ANALYSIS OF TWO-DIMENSIONAL ELECTROPHORESIS GEL IMAGES

Lars Pedersen

Informatics and Mathematical Modelling
Ph.D. Thesis No. 96
Kgs. Lyngby 2002

IMM

Printed by IMM/Technical University of Denmark

Preface

This thesis has been prepared at the Image Analysis and Computer Graphics section at the Informatics and Mathematical Modelling (IMM), Technical University of Denmark in partial fulfilment of the requirements for the degree of Ph.D. in engineering.

The general framework for this thesis is pattern analysis, digital image processing and computer vision with application in the field of proteomics. The subject is Analysis of Two-dimensional Electrophoresis Gel Images.

This work was carried out in close collaboration with *Centre for Proteome Analysis in Life Sciences* (CPA), University of Southern Denmark, Odense.

Part of this thesis is confidential and therefore Chapter 5 is omitted from this edition.

Kgs. Lyngby, February 2002

Lars Pedersen

Acknowledgements

I would like to thank the many people who have contributed their time to helping me with this thesis, for fruitful discussions and critical review. First, my thanks to my supervisors Associate Professor Bjarne Ersbøll, Associate Professor Stephen J. Fey, and Professor Knut Conradsen for their invaluable suggestions and constructive criticism.

I am grateful to *Centre for Proteome Analysis in Life Science* (CPA) for financial support and supplying of data material. In particular to Associate Professor Stephen J. Fey and Associate Professor Peter Mose Larsen, for making me realise the importance of proteomics and for teaching me some of the peculiarities of cell biology. Also at CPA, I wish to thank Arkadiusz Nawrocki and Adelina Rogowska for providing and preparing most of the data material used in this work.

Informatics and Mathematical Modelling at the Technical University of Denmark has provided me with an environment in which to carry out this work and for that I am thankful. At IMM I wish to thank previous and present members of the Section for Image Analysis and Computer Graphics and in particular my office-mates Klaus Baggesen Hilger, Mikkel B. Stegmann, and Rune Fisker who have provided a pleasant atmosphere throughout the years while also being inspiring in many ways and Associate Professor Jens Michael Carstensen for many inspiring discussions. Many thanks to our secretary staff Helle Welling, Mette Larsen, and Eina Boeck.

I am most thankful to Professor James Duncan for giving me the chance to work with his group; the Image Processing and Analysis Group at Department of Diagnostic Radiology, Yale University School of Medicine during my external research stay. Here, I in particular wish to thank Dr. Haili Chui and Associate

Professor Anand Rangarajan who have been of great inspiration and Reshma Munbodh and Larry Win for providing an always joyful environment. Thanks to Carolyn Meloling for secretary help.

I owe a great debt to Peter Chapman for his unique hospitality and to Camilla Hampton, Susan D. Greenberg, Robert Rocke and Matt Feiner for taking care of me and for introducing me to many aspects of New Haven and the American East coast culture.

Finally I wish to thank Stephen J. Fey, Bjarne Ersbøll, Mikkel B. Stegmann, Klaus Baggesen Hilger, Rasmus R. Paulsen, and Michael Grunkin for careful and patient review of the manuscript.

Lars Pedersen

Abstract

This thesis describes and proposes solutions to some of the currently most important problems in pattern recognition and image analysis of two-dimensional gel electrophoresis (2DGE) images. 2DGE is the leading technique to separate individual proteins in biological samples with many biological and pharmaceutical applications, e.g., drug development. The technique results in an image, where the proteins appear as dark spots on a bright background. However, the analysis of these images is very time consuming and requires a large amount of manual work so there is a great need for fast, objective, and robust methods based on image analysis techniques in order to significantly accelerate this key technology.

The methods described and developed fall into three categories: image segmentation, point pattern matching, and a unified approach simultaneously segmentation the image and matching the spots.

The main challenges in the segmentation of 2DGE images are to separate overlapping protein spots correctly and to find the abundance of weak protein spots. Issues in the segmentation are demonstrated using morphology based methods, scale space blob detection and parametric spot modelling. A mixture model for parametric modelling of several spots that may also be overlapping is proposed.

To enable comparison of protein patterns between different samples, it is necessary to match the patterns so that homologous spots are identified. Protein spot patterns, represented by the spot centre coordinates can be regarded as two-dimensional points sets and methods for point pattern matching can be applied. This thesis presents a range of state-of-the-art methods for this purpose and also suggests a *regionalised* scheme. The general point pattern matching methods focussed on are the *Robust Point Matching* methods and among the methods

developed in the literature specifically for matching protein spot patterns, the focus is on a method based on neighbourhood relations. These methods are applied to a range of 2DGE protein spot data in a comparative study.

The point pattern matching requires segmentation of the gel images and since the correct image segmentation can be difficult, a new alternative approach, exploiting prior knowledge from a reference gel about the protein locations to segment an incoming gel image, is proposed.

Resumé

Denne afhandling beskriver og foreslår løsninger til nogle af de vigtigste eksisterende problemer inden for mønstergenkendelse og billedanalyse af todimensional elektroforese gel (2DGE) billeder. 2DGE er den førende teknik til at separere de enkelte proteiner i biologiske prøver fra hinanden og teknikken har adskillige anvendelser inden for bioteknologi og farmakologi, f.eks. ved udvikling af nye lægemidler. Teknikken resulterer i et billede, hvor proteinerne fremstår som mørke pletter på en lys baggrund. Imidlertid er analysen af disse billeder særdeles tidskrævende og kræver en del manuelt arbejde, så der er et udtalt behov for hurtige, objektive og robuste metoder, baseret på billedanalyseteknikker med det formål at give 2DGE teknologien et væsentligt skub fremad.

Metoderne beskrevet og udviklet her kan inddeles i tre kategorier: billedsegmentering (dvs. adskillelse af billedet i protein-pletter og baggrund), punkt mønster parring og en forenet fremgangsmåde, der segmenterer billedet og samtidig parrer sammenhørende protein-pletter.

De vigtigste udfordringer i segmentering af 2DGE-billeder er at adskille tætliggende, overlappende protein-pletter og at detektere den store mængde af små, svage protein-pletter. Problemstillinger i segmenteringen er belyst vha. metoder i den matematiske morfologi, skalarumsbaseret klatdetektion (eng. *blob detection*), og parametrisk modellering af protein-pletter. Derudover foreslås en ny model baseret på vægtet superposition af parametriske plet-modeller. Denne *mixture* model kan modellere flere, evt. overlappende, pletter.

For at kunne sammenligne protein mønstre fra forskellige biologiske prøver er det nødvendigt at parre mønstrene så homologe protein-pletter kan identificeres. Repræsenteres protein mønstrene vha. pletternes center-positioner, kan disse betragtes som punktmængder i to dimensioner og så kan metoder til parring

(eng: *matching*) af punktmængder anvendes. Denne afhandling præsenterer en række førende, generelle metoder til dette formål og foreslår også en regionaliseret fremgangsmåde. Blandt de generelle metoder til punkt-parring fokuseres på familien af metoder kaldet *Robust Point Matching* og blandt metoderne i litteraturen, specielt udviklet til parring af protein-plet-mønstre, ligger fokus på en metode baseret på naboskabsrelationer. Metoderne er i et sammenlignende studie anvendt på en række 2DGE protein-plet data.

Parring af punktmængder forudsætter en segmentering af gelbilledet og en sådan segmentering kan være vanskelig at udføre korrekt. Derfor er der her udviklet en alternativ fremgangsmåde, der i segmenteringen af et nyt gelbillede drager nytte af forhåndsviden om proteinernes position fra en reference gel.

Contents

List of Tables

List of Figures

List of Algorithms

Chapter 1

Introduction

The field of proteomics or proteome analysis has become an increasingly important part of the life sciences, especially after the completion of sequencing the human genome. Proteome analysis is the science of separation, identification, and quantitation of proteins from biological samples with the purpose of revealing the function of living cells. Applications range from prognosis of virtually all types of cancer over drug development to monitoring of environmental pollution.

Currently, the leading technique for protein separation is two-dimensional gel electrophoresis (2DGE), resulting in grey level images showing the separated proteins as dark spots on a bright background (see Fig. 1.1). Such an image can represent thousands of proteins.

In order to identify the protein *diversity* and to quantitate the protein *amount* in a biological sample, pattern analysis and recognition can be of help. It also seems natural to apply pattern analysis in the task of comparing this information with similar information from other samples or a database. A small region of 150×150 pixels is shown in detail.

Pattern analysis methods *are* currently applied in order to automate and ease the task of analysing gel images and comparing images from different biological samples, but with the current methods this part of the process requires large

amounts of human-assisted work and it can be identified as the *major bottleneck* in the total process from biological sample to protein identification and quantitation.

The most important breakthrough in proteomics have been:

- introduction of immobilised pH gradients (1988) and
- introduction of mass spectrometry in the 1990's.

What would lead to an equal breakthrough would be improved pattern recognition methods for analysis of the gel images, reducing the large amount of resources spend on human-assisted analysis of the gels. In other words, there is a great need for effective, reliable, and objective methods to analyse the enormous amounts of data coming from the proteomics research.

The pattern analysis of the 2DGE data is traditionally divided into two parts, namely the *segmentation* of the 2DGE images into what is protein spots and what is background, and the process of *matching* protein spot patterns across two or more gels. Correct segmentation results in quantitation of the spots that reflects accurately the amount of protein present. The matching enables to detect changes in protein expressions across samples, or even to identify new proteins.

This thesis provides, with the main focus on protein spot matching, an overview of the pattern analysis related issues and open problems in the field of analysis of 2DGE images. State-of-the-art methods for point matching are presented, extended and tested on 2DGE data as well as a new method, combining segmentation and matching, is proposed.

These contributions will most likely open the above-mentioned bottleneck and enable to reduce the resources currently spend in the analysis of 2DGE images, hence take proteomics a step further.

1.1 Thesis Overview

The thesis is structured in the following manner:

- § 2 provides the motivation for this work. An introduction to the biological background is given and the interesting issues in the two pattern analysis related areas, 2DGE image segmentation and spot pattern matching, are

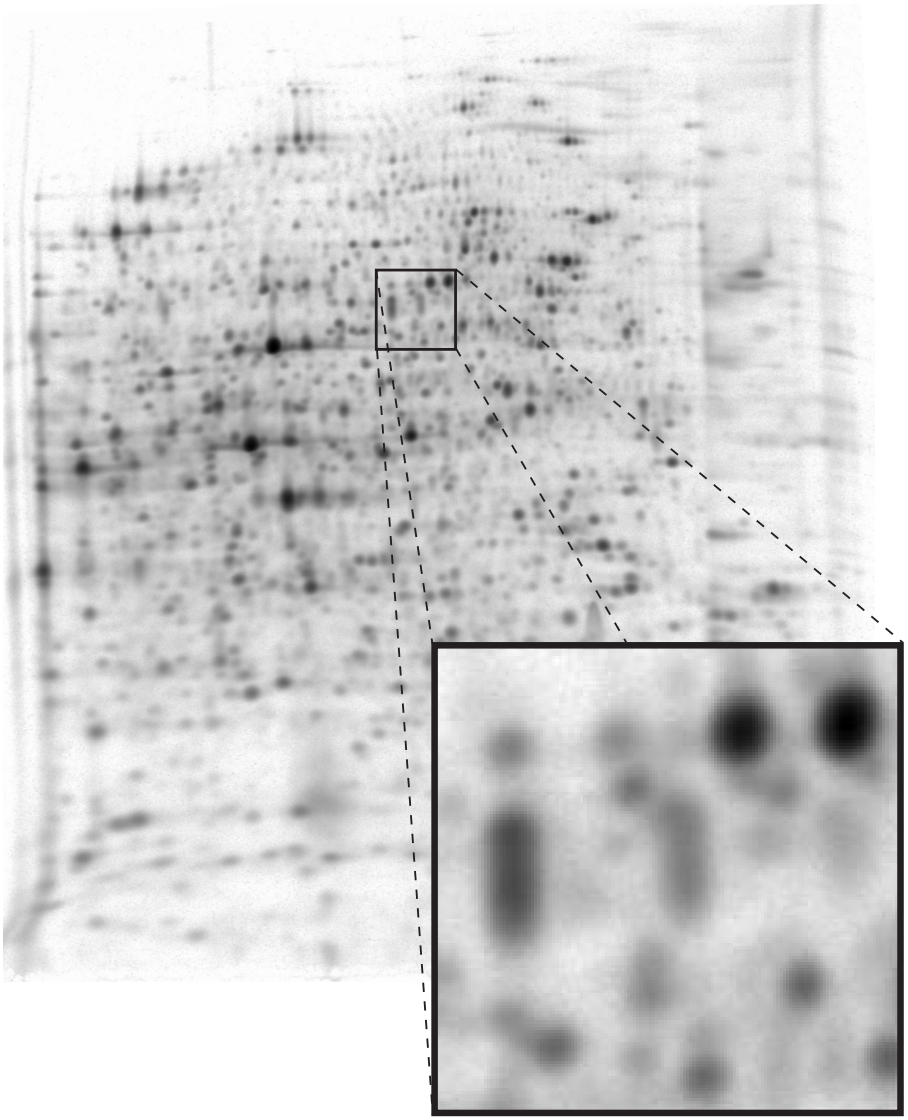


Figure 1.1: Two-dimensional electrophoresis gel image of baker's yeast *Saccharomyces cerevisiae* strain Fy1679-28C EC [pRS315]. Detail of 150×150 pixels region.

described. The nature of the spot patterns are discussed as well as a set of requirements for spot matching methods is defined.

- § 3 deals with issues such as low signal-to-noise ratio and spot overlap in the non-trivial task of segmenting the protein spots from the background. The subjects discussed are classical mathematical morphology, scale space based blob detection, and parametric spot modelling, all for the purpose of 2DGE image segmentation.

- § 4 presents a variety of methods for general point pattern matching succeeded by a range of methods designed for the matching of proteins spot patterns. A number of experiments on real 2DGE data is used for method comparison.

- § 5 proposes an alternative to the classical two-step procedure (segmentation succeeded by matching), namely a unified approach that simultaneously estimates the spot correspondence and segments the gel image.

1.1.1 Notation

When matching data from two 2DGE images, the task is usually to match an incoming new gel, Ω_q to a well-known gel, Ω_p. Hence the names *reference* gel (Ω_p) and *match* gel (Ω_q). The grey level intensity image of the reference gel is referred to as I_p and the set of points representing the protein spot centres is denoted $\mathcal{P}$. Similarly for the match gel, the intensity image is I_q and the point set is $\mathcal{Q}$. The correspondence between homologous points in the two point sets can be thought of as a field of disparity, describing the deformation from one set to the other. This disparity field is denoted δ.

A few abbreviations used most often in this thesis are:

2DGE two-dimensional gel electrophoresis.

RPM robust point matching.

TPS thin-plate spline.

1.2 Thesis Contributions

The main contributions of this thesis can be summarised in order of appearance as follows:

- extension of Sinkhorn's matrix normalisation method § 4.2.4,
- extension of RPM-TPS to include attribute information in the energy function § 4.2.4,
- regionalised point matching § 4.3.4,
- application and comparison of state-of-the-art point matching methods to real 2DGE data § 4.5 and
- EGM – elastic graph matching § 5.

The Sinkhorn's matrix normalisation method used in the Robust Point Matching (RPM) methods has been extended so that outlier rows and columns in the match matrix are not normalised and the method has also been extended to robustly handle non-square matrices.

In point matching, the points' spatial locations are used to determine the matches. However, if other information than the spatial location is available about each point, this can be used to ease the matching process if there is some correlation between the corresponding points and their attribute information, i.e., corresponding points should have similar attribute information. The Robust Point Matching (RPM) method with thin-plate-spline (TPS) has been extended to include attribute information in the energy function. This enables to exploit extra information available for each point.

A regionalisation scheme to break down a large, complex matching problem into several smaller problems and finally combine the sub-results has been proposed. The regionalised point matching serves two purposes. 1) to simplify a dense and locally varying disparity field relating corresponding points into several simpler disparity fields and 2) to reduce the number of points in the matching process and thereby reduce the computational cost. This is based on an assumption that a matching point should be found in the neighbourhood and therefore it is not necessary to attempt to match all points to all other points. The regionalised approach is suitable for point pattern matching methods robust to a large number of outliers.

A comparative study of three methods for protein spot matching, of which two have been proposed here, has been conducted on a number of real 2DGE image spot data.

A new method based on simultaneous segmentation and match of protein spots have been proposed and is also the main contribution of this thesis. The method, Elastic Graph Matching, uses the *a priori* knowledge of the spots location and neighbourhood interrelations from the reference gel as well as the new 2DGE

grey level image information is exploited. Most importantly, the prior image segmentation, necessary for the point matching methods, is not needed here.

Chapter 2

Motivation

Proteomics is an increasingly important part of cell biology and the efforts to understand the basic principles of life – how the living cell works. This chapter will give some basic introductory knowledge to proteomics, the process of two-dimensional gel electrophoresis for protein separation, and the motivation for applying image analysis in the field of proteomics will be further explained.

In proteome analysis, gel electrophoresis is a technique to separate proteins in a biological sample on a gel. The resulting gel images are by captured as a digital image of the gel. This image is then analysed in order to quantitate the relative amount of each of the proteins in the sample in question or to compare the sample with other samples or a database. The task of analysing the images can be tedious and is subjective (dependent on the human operator) if performed manually.

The use of digital image analysis in the field of proteomics is primarily motivated by the need to improve speed and consistency in the analysis of two-dimensional electrophoresis gel (2DGE) images.

