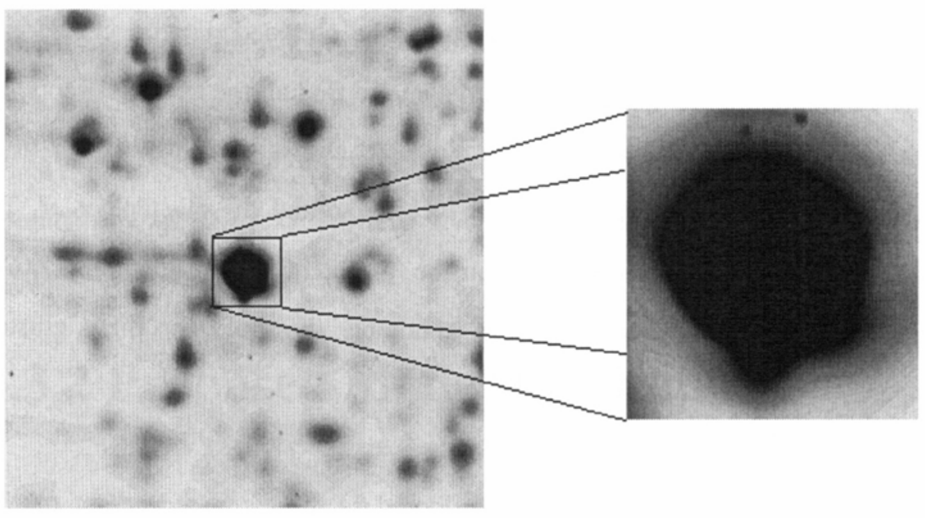


Figure 4-1 2DGE image and a saturated spot

With thresholding method, we can get the completely saturated region (Figure 4-2) from the original spot.

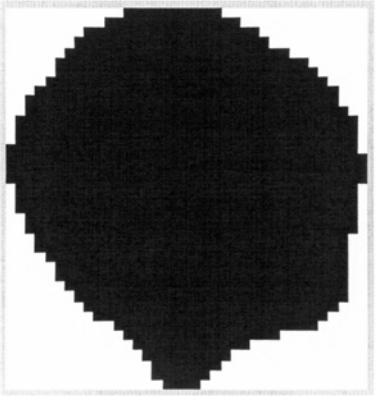


Figure 4-2 Saturated region of the spot

We use subtractive clustering method to find the center of the main spot, which is just the first cluster center (shown as a weight point on the spot in Figure 4-3).

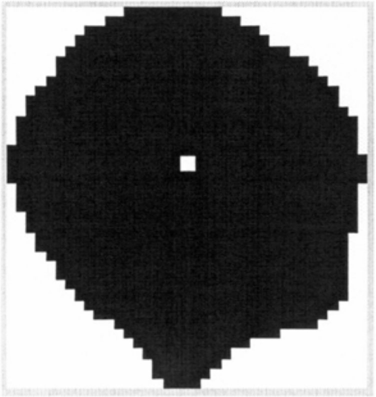


Figure 4-3 Saturated region with first cluster center

Then we perform Particle Swarm Optimization to search the best fitting ellipse for the main spot shown in Figure 4-4.

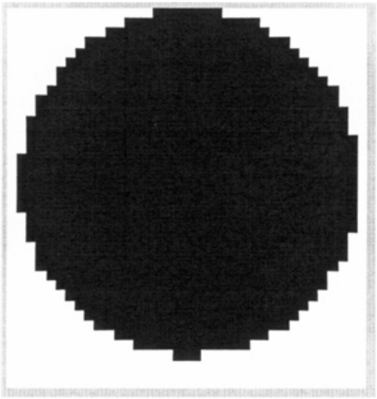


Figure 4-4 The fitting ellipse for the main spot

From the left saturated regions, we can get other cluster centers associated with the potential protein spots. They are shown as the white points on the protein spot in Figure 4-5.

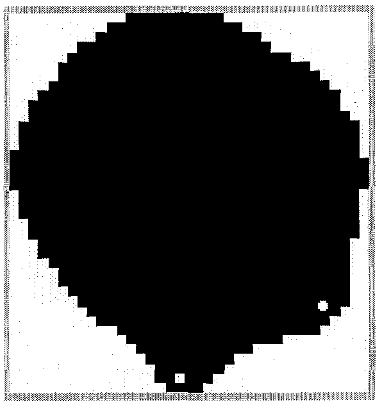


Figure 4-5 Other cluster centers associated with the potential protein

spots

Based on these centers, we use PSO again to search fitting ellipses for all the spots left shown in Figure 4-6 and 4-7.