The most important issues and challenges related to digital image analysis of the gel images will be addressed, namely the *segmentation* of the images and the *matching* of corresponding protein spots.

2.1 Biological Background

Knowledge of the basic principles in proteome analysis and gel electrophoresis provides a good background to understand the issues related to the image analysis part of the process – the main focus of this thesis. Readers familiar with the biological concepts and techniques may safely skip this part. Sections 2.2-2.4 where problems faced in image analysis are addressed should still be interesting.

2.1.1 Proteome analysis

A short definition of proteome analysis is: *identification, separation and quantitation of proteins*. The first publication of the word proteome was in 1995 by Wasinger et al. [85], and Wilkins [89] defines the concept of proteome analysis:

Proteome Analysis: the analysis of the entire PROTEin complement expressed by a genOME, or by a cell or tissue type.

In other words, the proteome is the complete set of of proteins that is expressed, and modified following expression, by the genome at a given timepoint and under given conditions in the cell.

The proteome provides us with much more information about the working of the living cell than the genome does. The genome is static and essentially identical in all somatic cells of an organism [32], where the proteome is constantly changing, reflecting the cell environment and also responding to both internal and external stimuli. The complete sequencing of the genome is not able to tell much about the function of the cell but analysis of the proteome is.

The techniques focused on here are *two-dimensional gel electrophoresis* (2DGE) combined with *mass spectrometry* (MS) and a general methods description for 2DGE and MS is given in Fey et al. [33]. A general introduction to the science of proteomics can be found in [1].

In the past years the extensive DNA sequencing efforts have provided hundreds of thousands of open reading frames in international databases. Unfortunately, a large proportion of this information has no or very little homology to any known protein. As one goes up the evolutionary tree this proportion increases (see Fig. 2.1) and even for one of the most extensively studied organisms, the relatively simple humble baker's yeast (*Saccharomyces cerevisiae*) 63% of the genes have either no or only limited homology to known proteins.

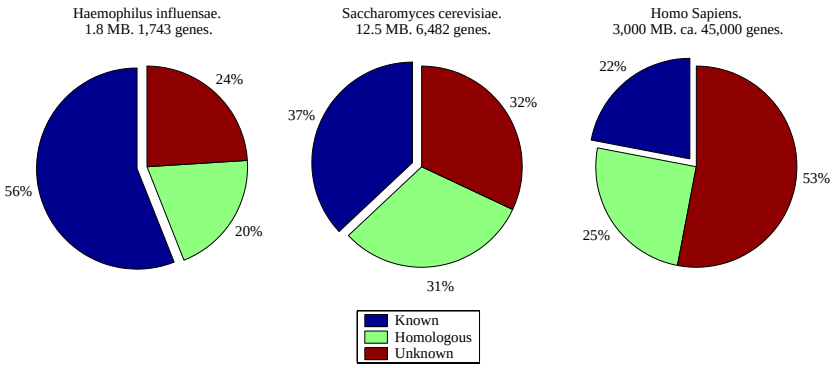


Figure 2.1: Division of sequenced genes into known, homologous and unknown categories for three different organisms. Data from CPA.

Furthermore, even when the open reading frame has some homology and the protein's function can be guessed at, many questions remain unanswered. These are questions as: Under what condition is the protein expressed? Where is it expressed in the organism? Where in the cell is it used? How is its expression regulated? Is the protein's expression affected in diseases (e.g. cancer, cardiovascular, auto-immunity or inflammatory diseases)?

To find answers to these questions is basically the motivation for studying the function of the gene products, namely the function of the proteins. By analysing expression patterns of the proteins under different conditions the function of particular genes can be determined and some of the questions posed above may be answered.

In proteome analysis, the technique of two-dimensional gel electrophoresis (2DGE) enables biotechnologists to generate protein expression patterns that can be digitised into images and analysed. Proteome analysis can provide a shortcut to identification of certain genes or groups of genes involved in, e.g., the development of severe illnesses. This is because the *differences*, quantitative and qualitative, in protein spot patterns between gels are related to the disease or treatment investigated.

Biological applications

Proteome analysis has a number of biological applications, examples include

- understanding of the basic principles of life,

- relating the genome and the environment to the organism's phenotype,
- drug development/evaluation (including toxicology and mechanism of action),
- disease prognosis, diagnosis, screening, monitoring of e.g.,
  - diabetes, all types of cancer, cardiovascular, and many more
- identification of new drug or vaccine targets,
- improvement of food quality,
- monitoring environmental pollution, and
- prevention of micro organism/parasite infections.

For instance in drug development, pharmaceutical companies spend large amounts[1] of resources on studying the drug effect in animal experiments. Some of these effects can be assessed by measuring changes in protein levels across different tissue samples.

2.1.2 Two-dimensional gel electrophoresis

Two-dimensional gel electrophoresis (2DGE) enables separation of mixtures of proteins due to differences in their isoelectric points (pI), in the first dimension, and subsequently by their molecular weight (MWt) in the second dimension as sketched in Fig. 2.2.

Other techniques for protein separation exist, but currently 2DGE provides the highest resolution allowing thousands of proteins to be separated. For a review of the latest developments in the proteomics field, please refer to Fey and Mose Larsen [32], where 2DGE and some of the candidate technologies to potentially replace 2DGE are presented along with their advantages and drawbacks.

The great advantage of this technique is that it enables, from very small amounts of material, the investigation of the protein expression for thousands of proteins simultaneously. After protein separation an image of the protein spot pattern is captured. Proper finding and quantitation of the protein spots in the images and subsequent correct matching of the protein spot patterns allows not only for the comparison of two or more samples but furthermore makes the creation of an image database possible.

[1] The cost of developing one new drug compound amounts to $\sim$ 300 million USD [30].

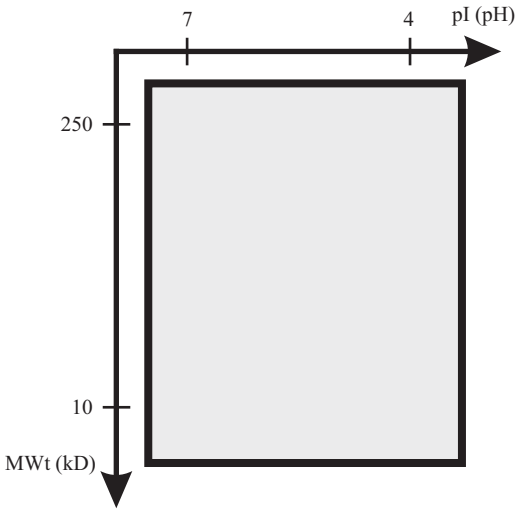


Figure 2.2: Schematic two-dimensional electrophoresis gel. Proteins are separated in two dimensions; horizontally by iso-electric point (pI) and vertically by molecular weight (MWt). No proteins shown. The pI and MWt ranges are example values.

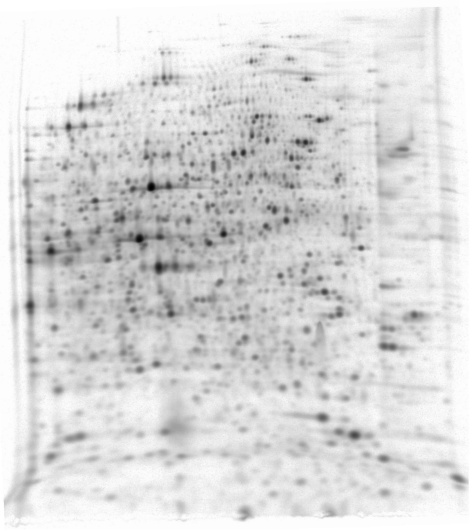
The changes in protein expression, for example in the development of cancer are subtle: a change in the expression level of a protein of a factor 10 is rare, and a factor 5 is uncommon. Furthermore, few proteins change: usually less than 200 proteins out of 15,000 would be expected to change by more than a factor 2.5. Multiple samples need to be analysed because of the natural variation, for example between individuals and therefore it is necessary to be able to rely on perfect matching of patterns of the new images.

Even though promising attempts have been made [13] to make the technique as reproducible as possible there are still differences in protein spot patterns from run to run. Also due to improvements in the composition of the chemicals used to extract as many proteins as possible the patterns become so dense (crowded) that locating the individual protein spots is a non-trivial task.

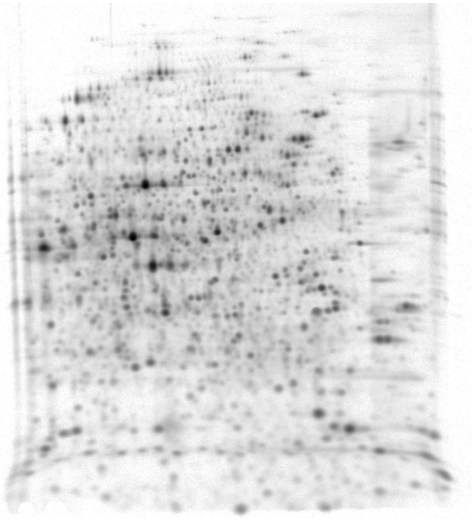
The laboratory process

The laboratory process as it is practiced in CPA is roughly sketched in the following. Some steps have been left out but please refer to a detailed description in [33].

Given a biological sample of living cells, e.g., a biopsy or a blood sample the process from the living cells to separated proteins on a gel will be explained. The



(a) Gel 1.



(b) Gel 2.

Figure 2.3: Two-dimensional electrophoresis gel images. Two different gels of baker's yeast *Saccharomyces cerevisiae* strain Fy1679-28C EC [pRS315].

procedure described here uses radioactive labelling, IPG for the first dimension, SDS polyacrylamide gels for the second dimension, and phosphor imaging to capture digital images of the protein patterns. Alternative visualisation methods will be described in § 2.1.2.

The 1st dimension, the incubation and, the 2nd dimension steps are illustrated in Fig. 2.4.

Labelling. A radioactive amino acid is "fed" to the living cells and all the proteins synthesised *de novo* may then contain the radioactive amino acid ([^{35}S]-methionine) in place of the non-radioactive one. The radioisotope used for the labelling is typically [^{35}S], but other radioisotopes, e.g., [^{32}P] or [^{14}C] can also be used. The radioactive labelling enables detection of the proteins later on. Duration: 20 hrs is the usual labelling interval used but this can be changed for specific purposes or situation.

Solubilisation. The cells' structures are broken down (killing the cells) and the proteins are dissolved in a detergent lysis buffer. The lysis buffer contains urea, thiourea, detergent (NP40 or CHAPS), ampholytes, dithiothreitol, all with the purpose of dissolving the proteins, unfolding them and preventing proteolysis. The actual procedure used depends on the sample itself and can take from less than 1 minute to 2 days.

1st dimension – isoelectric focusing. On an immobilised pH gradient (IPG) gel, in glass tube or on plastic strip, the proteins are separated according to their isoelectric point (pI). An electric field is applied across the gel and the 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 gel is the same as the pI of the protein, i.e., the pH value at which the number of positive and negative charges on the protein are the same. At this point no net electrical force is pulling the protein. See Fig. 2.4. Eventually all proteins will have migrated to their pI – their state of equilibrium. Duration: from 8-48 hrs. depending on the pH range of the IPG gel, e.g., 17.5 hrs for IPG pH range 4-7.

Incubation. In the incubation step the 1st dimension gel is "washed" in a detergent ensuring (virtually) the same charge on all proteins per unit length. Proteins are linear chains of amino acids. These fold up and can be cross-linked by disulphide bridges. The solutions that are used at CPA contain urea, thiourea and detergents which cause the proteins to unfold into long random-coil chains. Duration: 2×15 min.

2nd dimension – MWt separation. The incubated 1st dimensional gel strip is positioned on the upper edge of a polyacrylamide gel slab. See Fig. 2.4.

The second dimension acts like a molecular sieve so that the small molecules can pass more quickly than the large. Again, an electrical field is applied,

this time in the perpendicular direction, and proteins migrate into the gel. As all proteins have the same charge per unit length now, the same electrical force is pulling them. However, small ($\sim$ light) proteins meet less obstruction in the gel and will migrate with higher velocity through the gel. The larger proteins meet more resistance and migrate slower. Proteins with the same pI will migrate in the same "column" but will now be separated by *molecular weight* (MWt). As opposed to the 1st dimension process, the 2nd dimension has no equilibrium state because the proteins keep moving as long as the electric field is applied. The small proteins reach bottom of the gel first and the process has to be halted before they migrate out of the bottom of the gel. Duration: approx. 16 hrs.

Drying etc. The gel is dried on paper support requiring some manual handling. Duration: 20 min.

Image generation. The dry gel is put in contact with a phosphor plate which is sensitive to emissions. The radiation from the labelled proteins excites the electrons of rare earth atoms in the plate at positions where there is protein present in the gel. The larger amount of protein present at a specific location in the gel, the more electrons in the plate will be excited at that location. The amount of radioactive protein in the samples can be quite small (at the picogram level) hence the level of radiation is also small and the time required to expose the phosphor plate is long. After exposure, the phosphor plates are "read" using phosphor imaging technology where a laser beam excites the (already excited) electrons to an even higher energy state. The electrons return to their normal state while emitting electro-magnetic radiation (light). A CCD chip captures the light and a digital image is generated. Exposure time: usually 5 days. Image capture: 1 minute.

Alternative image generation

The radioactive [^{35}S]-methionine labelling described above is not the only technique to capture images of the separated proteins, although it is the most sensitive. Older methods using X-ray film to capture the image are still used. Staining with the Coomassie blue dye, silver or fluorescent dye can also visualise the proteins using spectroscopic techniques. Fig. 2.5 shows a comparison of four different visualisation methods on the same cell sample. Note how the [^{35}S]-methionine labelling technique (top left) results in an image with much more detail than the other techniques. Many more weak proteins are revealed using this technique. The most important methods for image generation are:

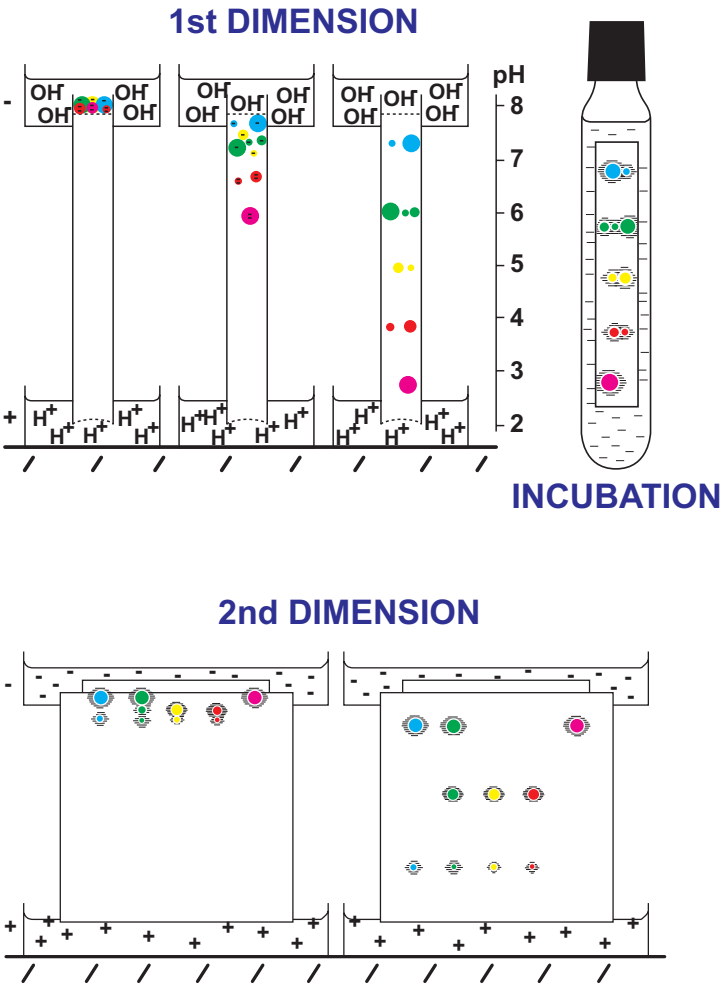


Figure 2.4: Diagram of the 2DGE process. By courtesy of CPA.

X-ray – autoradiography. Direct capture of irradiation on X-ray film by contact print (gel and film in contact).

X-ray – fluorography. Where the gel is impregnated with PPO (2,5 - diphenyloxazole) to amplify the signal. Contact print as above but the gel/film have to be placed at -70°C to speed up exposure.

Phosphorimager – autoradiography. Contact print where irradiation energy is captured by a rare earth complex (irradiation lifts electron into a higher orbital (meta-stable)) and then the plates are discharged pixel by pixel in the phosphorimager by a laser. The laser lifts the electron to an even higher, unstable state. The electron falls back to its normal orbital and the combined (irradiation plus the laser) energy is read.

Fluorescence. Monobromobimane binds covalently to cysteine (and in doing so becomes more strongly fluorescent) and is used to stain the proteins *before* electrophoresis. SyproRuby® binds to proteins in the gel after electrophoresis – contains some rare earth elements.

Silver staining. Gels are chemically treated in a similar fashion to photography in order to bind silver atoms to the proteins.

Image analysis

The protein pattern differences between gel images can be very subtle and tedious to detect by eye and therefore digital image analysis is a natural part of this process. By means of digital image analysis speed and objectivity can be greatly improved. Still, most existing commercial software for analysis of two-dimensional electrophoresis gels require a large amount of manual editing and correction of the spot segmentation and matching results. There is a need for development of better image segmentation and protein spot matching methods [83],[32], and this is the main motivation for this work. §§2.2, 2.3 and 2.4 present some of the issues and challenges that will be discussed in the remaining of the thesis.

2.2 Issues in Image Segmentation

The segmentation of an electrophoresis gel image basically consists of distinguishing the protein spots from the background. There are however several issues that make the segmentation process non-trivial. In a typical gel image with 1900+ protein spots, the strongest third of the spots account for more

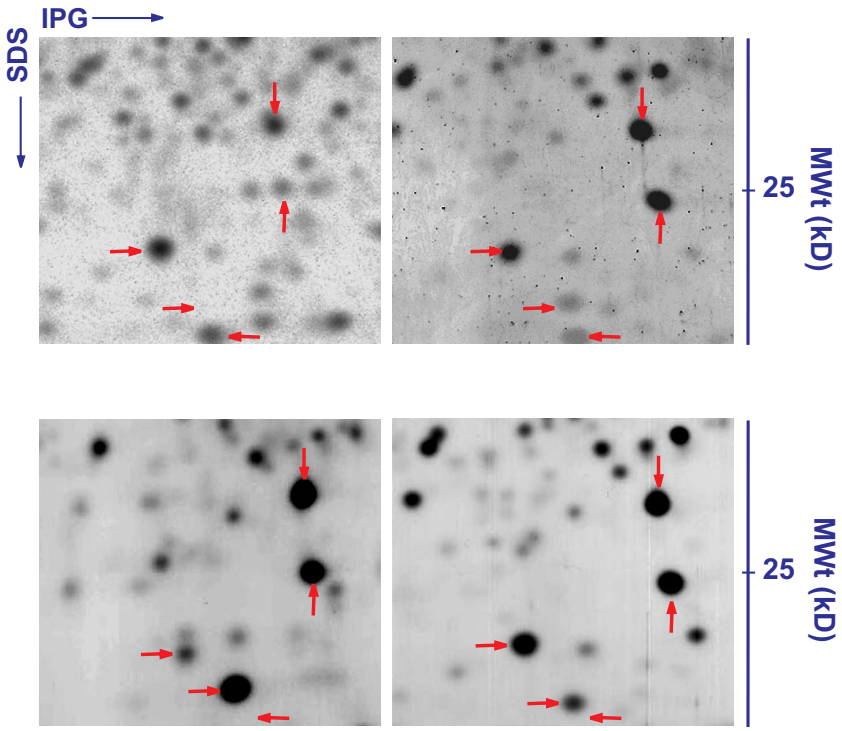


Figure 2.5: Comparison of four different protein visualisation methods. The same sample (HeLa cells) is used for all four visualisation methods. Top left: $[^{35}S]$-methionine labelled (2 mio. cpm). Top right: SyproRuby® stained (100 μg protein). Bottom left: Mono Bromo Bimane labelled (100 μg protein). Bottom right: Silver stained. Only a small part of the gels is shown. By courtesy of CPA.

than 75% of the total amount of protein in the sample and the weakest third of the spots account for less than 6% of the total protein amount. The distribution of protein is in other words very skew and Fig. 2.6 shows the histogram of the so called "percentage integrated intensity" (%II, see §C.2) for the protein spots in a gel image. For [^{35}S]-methionine labelled proteins %II is proportional to the amount of protein present, the rate of turnover of the protein and the number of methionine residues in the protein. A protein without methionine will not be detected irrespective of its amount. Similarly, a protein with a high rate of turnover will appear to be more abundant if the labelling interval is short (e.g., less than 2 hours). For bimane it is proportional to the number of cysteines. About 2% of human proteins do not have methionine and 8% do not have cysteine. For SyproRuby® and for silver staining, it is not well defined.

The integrated intensity (II) is calculated as the sum of pixel values inside the spot borders in an "inverted" gel image, i.e., when spots appear bright on a dark background. The %II for a spot is the II normalised with the sum of IIs inside all spots in the gel image. The relatively large number of weak spots combined with a high spatial density of spots is one of the main challenges in the image segmentation.

Accurate quantitation is very important because, as mentioned earlier it is often subtle changes that are seen in comparing two samples from for example normal and cancerous tissue.

To demonstrate how weak spots appear, Fig. 2.7 shows three small example gel regions from the same gel image. The top row is the image regions and in the second row, the same regions are shown with spot centres overlaid. Note the high level of noise compared to the weak spots.

A second challenge, in segmentation of the image into (separate) spots and background, is the fact that overlapping spots is not a rare phenomenon. At CPA mass spectrometry has shown that in standard gels covering the pH range from 4 to 7, more than 60% of the spots represent more than one protein. For this reason CPA is moving towards running gels covering single pH regions e.g., 5.0-6.0. For this type of gels, it is known that only 5% of the spots have more than one protein present (with current sensitivities for the mass spectrometry). Fig. 2.8 displays three example regions with typical cases of neighbouring spots that are located so close that they overlap each other. Overlapping spots are naturally harder to detect (and separate) than isolated ones.

The intensity of the image background can vary across the image. A typical gel image with varying background is shown in Fig. 2.9. Intensity profiles are picked up along the horisontal ($y = 250$) and vertical ($x = 200$) lines and shown in Fig. 2.10. The trends in these lines show generally a larger background variation in the vertical direction and higher background intensity at the edges

of the gel than in the gel centre. The latter is probably due to a larger spot intensity in the centre of the gel. Thus, the main challenges in segmentation of electrophoresis images are:

- noise / very weak (low intensity) spots,
- overlapping spots, and
- varying background.

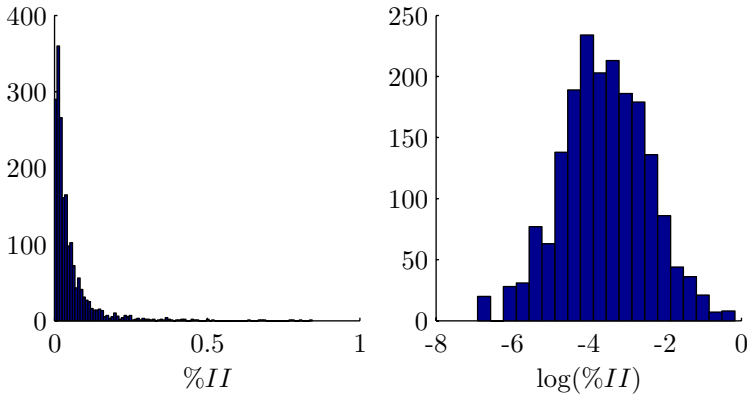


Figure 2.6: Histograms of %II and log(%II) for a gel image with 1919 spots.

§3 will present some general segmentation techniques based on mathematical morphology and scale space blob detection. Furthermore, some parametric spot models are investigated.

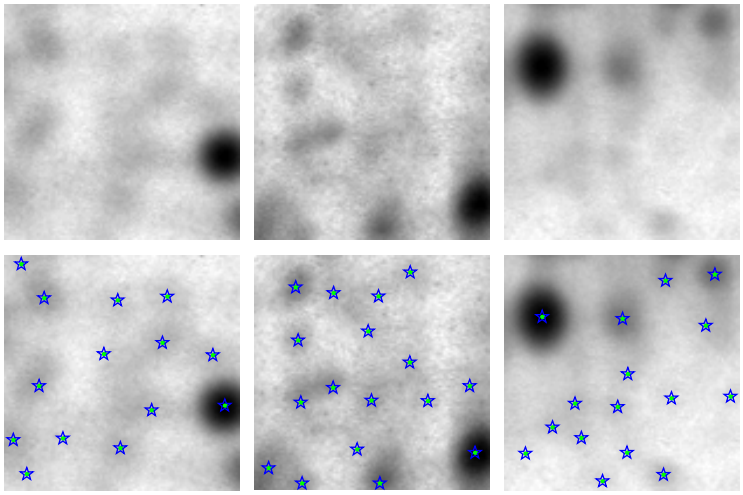


Figure 2.7: Example images of gel regions with low signal to noise ratio – *low intensity spots*. Top row: region 1, 2, and 3 from same gel image. Bottom row: same regions with known spot centres overlaid. The regions are 100×100 pixels and the grey level range in each region has been scaled appropriately to improve visual inspection. Region 1, 2, and 3 contain 14, 19, and 17 spots, respectively.

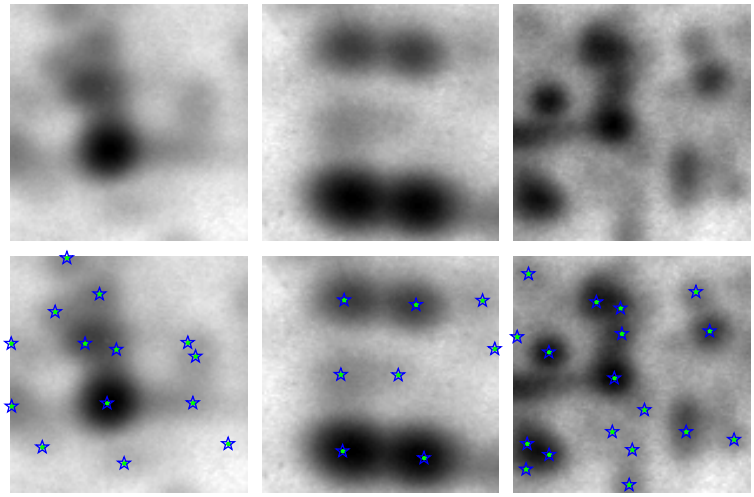


Figure 2.8: Example image of gel regions with *overlapping spots*. Top row: region 1, 2, and 3 from same gel image. Bottom row: same regions with known spot centres overlaid. The regions are 100×100 pixels and the grey level range in each region has been scaled appropriately to improve visual inspection. Region 1, 2, and 3 contain 11, 8, and 19 spots, respectively.

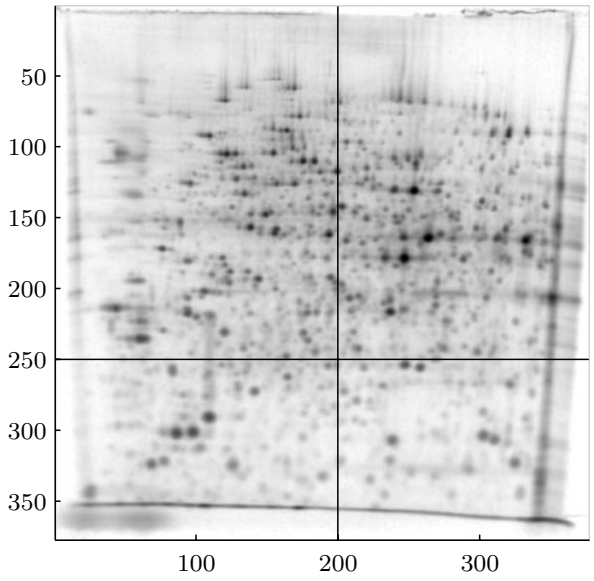


Figure 2.9: Example image of gel with typical varying background. Intensity profiles are picked up along the horizontal ($y = 250$) and vertical ($x = 200$) lines and shown in Fig. 2.10.

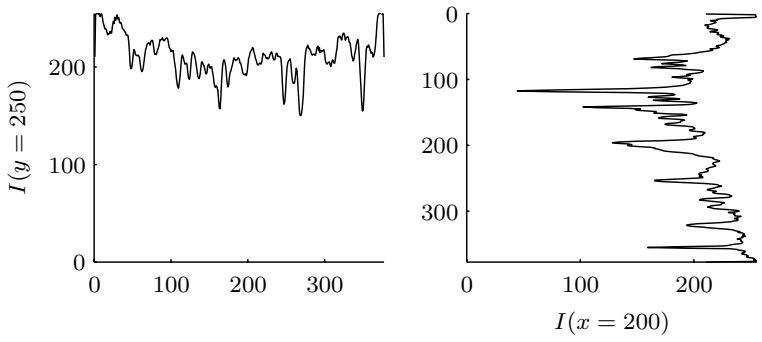


Figure 2.10: Intensity profiles along the horizontal and vertical lines, respectively in Fig. 2.9.

2.3 Issues in Spot Matching

Spot matching is a central issue in electrophoresis [83] and is also the main focus of this work. The goal is to establish protein spot *correspondence* between gel images in order to *detect changes* in protein expression levels or discover *new* proteins that are only detected in one of the images. For comparison of protein levels across several gels a correct match or correspondence between the protein spots is necessary. In matching up the protein spot patterns from two gels it is necessary to solve the *correspondence problem*. The task is to determine the exact (correct) correspondence between known spots in a *reference* gel image and the spots in an "incoming" gel image with protein spots. The new incoming gel is referred to as the *match* gel. Fig. 2.11 shows a sketch of the correspondence concept.

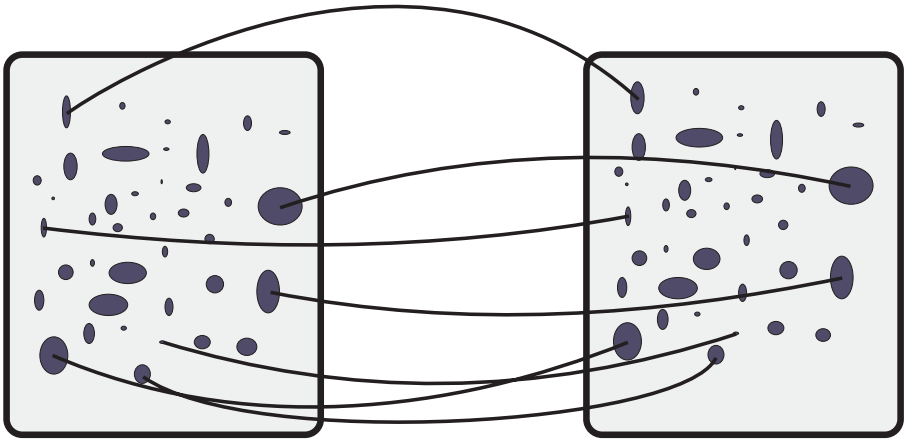


Figure 2.11: Principal sketch of (partial) correspondences between protein spots in two gel images. In order to compare protein expression levels between two gels the correct correspondence between matching spots is necessary.

For the purpose of spot matching, the problem of matching *points* instead of spots, i.e., matching the spot centres instead of the entire spots, is regarded most often. Also the main focus will be on matching *two* set of spots from two different gel images.

In Fig. 2.12 two electrophoresis gel images are shown with known protein spot centres overlaid as point sets. These two point sets (or point patterns) are shown together in Fig. 2.13 where corresponding points are connected with small arrows. The arrows can be interpreted as a disparity field. This will be the standard way of displaying the point correspondence throughout the thesis. To the left in Fig. 2.13 is shown the correspondence when the point sets are simply plotted together. Clearly a large contribution to the disparity field

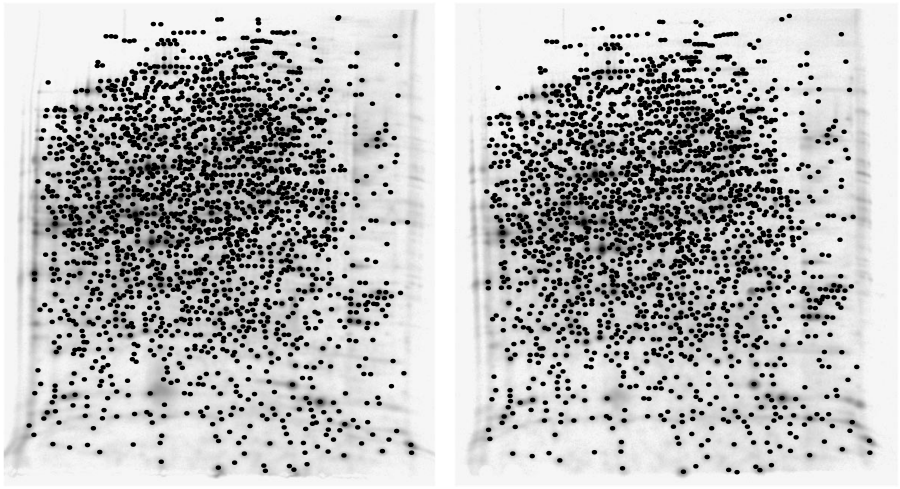


Figure 2.12: Gel images with known spot centres overlaid as points. Left: gel A (1919 spots), right: gel B (1918 spots). 1918 spots in common, which means that one spot present in gel A is missing in gel B.

stems from a rotation and a translation, probably due to the handling of the gels. To the right 20 landmarks have been hand-picked and the parameters in a first order polynomial transformation has been computed (see §4). The entire point set from gel B has been transformed according to the parameters found from the landmarks. The plot to the left shows the residual after this transformation (mainly a translation and a rotation) has been removed. The residual disparity field exhibit local, highly non-linear behaviour. Together with the high denseness of the points this constitutes the main challenges in the point matching task at hand.

The gels in Fig. 2.12 contain a different number of spots (1919 and 1918) respectively. One spot present in gel A is missing in gel B. If protein expression is very low this can cause the spot not to show up in the gel. This situation of extra or missing spots is not unusual and, in fact, very interesting from a biological point of view. A point occurring in only one of the gels will be referred to as *outliers* or *singles* and a pair of matching spots is referred to as a spot *pair*.

As seen from the Figs. 2.12 and 2.13 the point patterns to be matched possess no easily recognisable shape structure. Often, in other point pattern matching applications the exact match of certain points is not important, instead a good match of shape structures is sufficient.