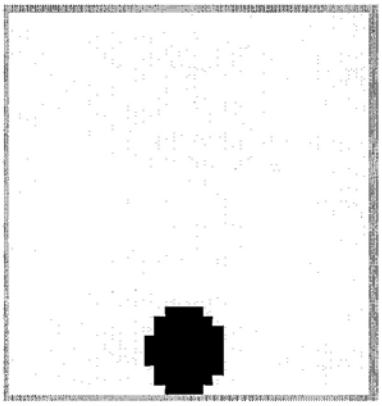


Figure 4-6 The fitting ellipse for another spot

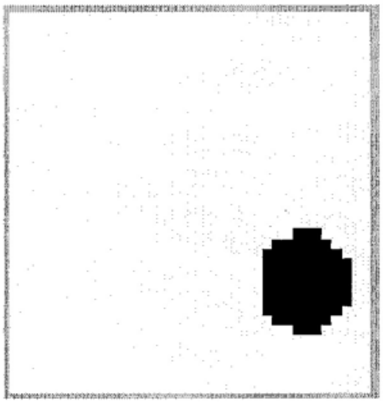


Figure 4-7 The fitting ellipse for another spot

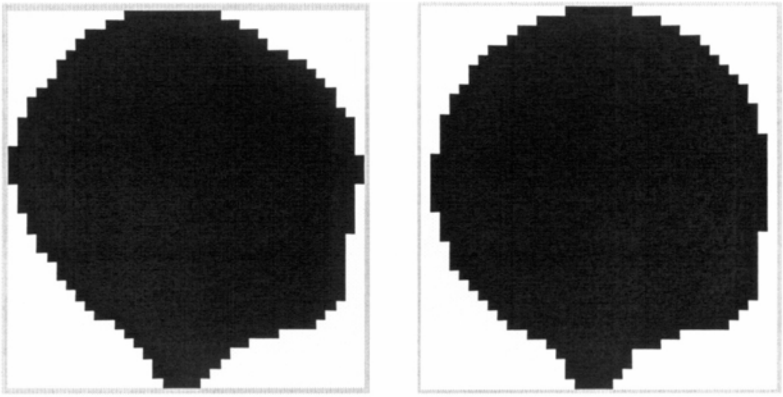


Figure 4-8 The left is original saturated region. The right is modeling result.

With the combination of covering of all the fitting ellipses , we get the model for the saturated protein spot. The modeling result is shown in Figure 4-8.

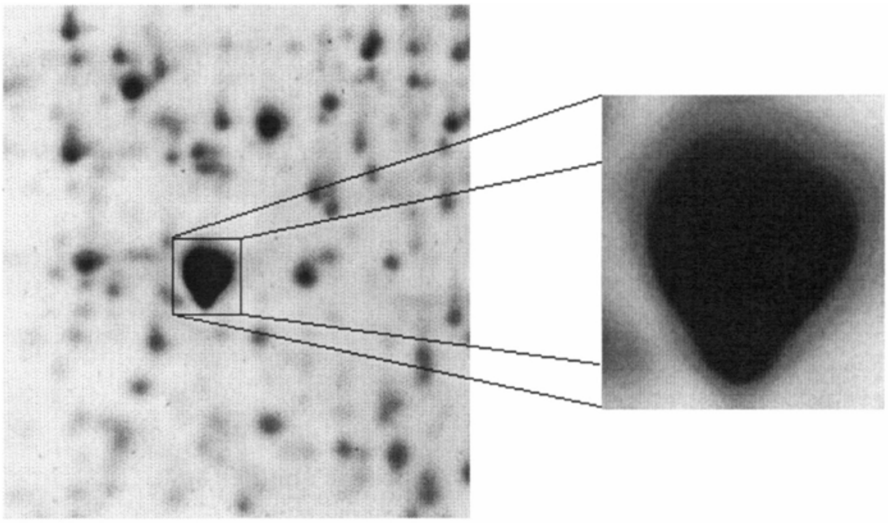


Figure 4-9 Saturated region of the spot

With thresholding method, we can get the completely saturated region (Figure 4-10) from the original spot.