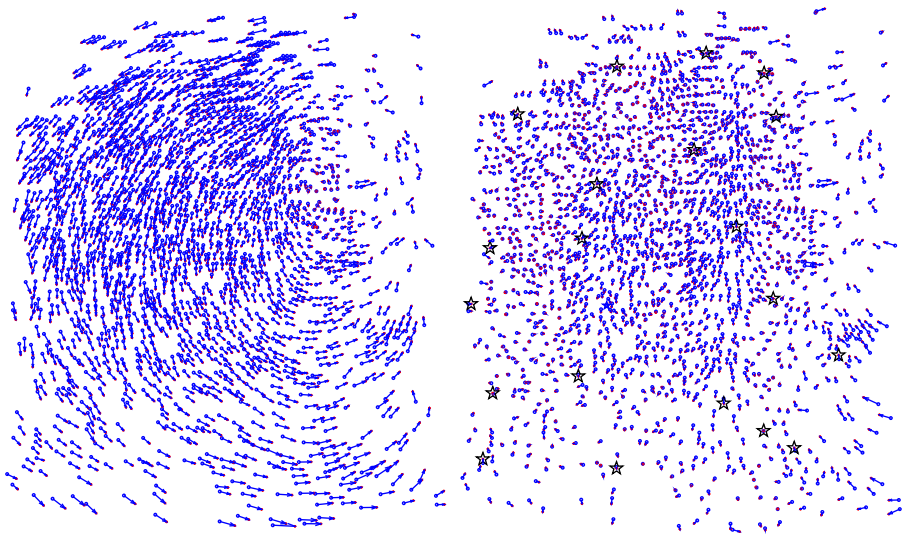


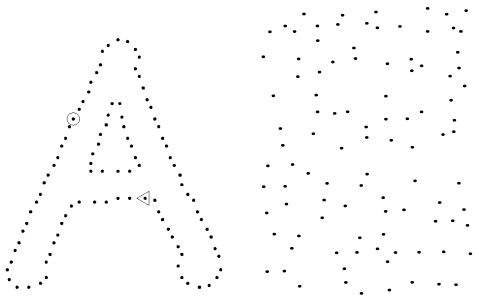
Figure 2.13: Known correspondence between spots in gel A and gel B. Corresponding spots are connected with small arrows. Left: Before initial alignment. Right: Residual after correction for 1st order polynomial transformation using landmarks. Landmarks are manually defined and marked with stars.

2.3.1 Properties of 2DGE spot patterns

It seems that some point patterns have more structure or shape than others. E.g., a pattern of the letter "A" shown in Fig. 2.14(a) exhibits far more structure than the sub pattern of a 2DGE spot pattern in Fig. 2.14(b). In texture analysis [18], the notion of more or less *stochastic* textures is used to describe how well-ordered the texture is. This terminology is adopted for point patterns. Similarly, point patterns can be described as more or less stochastic ranging from pure stochastic to pure deterministic. No quantitative measure for the degree of stochasticness is defined, but one could imagine an entropy-based measure to be suitable. In Fig. 2.14(a), the "A" is said to have a deterministic nature and the spot pattern (Fig. 2.14(b)) is more stochastic or amorphous.

The neighbour relations may be useful in describing the degree of stochasticness. In the "A" each point have two or three natural neighbours, because the points form a shape. In the 2DGE case there are none such shape and all neighbours are equally important.

It is hard to quantitate the idea into a measure, but the purpose of these remarks is to underline the difference in point patterns that make different point matching methods more or less suitable.



(a) Deterministic pattern. Disregard the circle and the triangle markers. From Belongie et al. [7].

(b) Stochastic pattern. 2D electrophoresis gel spot centres.

Figure 2.14: Deterministic and stochastic point patterns. Both patterns contain 100 points.

It seems acceptable that the more stochastic (less shape structure) the pattern is the more difficult it is to obtain correct matches.

Methods for protein spot matching should not specifically rely on the fact, that the patterns are deterministic.

Furthermore, when matching shape structures, it is usually an acceptable result when the shapes have a reasonable overlap. In the case of matching 2DGE spot patterns, a *correct* match of each point is a requirement.

The above considerations lead to five main requirements of a spot matching method. It must be able to:

- exactly and robustly match protein pairs,
- allow for non-linear distortions/transformations,
- robustly handle outliers in both sets,
- be able to handle point sets of stochastic/amorphous nature, and
- robustly match dense point sets.

In § 4 a number of point matching methods are presented for general point matching purposes, as well as for matching protein spot patterns.

2.4 Unified Approach

As the previous sections have implied there might be good reasons to combine the segmentation and the matching into one method, i.e., to find (locate) the spots in a new gel while simultaneously matching up the spots with spots already known in a reference gel. § 5 will discuss such an approach.

2.5 Protein Pattern Databases

The large amount of gel data can be organised in image databases and Fig. 2.15 shows an example of the construction of such a database. From a set of normal gels a normal composite gel is generated and similarly for a set of gels representing diseased subjects. The composite gels are formed as the *union* of the contributing gels. From the normal and diseased composite gels the database gel is formed and marker proteins can be identified.

One of the more important data often missing in most image databases is protein expression data (under given environmental growth conditions) [32], i.e., only gel images are presented. There exist also many other databases of biological data. These may include gene and protein sequence data, protein identification (unique protein code), protein function, theoretical values for the isoelectric point (pI) and the molecular weight (MWt), biochemical pathways, chemical data and the scientific literature all of which can be very useful in interpreting the data from the 2D gel image databases.

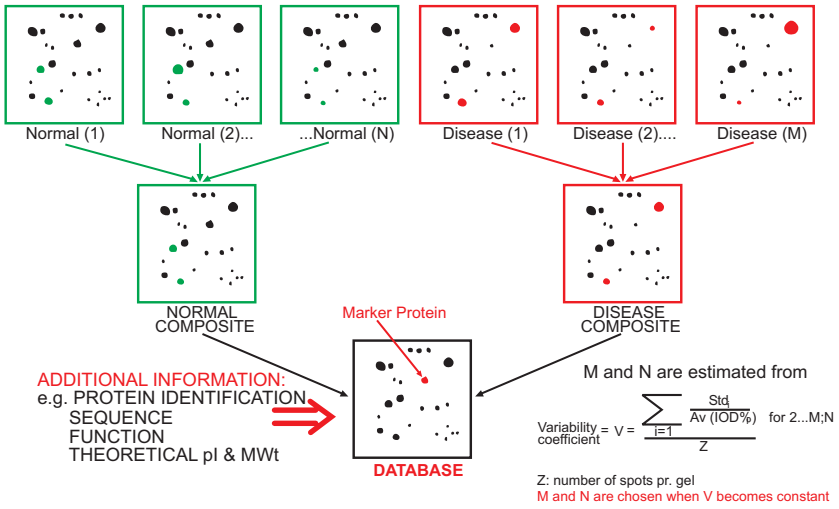


Figure 2.15: Construction of a 2D gel image database. By courtesy of CPA.

Chapter 3

Gel Segmentation

In a recognition system a preprocessing step to segment the pattern of interest from the background, noise etc. usually precedes [44] the actual recognition process and for the current task this is no exception. The two-dimensional electrophoresis gel images show the expression levels of several hundreds of proteins where each protein is represented as a blob shaped spot of grey level values.

In order to apply point pattern matching methods to solve the problem of matching spots from different images each spot must be reduced to a pattern (e.g., a point – the spot centre). It is of crucial importance that the segmentation is correct in order to obtain correct quantitation of protein expression and a successful matching result. The matching becomes meaningless if the input is an erroneous segmentation. The segmentation task at hand consists of a separation of the image into what is background and what is spots and the challenging part is the cases of overlapping spots, varying background and a high level of noise in the images. Please refer to § 2.2 for examples.

Although the segmentation is an extremely important step, it is not the main focus of this thesis. Therefore, this chapter will only touch upon a few approaches to segmentation of gel images. These include methods based on mathematical morphology, parametric spot models and a Gaussian scale space blob detector.

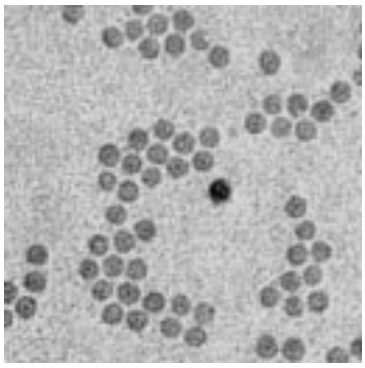


Figure 3.1: Original ferrit nanoparticle image. Microscope image of nano particles.

To illustrate the methods an image of nanoparticles with a number of distinct dark blobs will be used (Fig. 3.1). The particles are of relatively uniform size, intensity and shape. The nanoparticle image is overly simple compared to the 2DGE images and it is used here for illustration purposes only. The almost constant background, and blobs of almost identical shape and intensity facilitates the illustration of ideas in the segmentation process.

3.1 Mathematical Morphology Based Methods

Mathematical morphology in image analysis is a vast field of research and even though it is beyond the scope of this thesis some selected topics will be discussed. Some of the earliest work on segmentation of electrophoresis gels using mathematical morphology is by Beucher et al. [11, 12] who proposed to use a watershed based method for the segmentation of the images. These ideas are now well known and commonly used to segment images of electrophoresis gels. Other more recent approaches [74, 81] deal with the problems of oversegmentation by using marker controlled watersheds.

Another technique from the mathematical morphology is the so called *h-domes*, which is a grey-scale reconstruction method. After a brief introduction to the method some examples of gel image segmentation will be showed.

3.1.1 Watersheds

In geoscience terminology, a watershed line is the outline of a catchment basin, which again, is an area of land that drains to a common point. When it rains on an area all drops landing within the same watershed lines will eventually drain to the same point – the minimum of the catchment basin. Viewing grey-scale images as landscapes, i.e., as topographic reliefs where the pixel values represent the surface height, notions as valleys, tops, ridges, catchment basins and watershed lines can be introduced. In the fields of image analysis and mathematical morphology various methods using watersheds as a segmentation tool have emerged. One of the fastest techniques developed by Vincent and Soille [82] is based on the immersion principle.

The immersion principle

Imagine a grey-scale image as a landscape with tops, valleys etc. where the pixel grey value corresponds to the terrain height, i.e., dark areas of the image (low pixel values) correspond to a low altitude area in the landscape and vice versa. Now pierce a hole in *all* local minima of the surface and slowly immerse the landscape model in water. The water will trickle out from the minima starting with the global minimum. At some point, as the water level rises from different minima, two neighbouring basins will meet and merge. At the pixels where the water from the two neighbouring basins meet a dam is raised. Continuing like this until the entire landscape is immersed in water will result in a partitioning of the grey-scale image into a large number of catchment basins (as many as the number of local minima in the image). Each catchment basin associated with a local minimum is now bound by a dam and these dams constitute the watershed lines.

Over-segmentation

A major disadvantage of the watershed segmentation is its tendency to over-segmentation. A noisy image with many local minima will segment into a large number of sections. Among ways to overcome this are *marker controlled watersheds* [81] and Gaussian scale space based multi-scale techniques [63].

The marker controlled watershed transform is a restricted form of the watershed method where holes are pierced in selected minima *only* (the markers). This way the number of catchment basins is controlled and over-segmentation is avoided. How to automatically choose a good set of markers is, however, seldom trivial. § 3.2 describes a blob detector that is suitable for this purpose in the case of

segmentation of 2D electrophoresis gel images. § 3.4 contains experiments of watershed segmentation with and without markers of electrophoresis images.

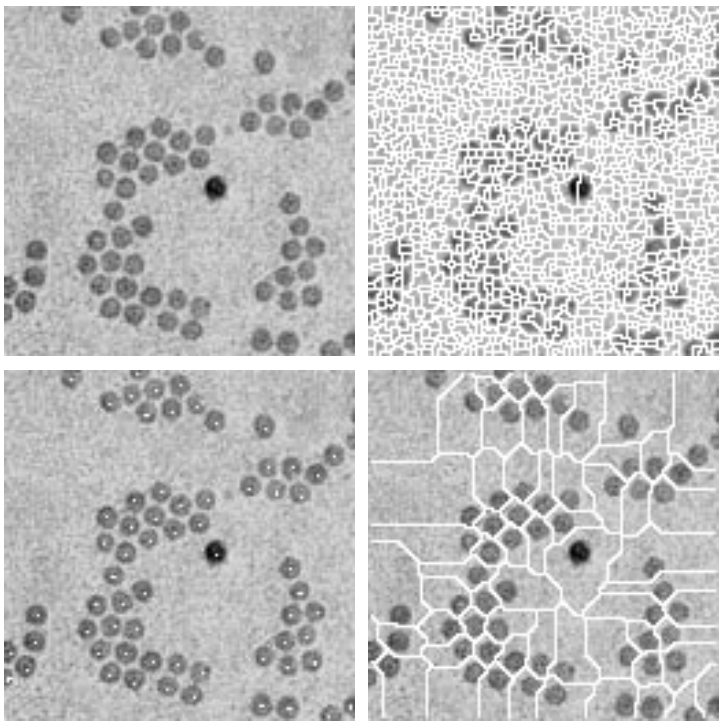


Figure 3.2: Example of watershed with and without markers. The goal is to find the dark circular nanoparticles. Top left: original image. Top right: watersheds of original image. This result is a severe over segmentation. Bottom left: original image with markers obtained using the method described in § 3.2. Bottom right: result of marker controlled watershed segmentation. Each particle has its own catchment basin.

3.1.2 H-domes

Another morphology based method the so called h-dome transformation is a technique to determine "maximal structures" in grey-scale images. Vincent [81] defines the h-dome image $D_h(I)$ of a grey-scale image I as

$$D_h(I) = I - \rho_I(I - h), \tag{3.1}$$

where $\rho_X(Y)$ is the grey level reconstruction of X from Y. The grey level reconstruction is obtained by iterative geodesic dilations of Y under X until stability is reached. Please refer to [73, 76, 81] for a general introduction to the subject.

The principles of h-dome extraction is demonstrated in Fig. 3.3 where the gel image has been inverted for illustrational purposes. Note that not all blobs are detected as h-domes. From the profile lines it is seen that the most intense spot on the horizontal line has a "right shoulder" stemming from a neighbour *overlapping* spot. This overlapping spot is not a regional maximum and therefore it is not an h-dome.

Fig. 3.4 shows the h-dome extraction on the inverted gel image and § 3.4 presents more experiments with different values of h on a two-dimensional electrophoresis gel image.

The electrophoresis images are traditionally viewed as dark protein spots on a bright background. If preferred, the extraction of *minimal* structures (spots) can be done using the *h-basins* (reverse of h-domes) as demonstrated by Horgan and Glasbey [43] on non-inverted electrophoresis gel images.

3.2 Scale Space Blob Detection

Protein spots in electrophoresis images may be detected using a blob detector because the spots possess blob characteristics. They do however vary in size and therefore the scale space theory is a natural framework for this task. This theory provides several feature detectors [56] like corner, junction, ridge, and blob detectors. A thorough review of the scale space theory is out of the scope here so the reader is kindly referred to [55, 77].

Blob detection can be formulated in terms of local extrema in the grey-level landscape (chapters 7 and 10 in [55]), e.g., as spatial minima in the Gaussian or as extrema in the Laplacian operator (chapter 13 in [55] and [14]). More specifically, as described in [56] (p. 9, Eq. (18)), which is used in the present implementation of the latter.

This presentation is formulated for continuous signals in two dimensions (2D). The concepts used here do however translate to discrete signals [54].

Given an image I and an isotropic 2D Gaussian $g(\mathbf{x};t)$,

$$g(\mathbf{x};t) = \frac{1}{2\pi t} e^{-\frac{x^2+y^2}{2t}}.$$

t is referred to as the scale-parameter and L is defined as the convolution of I with $g(\mathbf{x};t)$:

$$L(\mathbf{x};t) = I * g(\mathbf{x};t).$$

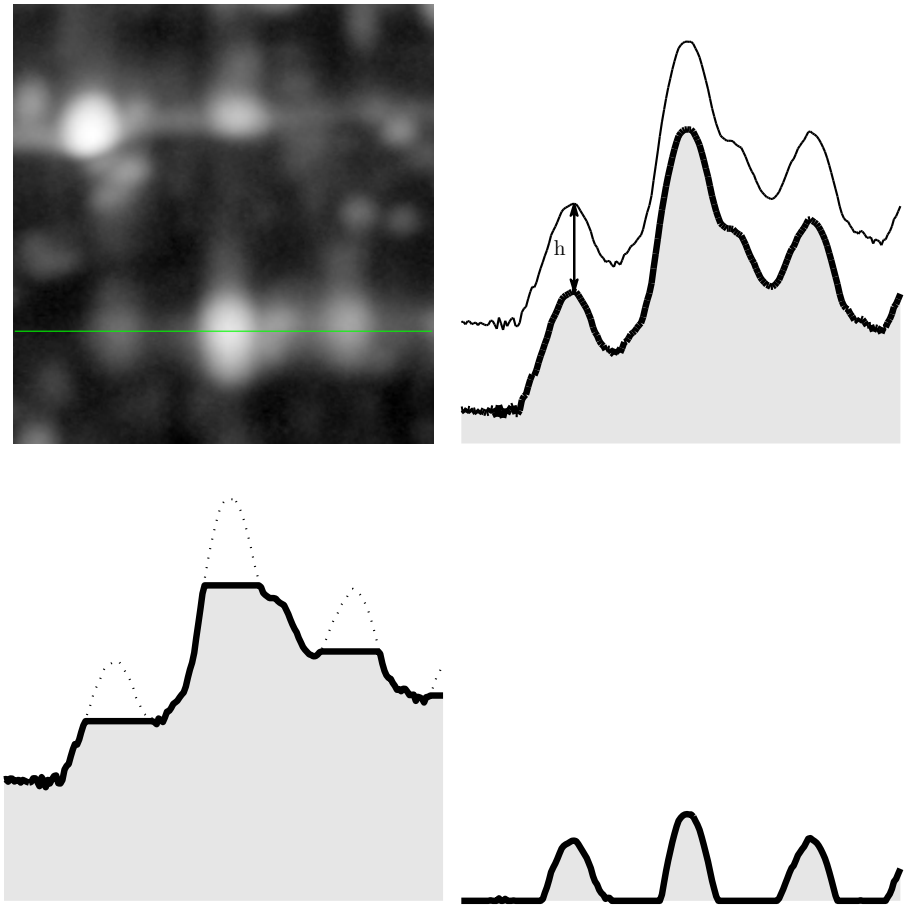


Figure 3.3: Principal sketch of h-dome extraction, $h = 0.25$. Top left: small region of two-dimensional electrophoresis gel image with horizontal profile line. Top right: intensity profile (thin line) and intensity profile minus h (bold line). Bottom left: grey level reconstruction (bold) and original intensity profile (dotted). Bottom right: *h-domes* result after subtraction. Note that not all blobs are detected as h-domes. It is seen that the most intense spot on the horizontal line has a "right shoulder" stemming from a neighbour *overlapping* spot. This overlapping spot is not a regional maximum and therefore it is not an h-dome.