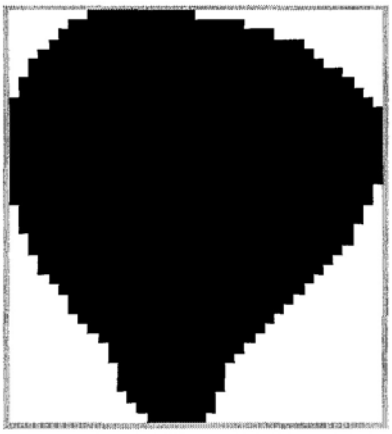


Figure 4-10 Saturated region of the spot

We use subtractive clustering method to find the center of the main spot, which is just the first cluster center (shown as a weight point on the spot in Figure 4-11).

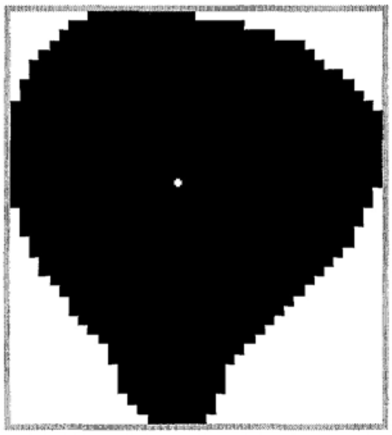


Figure 4-11 Saturated region with first cluster center

Then we perform Particle Swarm Optimization to search the best fitting ellipse for the main spot shown in Figure 4-12.

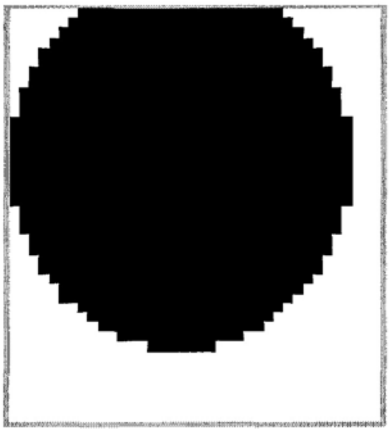


Figure 4-12 The fitting ellipse for the main spot

From the left saturated regions, we can get other cluster centers associated with the potential protein spots. They are shown as the white points on the protein spot in Figure 4-13.

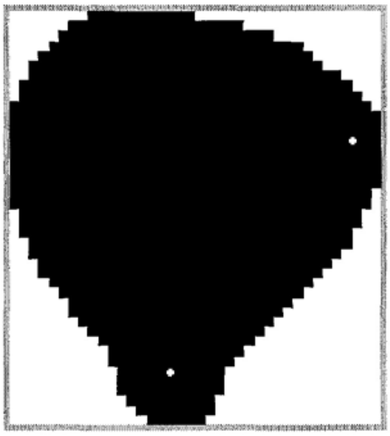


Figure 4-13 Other cluster centers associated with the potential protein spots

Based on these centers, we use PSO again to search fitting ellipses for all the spots left shown in Figure 4-14 and 4-15.

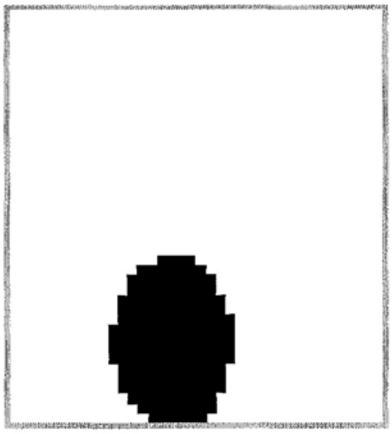


Figure 4-14 The fitting ellipse for another spot

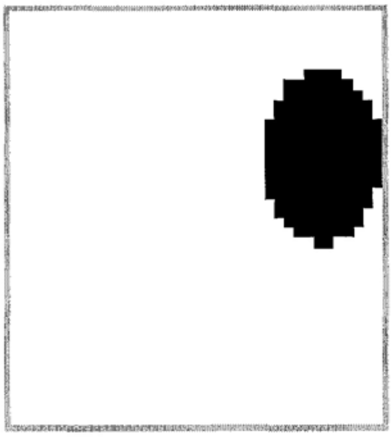


Figure 4-15 The fitting ellipse for another spot

With the combination of covering of all the fitting ellipses , we get the model for the saturated protein spot. The modeling result is shown in Figure 4-16.