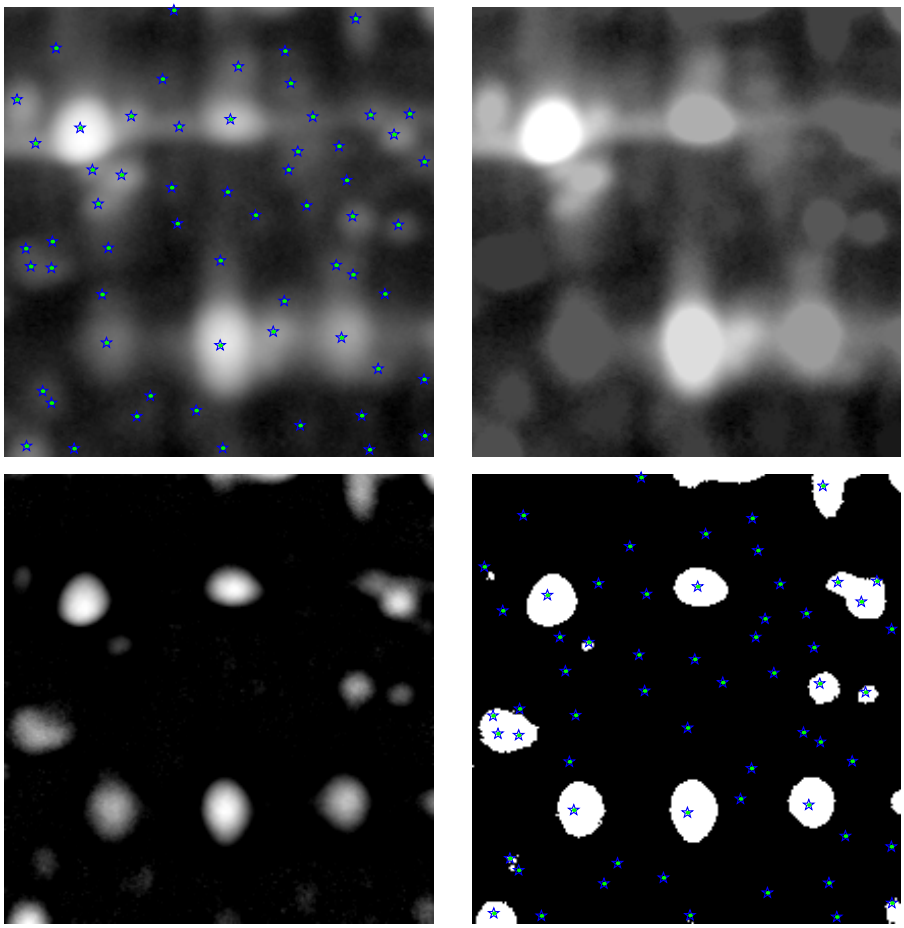


Figure 3.4: H-dome extraction of two-dimensional electrophoresis gel image, $h = 0.25$. Top left: small region of two-dimensional electrophoresis gel image I with ground truth operator specified spot centres (as specified in § C.2) overlaid. Top right: grey level reconstruction $\rho_I(I - h)$. Bottom left: H-domes of I, $D_h(I)$. Bottom right: $D_h(I) > 0.05$ with spot ground truth spot centres overlaid. Note how only major isolated spots are found and overlapping spots are either not found or detected as a single spot.

Scale-space derivatives in 2D at scale t are defined as the result of convolution with differentiated Gaussians:

$$L_{\mathbf{x}^\alpha}(\cdot;t) = \partial_{x^{\alpha_x} y^{\alpha_y}} L(\cdot;t) = \left(\partial_{x^{\alpha_x} y^{\alpha_y}} g(\cdot;t)\right) * I$$

where α is the order of differentiation.

With the Laplacian of L, F:

$$F \;=\; \nabla^2 L \;=\; \Delta L \;=\; L_{xx} + L_{yy}$$

and the Hessian $\mathcal{H}$ of F:

$$\mathcal{H}(F) = \begin{bmatrix} F_{xx} & F_{xy} \\ F_{yx} & F_{yy} \end{bmatrix} = \begin{bmatrix} L_{xxxx} + L_{xxyy} & L_{xxxy} + L_{xyyy} \\ L_{xxxy} + L_{xyyy} & L_{xxyy} + L_{yyyy} \end{bmatrix}$$

the determinant and the trace of $\mathcal{H}$ can be written as:

$$\det \mathcal{H}(F) = (L_{xxxx} + L_{xxyy})(L_{xxyy} + L_{yyyy}) - (L_{xxxy} + L_{xyyy})^2$$

and

$$\operatorname{tr} \mathcal{H}(F) = L_{xxxx} + L_{yyyy} + 2L_{xxyy}.$$

Critical points in the Laplacian of L, i.e., in F are found as zero crossings of F_x and F_y

$$F_x = 0 \quad \text{and} \quad F_y = 0. \tag{3.2}$$

To detect blobs it is furthermore required that the critical points are extrema:

$$\det \mathcal{H}(F) > 0, \tag{3.3}$$

and to ensure that the extrema are also minima

$$\operatorname{tr} \mathcal{H}(F) < 0 \tag{3.4}$$

must be fulfilled.

For this blob detector, the normalised *strength measure* for scale selection is [56]:

$$t\nabla^2 L.$$

Large blobs have maximum response at large scales and vice versa. Therefore, the scale can be used to characterise the size of a blob. Lindeberg [55] provides

an automatic scale selection based on maxima over scale in the normalised measure of blob strength.

Fig. 3.5 shows the steps of scale space blob detection of particles in the nanoparticle image. The top row shows to the left L, the Gaussian blurred version of the original image (a). Then follows in the same row the Laplacian of Gaussian, F, the determinant of the Hessian of F, and the trace of the Hessian of F. The second row displays to the left F (a) with contour lines where $F_x = 0$ (blue) and $F_y = 0$ (red). (b) is a binary image of the critical points of F (satisfaction of (3.2)), i.e. crossings of the contours in (a). The two last images in second row show binary images of the conditions in (3.3) and (3.4) respectively. The last row shows (a) the three binary condition images ((b), (c), (d) from second row) added. The brightest values indicate positions where all three conditions hold, i.e., where blobs are detected. Image (c) in the last row show a simple threshold of L to remove unwanted blobs. The last image is the original image with resulting blob markers overlaid.

This example is however only for a single fixed scale ($t = 3$) and the number of detected blobs greatly depends on the choice of scale. In this case of relatively uniformly sized spots all particles seems to be detected correctly at the same scale. Fig. 3.6 shows how the number of detected blobs decrease as the scale increases.

For electrophoresis gel images however, the protein spots exhibit large variations in size and automatic scale selection [57] is one approach that may be suitable. Linking of segments across multiple scales is another. Olsen et al. [63] study multi-scale watersheds and the structural changes as scale increases. Their scheme to link watershed segments across scales seems promising for segmentation of electrophoresis gel images. Scale space linking and automatic scale selection will, however, not be pursued further here.

§ 3.4 shows some examples of the described methods applied to electrophoresis images.

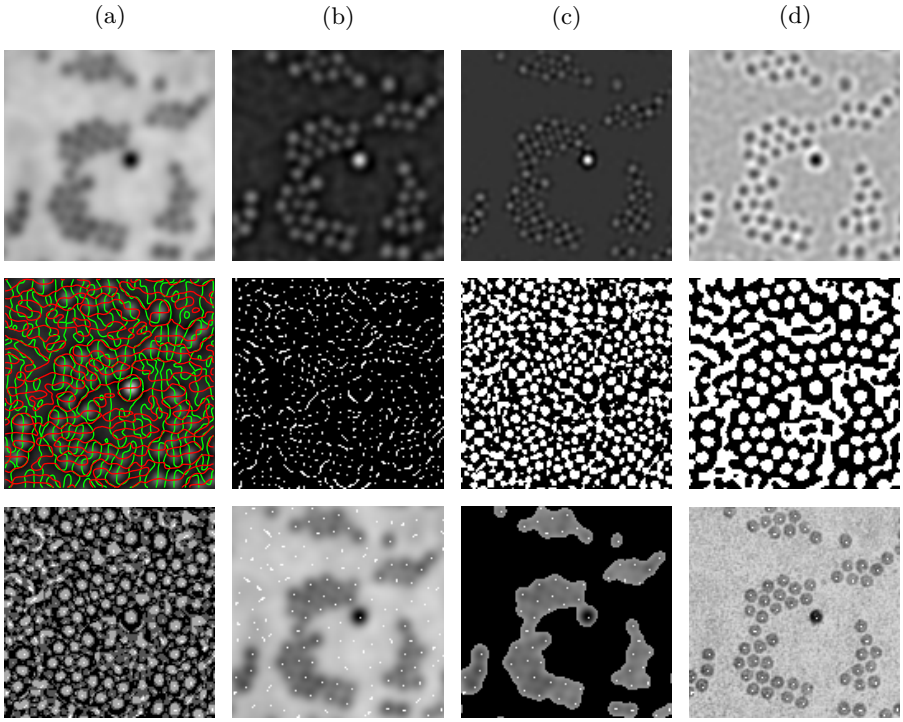


Figure 3.5: Example of scale space blob detection on nanoparticle image. **Top row:** (a) Original image blurred with a Gaussian filter ($t = 3$), L. (b) Laplacian of Gaussian, F. (c) $\det(\mathcal{H}(F))$. (d) $\mathrm{tr}(\mathcal{H}(F))$. Centre row: (a) Critical contours (3.2) of F, green: $F_x = 0$, red: $F_y = 0$. (b) Critical points of F. (c) Extrema condition (3.3) $\det(\mathcal{H}) > 0$. (d) Minima condition (3.4) $\mathrm{tr}(\mathcal{H}(F)) < 0$. **Bottom row:** (a) All 3 necessary blob conditions superimposed. (b) Blob markers detected at this scale. (c) Simple threshold to remove non-particles. (d) Markers of particles found overlaid the original image.

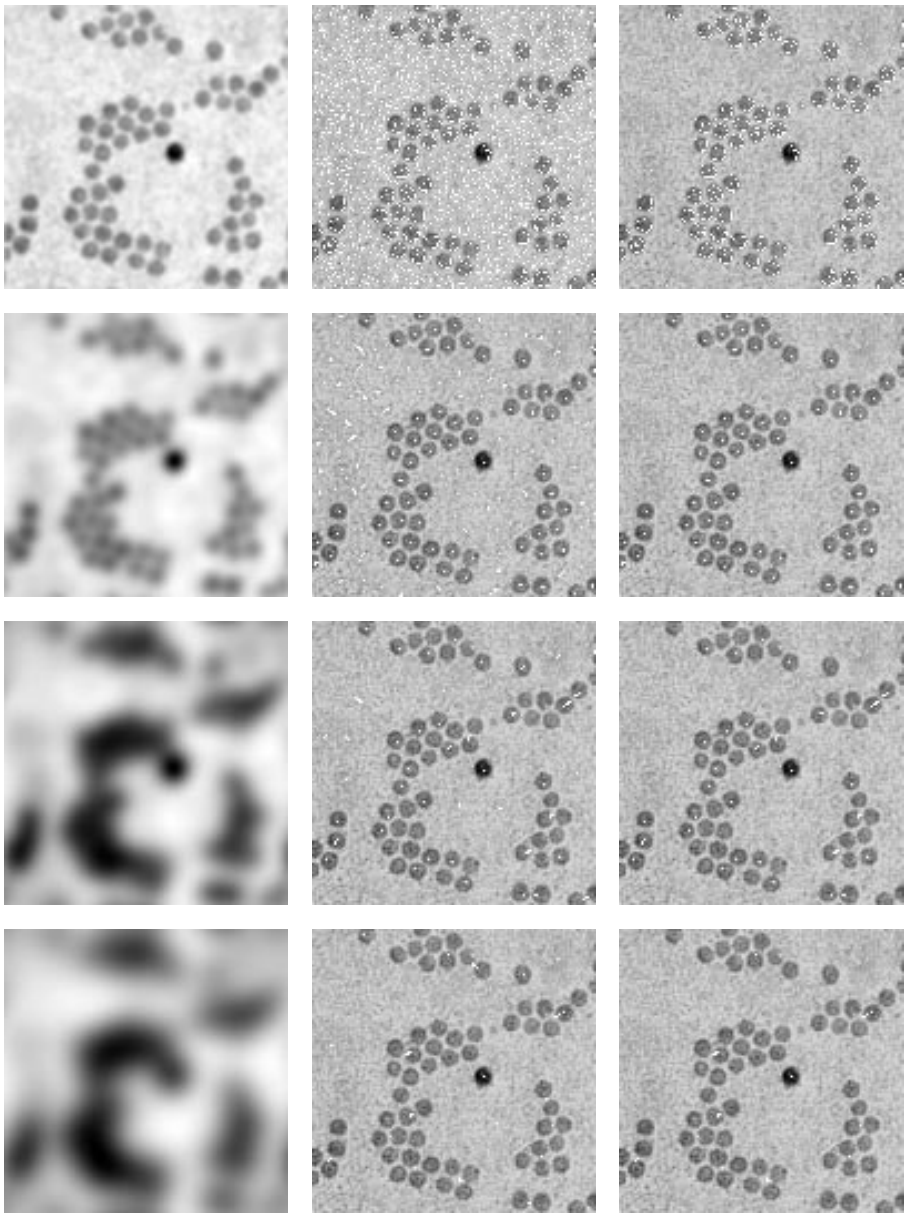


Figure 3.6: Scale space blob detection at 4 different scales. Scale t increases from top and down. $t = 1, 3, 5$, and 7. Left column: $L(\cdot; t)$. Centre column: Blobs markers at this scale. Right column: Blob markers inside threshold mask (see Fig. 3.5).

3.3 Parametric Spot Models

Parametric spot models rely on the assumption that the protein spots have some common characteristics that can be captured by a model. The idea is to define a suitable spot model $C(\mathbf{x}, \boldsymbol{\theta})$ and adjust the parameters $\boldsymbol{\theta}$ in the model so that the model fits the data (image $I(\mathbf{x})$) by some measure. $I(\mathbf{x})$ is defined on a square region, w around the protein spot in question and the range of $\mathbf{x}$ is denoted the *support* for the model. More formally: given an image $I : \mathbb{Z}^2 \mapsto \mathbb{R}$ and a model $C(\mathbf{x}, \boldsymbol{\theta}) : \mathbb{Z}^2 \times \mathbb{R}^p \mapsto \mathbb{R}$, $\mathbf{x} = (x, y)$, and $\boldsymbol{\theta} \in \mathbb{R}^p$ the goal is to minimise the sum of squared residuals within the support region, w:

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} \quad \sum_{\mathbf{x} \in w} \big(I(\mathbf{x}) - C(\mathbf{x}, \boldsymbol{\theta})\big)^2.$$

Attempts made to parametrically model the protein spots are most commonly based on a two-dimensional Gaussian spot model [10]. This model has some deficiencies which are sought overcome in more advanced models as the diffusion based model proposed by Bettens [10]. A description of both models will be given and in § 3.4 the models are fitted to a gallery of protein spots of various sizes and shapes.

In the optimisation of the parametric spot models the sum of residuals between the image data and the spot model is minimised within the support region, w. Ideally, the support region should be large enough to contain the entire (or almost all of the) spot. Also, it should not be too large, since this will include neighbouring spots in the support region. As the models can only model a single spot within the region, neighbouring spots should be avoided.

The spots vary greatly in size and without prior knowledge about the spot size it is difficult to determine the size of the support region for the spot in question.

A weighting function (e.g., a Tukey window) could solve some of the difficulties with neighbouring spots within the support region, but the size of such a window still remains unknown.

Also, overlapping spots is not an uncommon phenomenon and in such cases both models must be expected to perform less well. This can possibly be solved by additive mixture models as suggested in § 3.3.3.