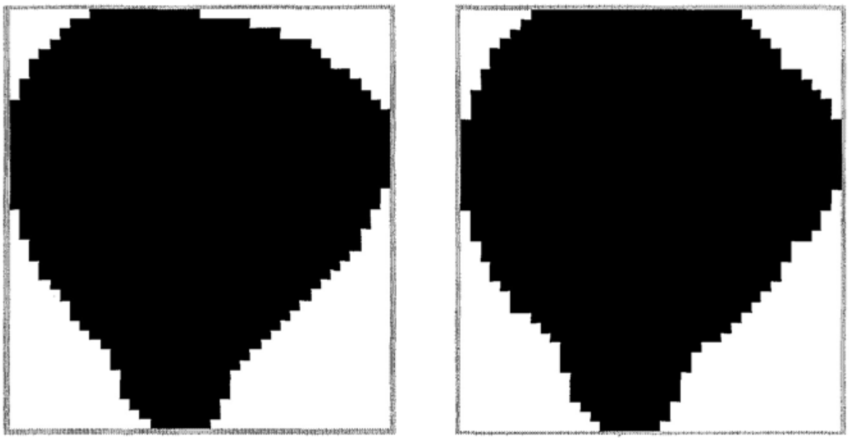


Figure 4-16 The left is original saturated region. The right is modeling result.

With view of two examples above for spots fitting of the saturated regions of protein spots, it is clear that the shapes of the saturated regions are well fit. The accuracy of the fitting to these two shapes is more than 96%. We have performed our method on the modeling of many saturated protein spots. For the complete saturated region of such spots, the accuracy of the most modeling is always more than 96%, which has shown that the performance of the fitting method just with only several overlapping ellipses shapes is satisfying. With the existing single ellipse modeling for those regions, even for quite simple shapes, the accuracy is not more than 89%. In little more complex spot regions, the accuracy of the single ellipse model will be much less than 80%.

Actually, for the spots covering with our method, the modeling accuracy can approach 100%because obviously we can use enough amounts of ellipses which are small enough to cover any leaving areas inside the spots. The tradeoff of the accuracy and the amount of cover ellipses is depended on the purpose of further applications. In most cases, three ellipses can get an accuracy of more than 96% with only a few overlapping ellipses shape.

4.4 Conclusion

The traditional spots modeling methods can only give good handling for the normal spots with Gaussian profile but not suitable for the saturated spots. Bettens proposed a diffusion model which assumes that the initial distribution should occupies a finite region within a circle but not in a point. Diffusing substance is initially distributed uniformly through a circle of radius$a$. But in practice, the diffusion process is not perfect and protein spots may be formed with unpredictable and unusual shapes, so the assumption of diffusing substance is not practicable.

To model the saturated region of the spots, the statistical model can make a satisfying description for the saturated region, but it ignores all the information about the construction of the saturated spot which is constructed from several protein spots. Some other methods use axis-parallel ellipse to cover the saturated spots region, but they use nearly complete random searching to generate the best covering ellipses for the saturated regions. The complete random method will use much more time and the covering results are usually not satisfying enough.

We have proposed a modeling method for the saturated region of the protein spots. We also use axis-parallel ellipses to make covering for the saturated region. The technologies of subtractive clustering and particle swarm optimization are utilized. The clustering method can be performed to estimate the position of the potential spots. With these estimations, the search of the covering ellipses for each merging spot will not be complete random. Then we perform PSO to find the best covering ellipses for all the spots. The simulation results have shown the satisfying performance for our covering model.

5. Conclusion and Recommendations

5.1 Conclusion

In this thesis, we have illustrated the problem of the traditional watershed algorithm when it is used in the protein spots detection processing. With the observation on the 3D profile and the section image of some protein spots, we find that some spots may be missed by basic watershed detection. Based on the properties of such spots, we proposed a clustering based watershed segmentation method for spots detection. If the pixels in the 2D gel electrophoresis image are regarded as cluster data, we utilize the subtractive clustering to detect the cluster centers which can be used as the internal markers in watershed segmentation. With the new markers, more potential protein spots can be detected by our methods. In the table of experiment results, when our method is performed in different 2D gel electrophoresis images, it has always got higher performance than the basic watershed segmentation in the same conditions.

We have proposed a modeling method for the saturated regions of the protein spots. We use the optimization and clustering method to search the axis-parallel ellipses and make covering on the saturated region. Particle swarm optimization and subtractive clustering are used to make the model. With clustering method, we can obtain good estimation for the positions of the potential merging spots which contract a large saturated spot. Then PSO is used to minimize the covering error and search the best covering ellipses for those large saturated protein spots. The combination of all the detected ellipses is just the model for the saturated region of protein spot. Simulations have shown the satisfying results of the covering model.

In comparison with existing works, our clustering based spots detection method and spots modeling methods are important because they make the detection and modeling of protein spots more accurate and faster, will help to save the time of researchers who do the job of protein spots detection and modeling and will improve the performance of the further processes such as spots matching and comparison.

5.2 Recommendations for Further Research

Based on the techniques and algorithms developed in this thesis, we will give the recommendations for further research areas.

In the spots detection of our research, we have given a good solution after the region segmentation from watershed algorithm. With observation of the 2D gel electrophoresis images, we can find that many streaks that cross the adjacent protein spots. From the view of gray level image, such streaks show nearly the same gray value with some protein spots. With basic image segmentation methods, it is very difficult to distinguish them from spots. Some new techniques should be explored to analysis the streaks and find more effective features which will be helpful to identify and then remove these streaks.

Our modeling method can give satisfying description for the saturated region of protein spots. Actually, for the modeling of normal spots which are not saturated, Gaussian model **is** frequently used, but most of the spots do not strictly follow this model, and the modeling error sometimes is very large. With the purpose to minimize the modeling error, new more complex spots model may be explored based on better gathering of features from the real protein spots.

Author's Publications

[1] Diao Xiaoning, Mao Kezhi, "Clustering Based Watershed Segmentation for Two-Dimensional Gel Electrophoresis Image", International Bioengineering Conference, 2004.

Bibliography

[1] Michael R.Barnes, Lan C.Gray, “Bioinformatics for Geneticists”, 2003.

[2] Timothy Palzkill, “Proteomics”, 2001.

[3] David W.Speicher, “Proteome analysis interpreting the genome”, 2004.

[4] Godley, A.A, and Packer, N.H., “Proteome Research: New frontiers in Functional Genomics”, Springer, Berlin, 65-91, 1997.

[5] Reiner Westermeier, “Electrophoresis in Practice”, 2000.

[6] Andrew J.Link, “2-D Proteome analysis protocols”, 1999.

[7] A.W.Dowsey, M.J.Dunn, G.Z.Yang, “The role of bioinformatics in two dimensional gel electrophoresis”,Proteomics, 3, 1567-1596,2003.

[8] R.D.Appe1, J.R.Vargas, P.M.Palagi, D.Walther, et al., “Melanie I1 – a third generation software package for analysis of two dimensional electrophoresis images: II. Algorithms”, Electrophoresis, 18, 2735–2748, 1997.

[9] Yecheng.Wu, P.F.Lemkin, K.Upton, “A fast spot segmentation algorithm for two dimensional gel electrophoresis analysis”, Electrophoresis, 14, 1351–1356, 1993.

[10] J.I.Garrels, “Two-dimensional Gel Electrophoresis and Computer Analysis of Proteins Synthesized by Clonal Cell Lines””, Biological Chemistry, Vo1.254, No.16, pp.7961-7977, 1979.

[11] K.Takahashi, M.Nakazawa, Y.Watanabe, A.Konagaya,

"Fully-Automated Spot Recognition and Matching Algorithm for 2-D Gel Electrophoretogram of Genomic DNA", Genome Inform., 9, 161-172, 1998.

[12] J.Prehm, P.Jungblut, J.Klose, "Analysis of Two-dimensional Electrophoretic Protein Using a Video Camera and a Computer Ⅱ. Adaptation of Automatic Spot Detection to Visual Evaluation", Electrophoresis, 8,562-572, 1987.

[13] G.W.Horgen, C.A.Glasbey, "Use of Digital Image Analysis in Electrophoresis", Electrophoresis, 16, 1995.

[14] J.B.T.M.Roerdink, A.Meijster, "The watershed transform: Definitions, Algorithms and Parallelization Strategies", Fundamental Informatics, 41, 187-228,2000.

[15] L.Vincent, P.Soille, "Watersheds in Digital Spaces: An Efficient Algorithm Based on Immersion Simulations", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.13, No.6, pp.583-598, 1991.

[16] R.R.Yager, D.P.Filev, "Approximate Clustering Via the Mountain Method", IEEE Transactions on Systems, Man, and Cybernetics, vol.24, N0.8, 1994.

[17] N.R.Pal, "Mountain and Subtractive Clustering Method: Improvements and Generalizations", Debrup Chakraborty, Jan 24, 2001.