In the following the (inverted) sub image (Fig. 3.7) of an electrophoresis gel image will be the source for real data examples. The image is shown with inverse grey levels so that spots appear bright on a dark background. The image contains a total of 61 protein spots of which there is several large, bright spots and quite a few smaller and less distinct spots. Six spots have been selected

to cover the range of strong, intermediate and weak spots, relatively isolated spots, spots with a dense neighbourhood and overlapping spots. The six spots will serve as examples in the following presentation of the two parametric spot models. Fig. 3.8 displays the spot gallery as mesh plots. Each mesh plot is *centred* around the spot in question and the centre of this spot is marked with a small flag (a star on top of a vertical bar). The spot centre locations stem from the protein spot attribute information (§ C.2). It appears that the spot centres are not located exactly at the local maximum of the spot. This is due to the attribute data definition of the spot centre, which is not necessarily at the maximum grey level intensity.

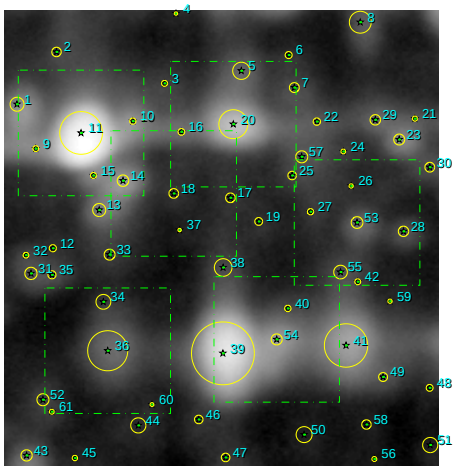


Figure 3.7: Selected spots in 2D gel image. The "ground truth" attribute information (as defined in § C.2) about spot centres and spot areas is marked on the image with stars and circles. The neighbourhoods of selected spots # (11, 18, 20, 36, 53 and 54) are marked with quadratic regions. Selected spots cover the range of strong, intermediate and weak spots, relatively isolated spots, spots with a dense neighbourhood and overlapping spots. Note that spot #54 is the "non-h-dome" from Figs. 3.3 and 3.4.

3.3.1 Gaussian spot model

The two-dimensional anisotropic Gaussian spot model with background level is on the form:

$$C_G(\mathbf{x}, \boldsymbol{\theta}) = B + c\, e^{-\frac{(x-x_0)^2}{2\sigma_x^2}} e^{-\frac{(y-y_0)^2}{2\sigma_y^2}} \tag{3.5}$$

with $\boldsymbol{\theta} = (B, c, x_0, y_0, \sigma_x, \sigma_y)$. B is the background intensity level, i.e., the intensity level in the image around the spot. c is the spot height, (x_0, y_0) are

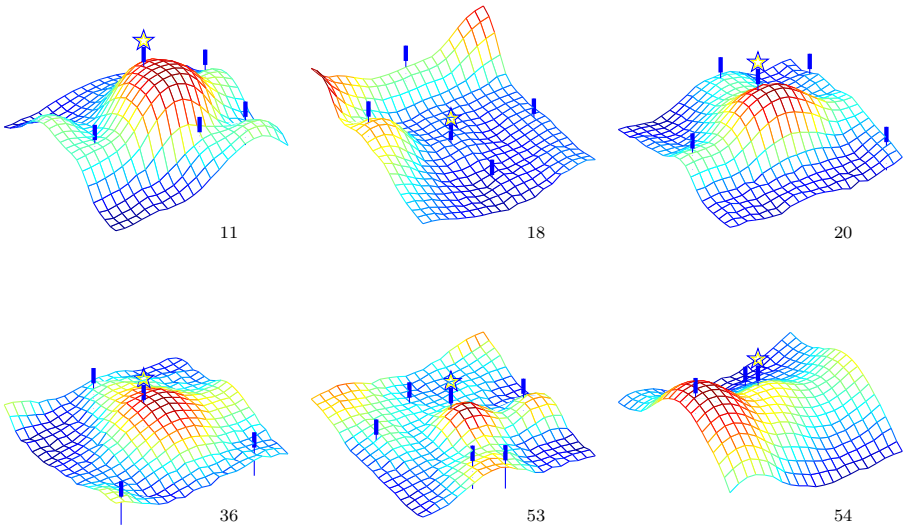


Figure 3.8: Gallery of 6 different protein spots. Top row: spot #11, 18, and 20. Bottom row: spot #36, 53, and 54. Each mesh plot is *centred* around the spot in question and the centre of this spot is marked with a small flag (a star on top of a vertical bar). The spots are viewed a little from above and from the bottom left. Furthermore, if other spots are present in the region, their centres are also marked with a flag (a vertical bar). Please compare to the corresponding quadratic regions around each spot in Fig. 3.7. All meshes are scaled equally so that spot heights and sizes are comparable across plots. Note e.g., how spot #18 is very weak and spot #11 is strong – again in agreement with Fig. 3.7.

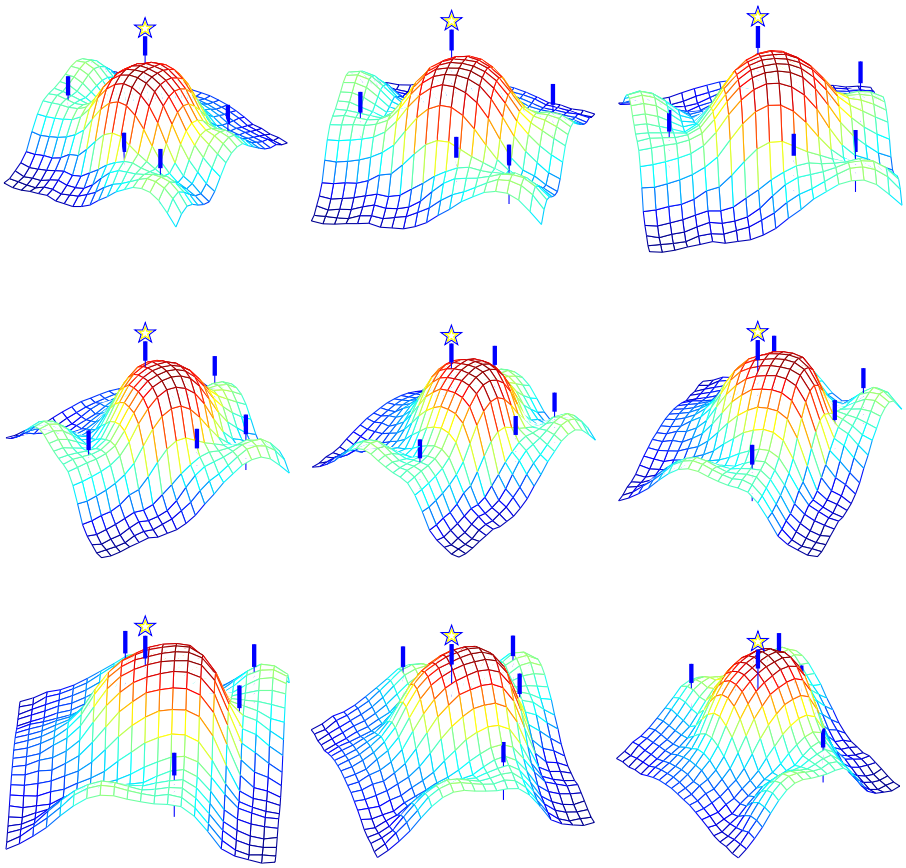


Figure 3.9: Spot 11 from different view points.

offset constants for the centre coordinates, σ_x is the standard deviation in the x-direction and σ_y is the standard deviation in the y-direction. Fig. 3.10 shows the Gaussian parametric model at four different configurations of the parameters. Bettens et al. [10] showed that this model is far too simple to model protein spots and proposed instead a diffusion model that attempts to model the diffusion process of protein migration in the 2D gel. The main drawback of the Gaussian model is its inability to model saturation of the spots. In order to be able to see very low intensity spots the phosphor plate is sometimes exposed for a long period of time, resulting in a saturation of the most intense spots.

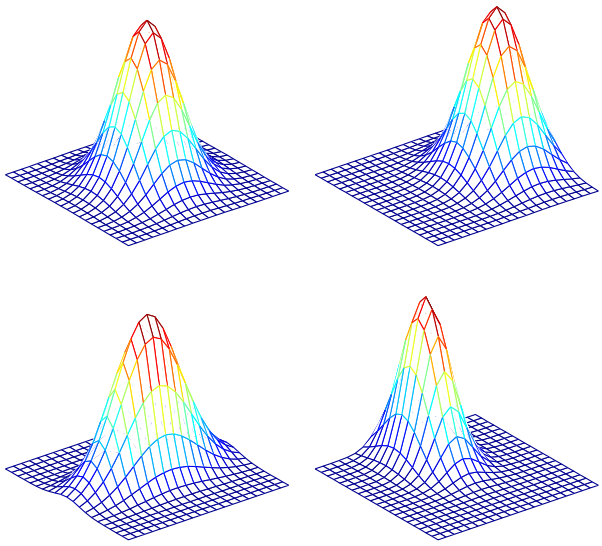


Figure 3.10: 2D anisotropic Gaussian parametric spot model shown at four different configurations of the parameters. Top left: $\boldsymbol{\theta} = (B, c, x_0, y_0, \sigma_x, \sigma_y) = (0, 1, 0, 0, 3, 3)$. Top right: $\boldsymbol{\theta} = (B, c, x_0, y_0, \sigma_x, \sigma_y) = (0, 1, 5, 0, 3, 3)$. Bottom left: $\boldsymbol{\theta} = (B, c, x_0, y_0, \sigma_x, \sigma_y) = (0, 1, 0, 0, 4, 2)$. Bottom right: $\boldsymbol{\theta} = (B, c, x_0, y_0, \sigma_x, \sigma_y) = (0, 1, 0, -5, 2, 3)$.

3.3.2 Diffusion spot model

The diffusion protein spot model proposed by Bettens et al. [10] is designed to model the actual diffusion process in the gel. The following are the assumptions about the process that defines the model: 1) the medium of the diffusion is two-dimensional and anisotropic, i.e., there are two main directions of diffusion (x and y) with different diffusion constants (D_x and D_y), 2) the diffusing substance is initially distributed uniformly on a disc with radius a. Bettens gives the

solution to the corresponding diffusion equation as

$$\begin{aligned} C_D(\mathbf{x},\boldsymbol{\theta}) &= B + \frac{1}{2}C_0\left\{ erf\left(\tfrac{a+r}{2\sqrt{Dt}}\right) + erf\left(\tfrac{a-r}{2\sqrt{Dt}}\right)\right\} \\ &+ \frac{C_0}{r}\sqrt{\frac{Dt}{\pi}}\left(e^{-\frac{(a+r)^2}{4Dt}} - e^{-\frac{(a-r)^2}{4Dt}}\right) \\ \text{with} \qquad r &= \sqrt{D\left(\frac{(x-x_0)^2}{D_x} + \frac{(y-y_0)^2}{D_y}\right)} \end{aligned} \tag{3.6}$$

where $\boldsymbol{\theta} = (B, C_0, x_0, y_0, a, D, D_x, D_y, t)$. B is the background intensity level and C_0 is the initial concentration in the circle. x_0 and y_0 are offset constants for the centre coordinates. *erf* is the error function defined as:

$$erf(z) = \frac{2}{\sqrt{\pi}}\int_0^z e^{-t^2}dt. \tag{3.7}$$

By elimination of symmetric parameters the final model reduces to:

$$\begin{aligned} C_D(\mathbf{x},\boldsymbol{\theta}') &= B + \frac{1}{2}C_0\left\{ erf\left(\tfrac{a'+r'}{2}\right) + erf\left(\tfrac{a'-r'}{2}\right)\right\} \\ &+ \frac{C_0}{r}\sqrt{\frac{1}{\pi}}\left(e^{-\frac{(a'+r')^2}{4}} - e^{-\frac{(a'-r')^2}{4}}\right) \\ \text{with} \qquad r' &= \sqrt{\frac{(x-x_0)^2}{D'_x} + \frac{(y-y_0)^2}{D'_y}}. \end{aligned} \tag{3.8}$$

Following [10] the 7 parameters in (3.8) to be estimated are

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

To give an impression of the model capabilities Figs. 3.11 and 3.12 show the model for a number of different parameter settings. In § 3.4 the two parametric models are tested on the six spots in the spot gallery (Fig. 3.8).

3.3.3 Mixture model

The the parametric spot models described so far do not handle overlapping spots in a well-defined manner. In general, the presence of other spots within

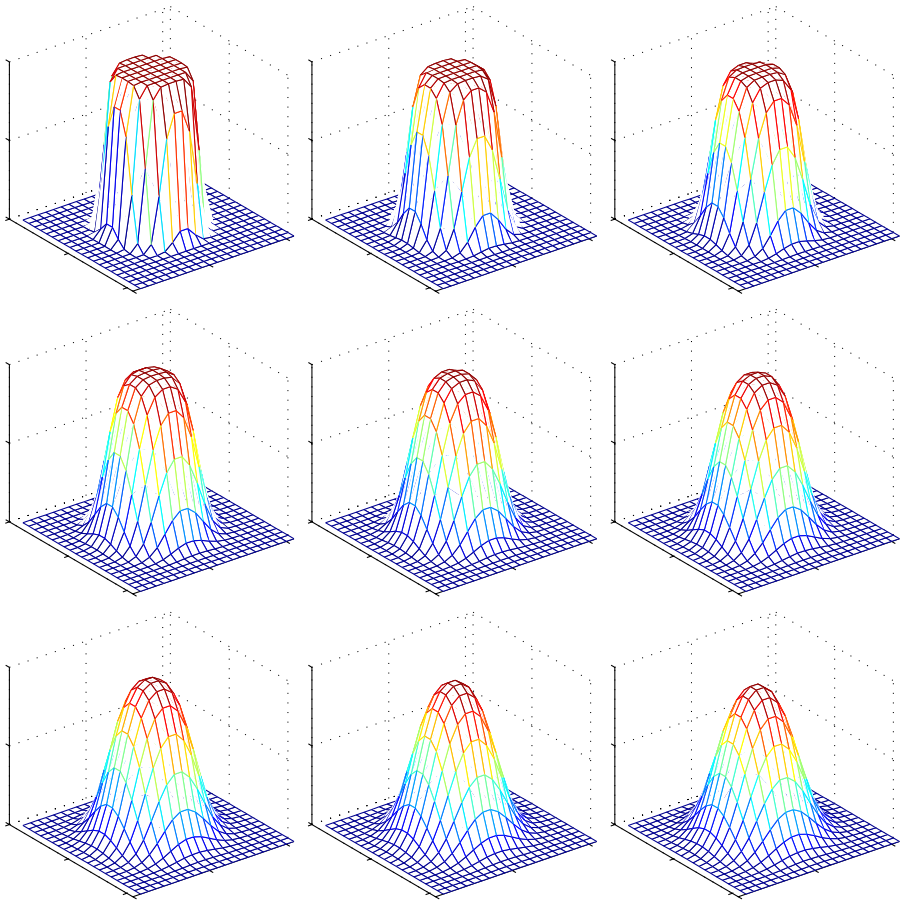


Figure 3.11: 2D diffusion spot model. Increasing t from top left ($t = 0.1, 0.3, 0.5, 0.7, 0.9, 1.1, 1.3, 1.5$), other parameters constant $(B, C_0, a, D, D_x, D_y, x_0, y_0) = (0, 1, 5, 1, 1, 1, 0, 0)$.

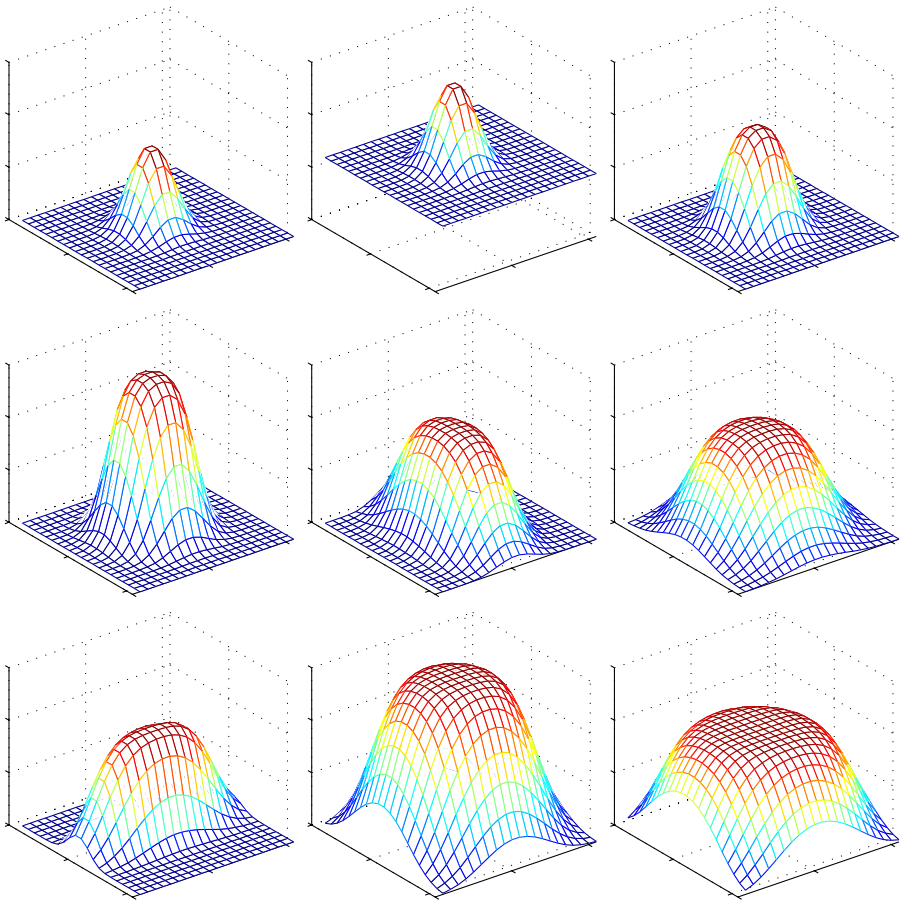


Figure 3.12: 2D diffusion spot model at different parameter configurations. The spot centre offset is constant $(x_0, y_0) = (0, 0)$ as well as time $t = 1$ for all plots. The parameters for each plot are listed in Tab. 3.1. Note how this model is able to model saturated spots.