[18] S.L.Chiu, "Extracting Fuzzy Rules for Pattern Classification by Cluster Estimation", Proc. Sixth Int. Fuz. Systs. Assoc, World Congress (IFSA'95), Sao Paulo, Brazil, ppl-4, July 1995.

[19] S.L.Chiu, "An Efficient Method for Extraction Fuzzy Classification Rules from High Dimensional Data", Advanced Computational Intelligence, vol. 1, No. 1, 1997.

[20] K.P.Pleissner, Helmut Oswald, Susan Wegner, "Image analysis of two-dimensional gels", Proteomics, 2001.

[21] Y.C.Wu, P.F.Lemkin, K.Upton, "A Fast Spot Segmentation Algorithm for Two-dimensional Gel Electrophoresis analysis", Electrophoresis, 14, 1351-1356, 1993.

[22] L.Z.Xia, "Digital image processing", University of South East, 1999.

[23] A.W.Dowsey, M.J.Dunn, G.Z.Yang, "The role of bioinformatics in two dimensional gel electrophoresis",Proteomics, 3, 1567-1596,2003.

[24] R.D.Appe1,., J.R.Vargas, P.M.Palagi, D.Walther, et al., "Melanie 11– a third generation software package for analysis of two dimensional electrophoresis images", Electrophoresis, 18,2724-2748, 1997.

[25] R.C.Gonzalez, R.E.woods, "Digital image processing – Second Edition", 2002.

[26] N.Ostu, "A Threshold Selection Method from Gray-Level Histograms", IEEE Transactions on Systems, Man, and Cybernetics, vo1.9, no. 1, pp.62-66, 1979.

[27] P.L.Lemkin, J.E.Myrick, K.M.Upton, "Splitting Merged Spots in Two-dimensional Polyacrylamide Gel Electrophoresis Images", Applied and Theoretical Electrophoresis, 3, 163-172, 1993.

[28] E.Bettens, "Peak Characterization Using Parameter Estimation Methods Ph.D. thesis", University of Antwep, 1999.

[29] E.Bettens, PScheunders, J.Sijbers, D.V.Duck, L.Moens, "Automatic

Segmentation and Modeling of Two-dimensional Electrophoresis Gels", Electrophoresis, 18, 792-798, 1999.

[30] R. C. Eberhart and J. Kennedy, "A new optimizer using particle swarm theory," Proc. of the Sixth International Symposium on Micromachine and Human Science, Nagoya, Japan, pp. 39-43, 1995.

[31] R.C.Eberhart, P.K.Simpson and R.W.Dobbins Computational Intelligence PC tools, Boston: MA: Academic Press Professional,

[32] **Y.** Shi and R. C.Eberhart, "A modified particle swarm optimizer," Proc. of the IEEE Congress on Evolutionary Computation (CEC 1998), Piscataway, NJ, pp. 69-73, 1998.

[33] H. Y.Fan **and** Y. **Shi,** "Study on Vmax of particle swarm optimization," Proc. of the Workshop on Particle Swarm Optimization 2001, Indianapolis, IN, 2001.

[34] **Y. Shi** and R. C. Eberhart, "Particle swarm optimization with fuzzy adaptive inerita weight," Proc. of the Workshop on Particle Swarm Optimization, Indianapolis, IN, 2001.

[35] P.J.Angeline, "Using selection to improve particle swarm optimization," Proc. of the IEEE Congress on Evolutionary Computation (CEC 1998), Anchorage, Alaska, USA, 1998.

[36] P.J.Angeline, "Evolutionary optimization versus particle swarm optimization: philosophy and performance differences," Evolutionary Programming VII: Proceedings of the Seventh Annual Conference on Evolutionary Programming, 1998.

[37] M.Rogers, J.Graham, R.P.Tonge, "Statistical models of shape for the analysis spots in two-dimensional electrophoresis gel images."

Proteomics, 3,887496,2003.

[38] A.Efrat, F.Hoffmann, K.Kriege1, C.Schultz, C.Wenk, "Geometric algorithm for the analysis of 2D-electrophoresis gels", the Fifth Annual International Conference on Computational Molecular Biology (RECOMB), 114-123, Montreal, Canada, 2001.

[39] Diao Xiaoning, Mao Kezhi, "Clustering Based Watershed Segmentation for Two-Dimensional Gel Electrophoresis Image", International Bioengineering Conference, 2004.