	B	C_0	a	D	D_x	D_y
1	0.0	1.0	3.0	1.0	1.0	1.0
2	0.6	1.0	3.0	1.0	1.0	1.0
3	0.0	1.0	3.0	2.0	1.0	1.0
4	0.0	1.5	5.0	1.0	1.0	1.0
5	0.0	1.0	3.0	3.0	2.5	1.0
6	0.0	1.0	5.0	1.0	2.5	2.5
7	0.0	1.0	5.0	1.1	0.5	2.5
8	0.0	1.5	5.5	1.1	2.5	2.5
9	0.0	1.0	5.5	1.5	2.5	2.5

Table 3.1: Parameters corresponding to plots in Fig. 3.12. Row 1-3 in the table corresponds to the 3 plots in the top row, row 4-6 in the table corresponds to the 3 plots in the centre row, and row 7-9 corresponds to the 3 plots in the bottom row.

the support region of a spot is not modelled. This may be overcome by simple superposition of diffusion spot models resulting in a mixture model on the form:

$$C_m(\mathbf{x}, \mathbf{\Theta}) = k_1 C_1(\mathbf{x}, \boldsymbol{\theta}_1) + k_2 C_2(\mathbf{x}, \boldsymbol{\theta}_2) + \ldots + k_n C_n(\mathbf{x}, \boldsymbol{\theta}_n), \tag{3.9}$$

where n is the number of diffusion spot models to constitute the mixture model and $\mathbf{\Theta}$ is a $n \times 7$ parameter matrix (7 parameters for each spot).

The sum of squared residuals inside the support region, w

$$\sum_{\mathbf{x} \in w} \big(I(\mathbf{x}) - C_m(\mathbf{x}, \mathbf{\Theta})\big)^2 \tag{3.10}$$

is to be minimised and by including enough spot models, i.e., for sufficiently large n (3.9) can model the data perfectly. Therefore, it is necessary to penalise more complex models (large n) in the optimisation of the model. The goal is to estimate n (the number of spots inside the support region, w) *and* the $(n \times 7)$ parameters in $\mathbf{\Theta}$ so that (3.10) is small while the model is not too complex. The following optimisation problem is proposed:

$$\mathbf{\Theta}^* = \arg\min_{\mathbf{\Theta}} \quad \sum_{\mathbf{x} \in w} \big(I(\mathbf{x}) - C_m(\mathbf{x}, \mathbf{\Theta})\big)^2 + \kappa_1 e^{\kappa_2 n}, \tag{3.11}$$

where the last term serves to penalise large models. κ_1 and κ_2 are constants that must be chosen according to the problem at hand.

3.4 Experiments and Results

This section presents some experiments and results of application of the methods described in the previous sections on a sub image (Fig. 3.7) and the selected spots (Fig. 3.8). First, some results on *scale space blob detection* on the sub image are presented and these will be used in *marker based watershed segmentation*. The same sub image is used in experiments with *h-dome transformation* at different values of h. Finally the *parametric spot models* (2D anisotropic Gaussian and 2D anisotropic diffusion) are fitted to the selected spots in the spot gallery.

3.4.1 Scale space blob detection

For the purpose of scale space blob detection the view of the spots is inverted so that spots appear dark on a bright background. Fig. 3.13 show the same region as Fig. 3.7 with the 61 known protein spot centres overlaid. Figs. 3.14-3.17 show the results of scale space blob detection at scales $t = 1, 2, 3, 4, 5, 6, 7,$ and 8. The number of detected distinct blobs decreases as scale increases (Tab. 3.2). At low scales the image noise dominates the blobs found. With respect to the number of blobs detected, the best result is in the top row of Fig. 3.17 (at $t = 7$) where 57 blobs are found. Compared to Fig. 3.13 most of the spot centres are found even many of the very weak ones. Note however, that even if the number of detected blobs is close to the correct number of proteins, still quite a few spots are not detected. In return other blobs, that are not proteins this method detects as blobs. Some of the very weak spots and spots located close to other spots are detected, only at a very low scale where far too many local minima are found. In other words, there exist minima stronger than the weakest spots and this fact makes the spot detection non-trivial.

scale, t	Number of blobs
1	4313
2	972
3	261
4	103
5	81
6	72
7	57
8	36

Table 3.2: Number of detected blobs at different scales.

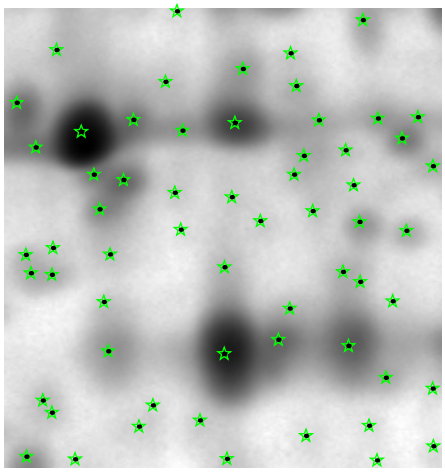


Figure 3.13: Sub region of electrophoresis gel. Spot centres of 61 known protein spots are marked.

3.4.2 Marker based watershed segmentation

In Fig. 3.18 left the attribute (known) spot centres are used as markers for the segmentations. Right shows the segmentation using the scale space blob detection markers ($t = 6$) from experiments in § 3.4.1.

3.4.3 H-dome transformation

This section shows experiments with the *h-domes* transformation (§ 3.1.2) on an electrophoresis gel image. Again the (inverted) gel image is used and Fig. 3.19 shows the h-domes for different values of h.

3.4.4 Parametric spot models

The two parametric protein spot models described in § 3.3 both have been optimised to fit the 6 spots in the spot gallery (see Fig. 3.8). The optimal fits were obtained using the non-linear least squares method *lsqnonlin* in the Matlab® Optimization Toolbox (The MathWorks, Inc.). Figs. 3.20 - 3.31 show the results of fitting the Gaussian model and the diffusion model to the six spots in the gallery. Every second figure shows the Gaussian model and every other second the diffusion model.

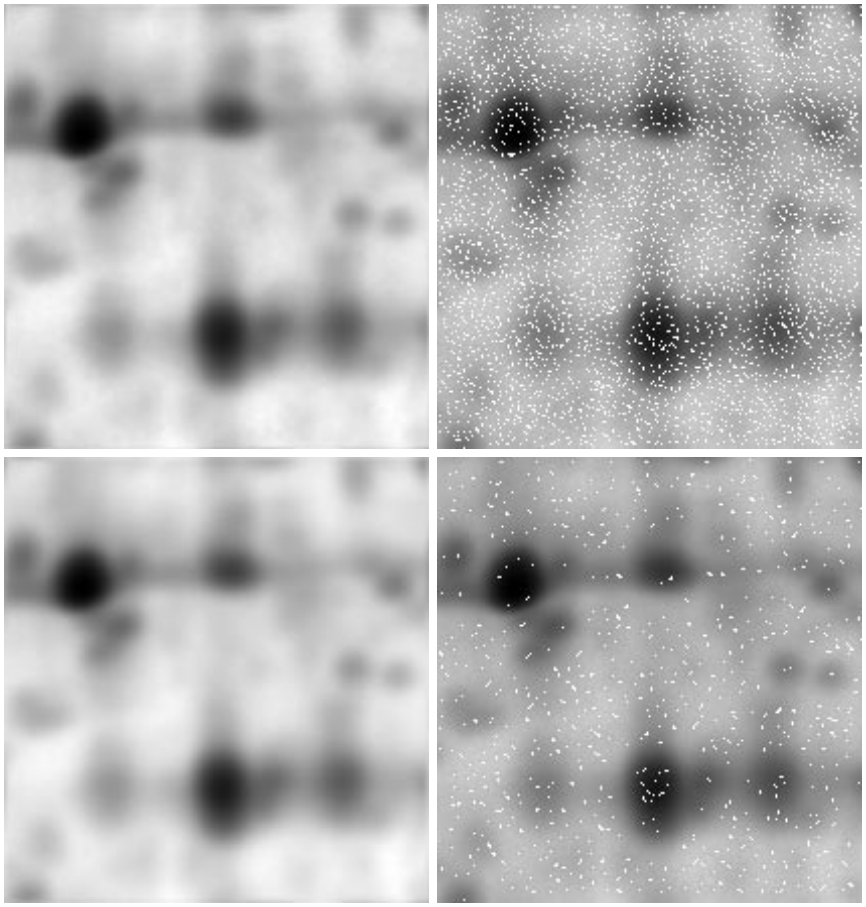


Figure 3.14: Scale space blob detection at different scales. Left column: Gaussian filtered versions of the original image (Fig. 3.13). Right column: original image with blob markers. Top: $t = 1$, 4313 distinct blobs were found. Bottom: $t = 2$, 972 blobs. The correct number of protein spots in this part of the gel is 61.

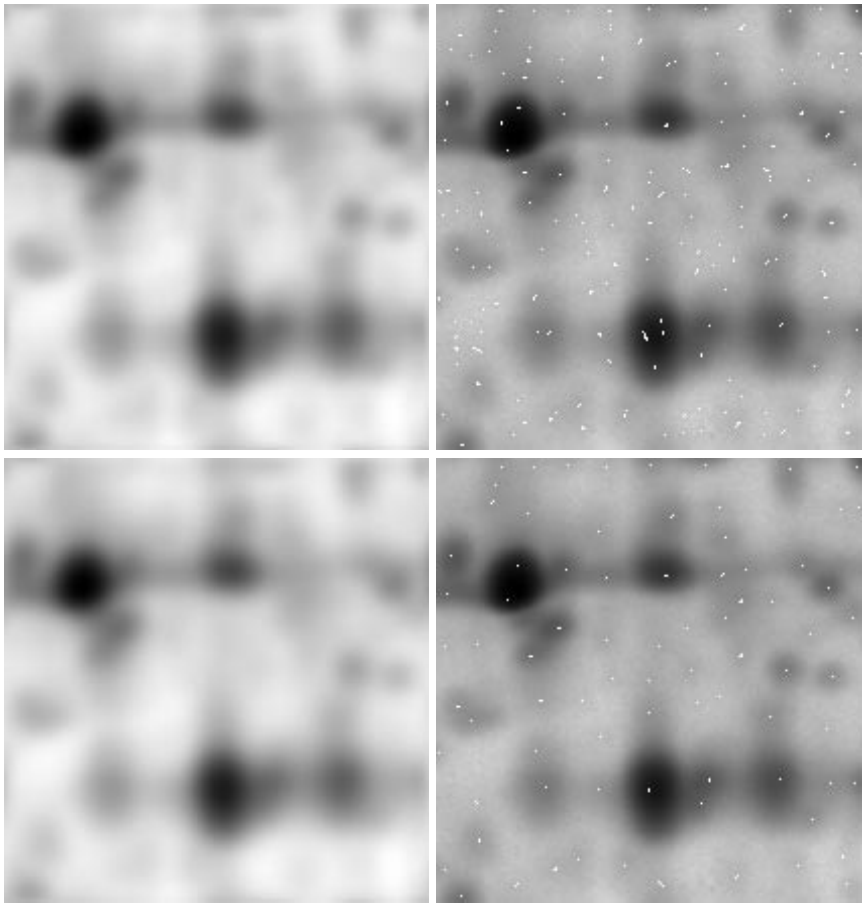


Figure 3.15: Scale space blob detection at different scales. Left column: Gaussian filtered versions of the original image (Fig. 3.13). Right column: original image with blob markers. Top: $t = 3$, 261 distinct blobs were found. Bottom: $t = 4$, 103 blobs. The correct number of protein spots in this part of the gel is 61.

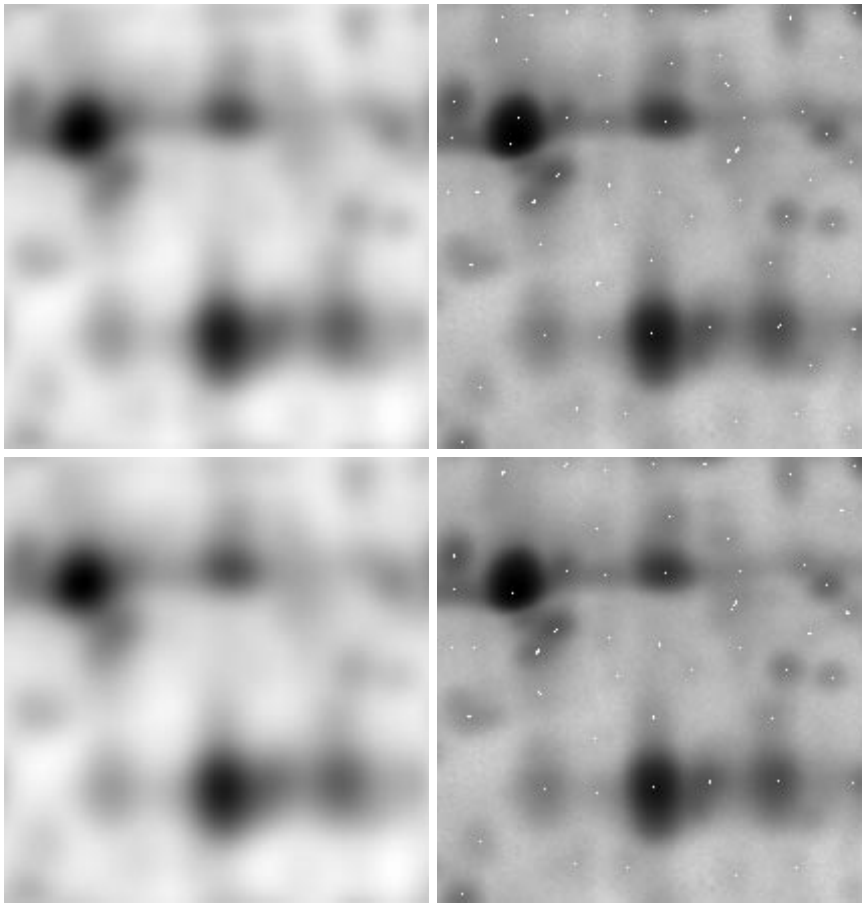


Figure 3.16: Scale space blob detection at different scales. Left column: Gaussian filtered versions of the original image (Fig. 3.13). Right column: original image with blob markers. Top: $t = 5$, 81 distinct blobs were found. Bottom: $t = 6$, 72 blobs. The correct number of protein spots in this part of the gel is 61.

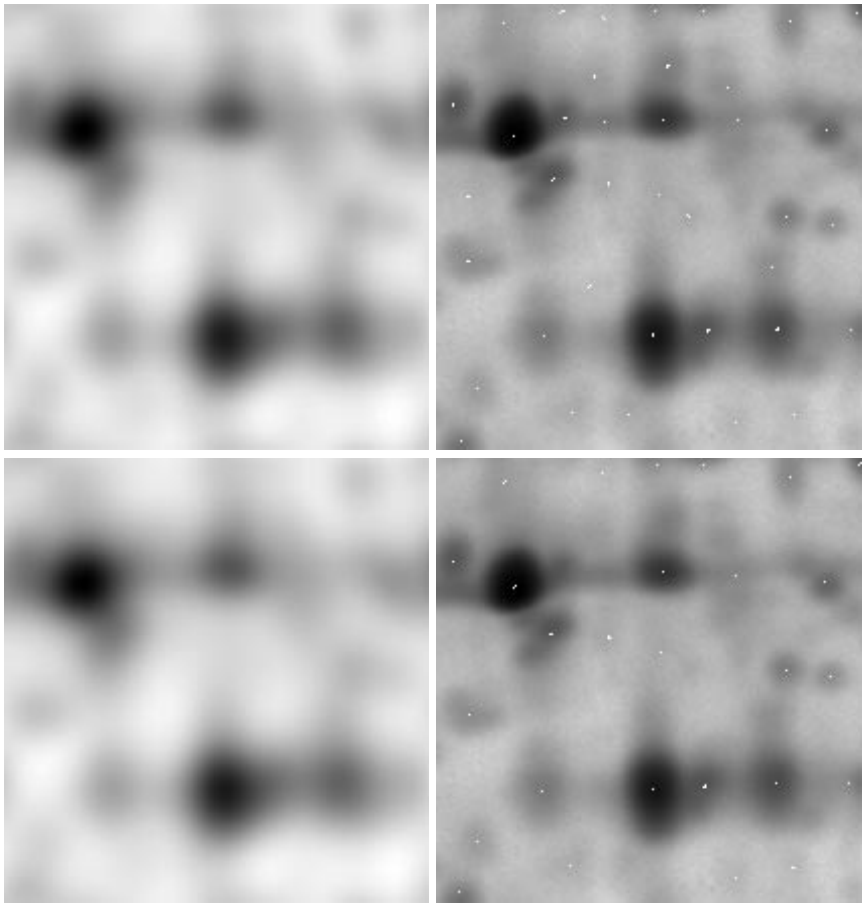


Figure 3.17: Scale space blob detection at different scales. Left column: Gaussian filtered versions of the original image (Fig. 3.13). Right column: original image with blob markers. Top $t = 7$, 57 distinct blobs. Bottom $t = 8$, 36 blobs. The correct number of protein spots in this part of the gel is 61.

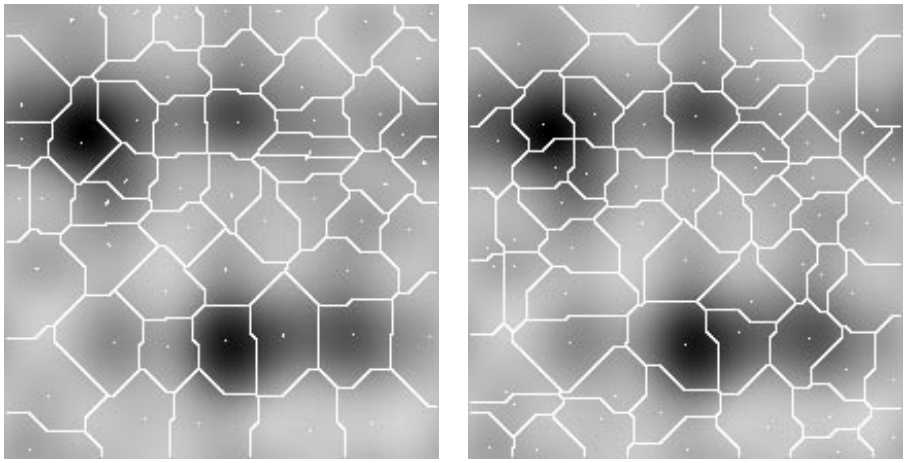


Figure 3.18: Watershed segmentation using different markers. Left: markers are the spot centres from the "ground truth" attribute information (see § C.2). Right: markers are the results of scale space blob detection at $t = 6$.

The gallery figures (3.7 and 3.8) show the spot centres of the neighbour spots within the square support region. The support region size has been chosen to be the same (61×61 pixels) for all spots in the gallery.

The evaluation of the fit of a model to the spot data can be done in different ways. One is by some quantitative measure, such as the sum of squared differences inside the support region. Another way of evaluation is by qualitative visual inspection. However, defining a good quantitative measure is not trivial. The reason is, that an ideal (single spot) model fits nicely to the spot and ignores eventual surrounding neighbours inside the support region and therefore the residual may (correctly) be large in areas of the support region where neighbours are present. This problem is closely connected to the size of the support region. Clearly there is a need for such a quantitative evaluation, but the following evaluation will be confined to the qualitative visual inspection.

Each of the figures consists of five components. In the top is seen the square region around the *known* spot centre displayed as a grey level image. To the left in the second row is the same information shown as a wire-frame surface. To the right in the second row is shown the resulting optimised model also as a wire-frame surface. The bottom row displays the residual (pixel-wise difference between data and model) in different ways. To the left, the residual is shown as a wire-frame surface together with the zero-plane as reference. To the right, the spot data is shown as a wire-frame surface. Together with the data, the model is displayed as a coloured solid surface. The colouring represents the residuals

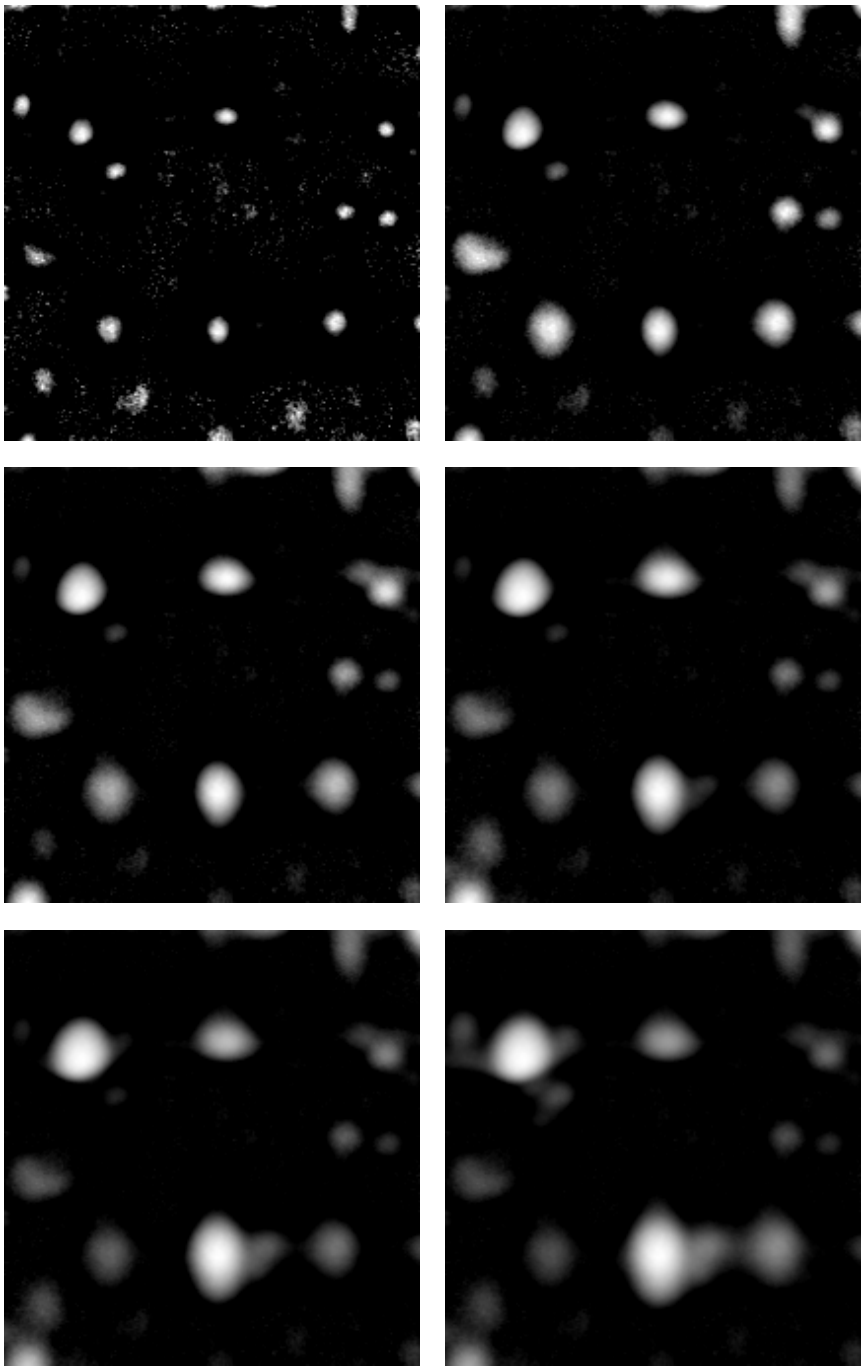


Figure 3.19: H-dome transformation of 2D gel image. From the top left: $h = 0.05, 0.15, 0.25, 0.35, 0.45$, and 0.50.

(signed). Note the colour at intersections between the data and the model (zero error).

- Spot 11 (figs. 3.20 and 3.21). This spot is relatively large and intense. Within the shown region it has 4 neighbouring spots that are all weaker than spot 11. Some of the neighbours are overlapping. From the wire-frame surface of the data (second row, left) it is seen, that the spot appears to be "flat" on the top, which indicates that the spot is *saturated*. Note how the diffusion model is much better at capturing this behaviour.
- Spot 18 (figs. 3.22 and 3.23). The spot is so small and weak, that it almost vanishes. Furthermore it has 4 (all stronger) neighbouring spots. None of the models are able to detect the spot and they both suggest a flat surface. The restrictions of x_0 and y_0 prevent the models from moving to one of the stronger neighbours.
- Spot 20 (figs. 3.24 and 3.25). This spot is of medium size and intensity. It has four neighbours, two strong and two weak. The spot is oblong in the x-direction, which may be due to overlapping neighbours. The Gaussian model seems to be more influenced by this than the diffusion model.
- Spot 36 (figs. 3.26 and 3.27) is a relatively large spot of medium intensity. The spot is relatively isolated, i.e., neighbours are few, small and far away. Both models seem to do well in this case.
- Spot 53 (figs. 3.28 and 3.29). This spot of medium strength is small and well separated from neighbouring spots. Both models fit this spot well.
- Spot 54 (figs. 3.30 and 3.31) is of medium size and intensity but it has a relatively large and intense neighbour spot to the left. The two spots have a severe overlap and both methods experience difficulty modelling the spot correctly. Another strong neighbour to the right also overlap, although it is not in the support region. Both models try to model all three spots as one big spot.

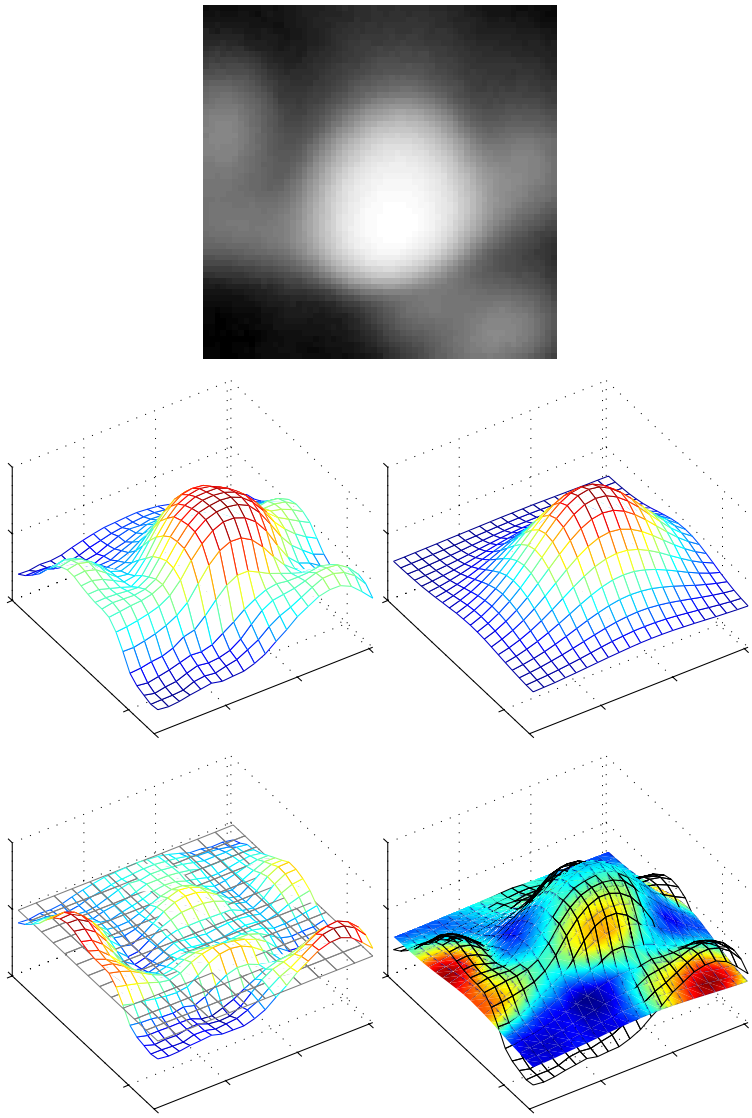


Figure 3.20: Parametric fit of Gaussian spot model to spot 11. **Top**: Spot in question shown as grey-scale image. **Second row:** *left* – spot data in question shown as wire-frame, *right* – fitted model. **Bottom row:** *left* – residual shown with zero-level plane, *right* – spot data in question shown as wire-frame and the fitted model shown as a solid surface, coloured according to the residual.

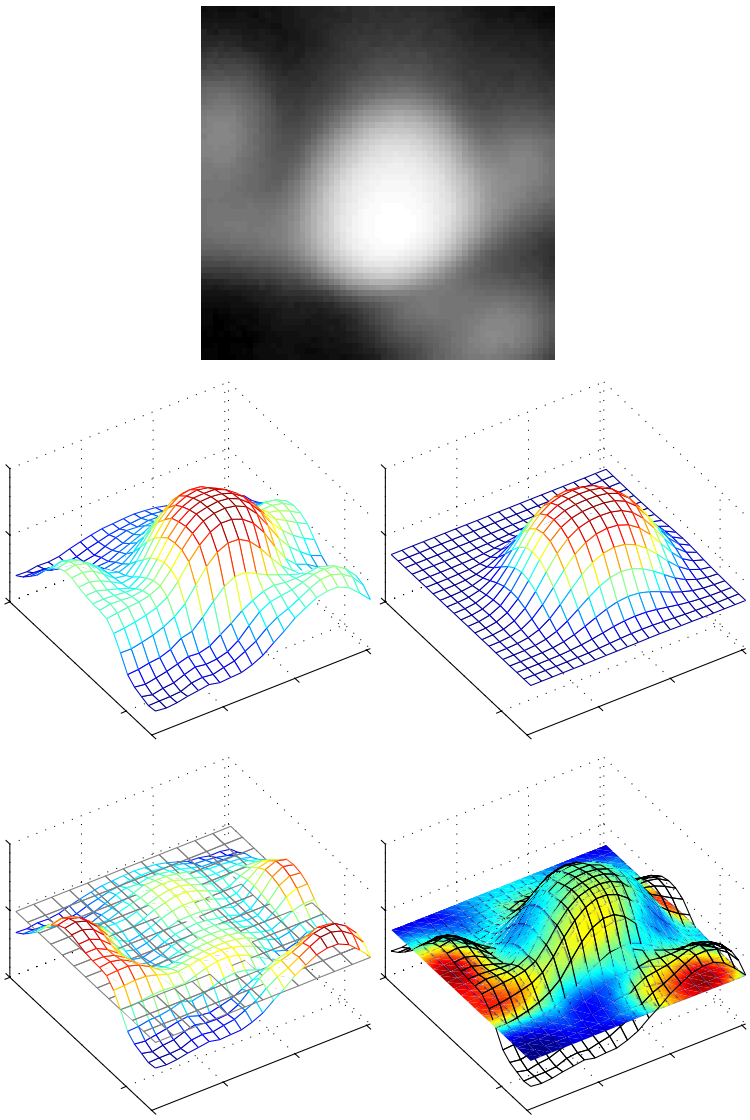


Figure 3.21: Parametric fit of diffusion spot model to spot 11. **Top**: Spot in question shown as grey-scale image. **Second row:** *left* – spot data in question shown as wire-frame, *right* – fitted model. **Bottom row:** *left* – residual shown with zero-level plane, *right* – spot data in question shown as wire-frame and the fitted model shown as a solid surface, coloured according to the residual.

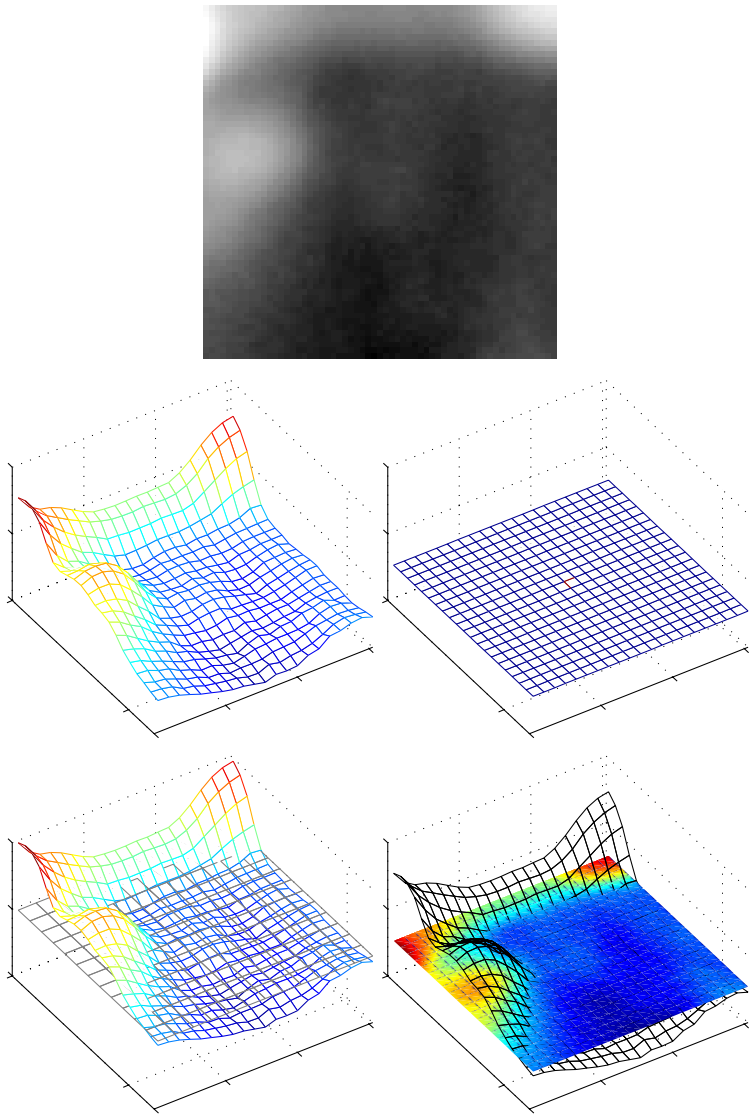


Figure 3.22: Parametric fit of Gaussian spot model to spot 18. **Top**: Spot in question shown as grey-scale image. **Second row:** *left* – spot data in question shown as wire-frame, *right* – fitted model. **Bottom row:** *left* – residual shown with zero-level plane, *right* – spot data in question shown as wire-frame and the fitted model shown as a solid surface, coloured according to the residual.

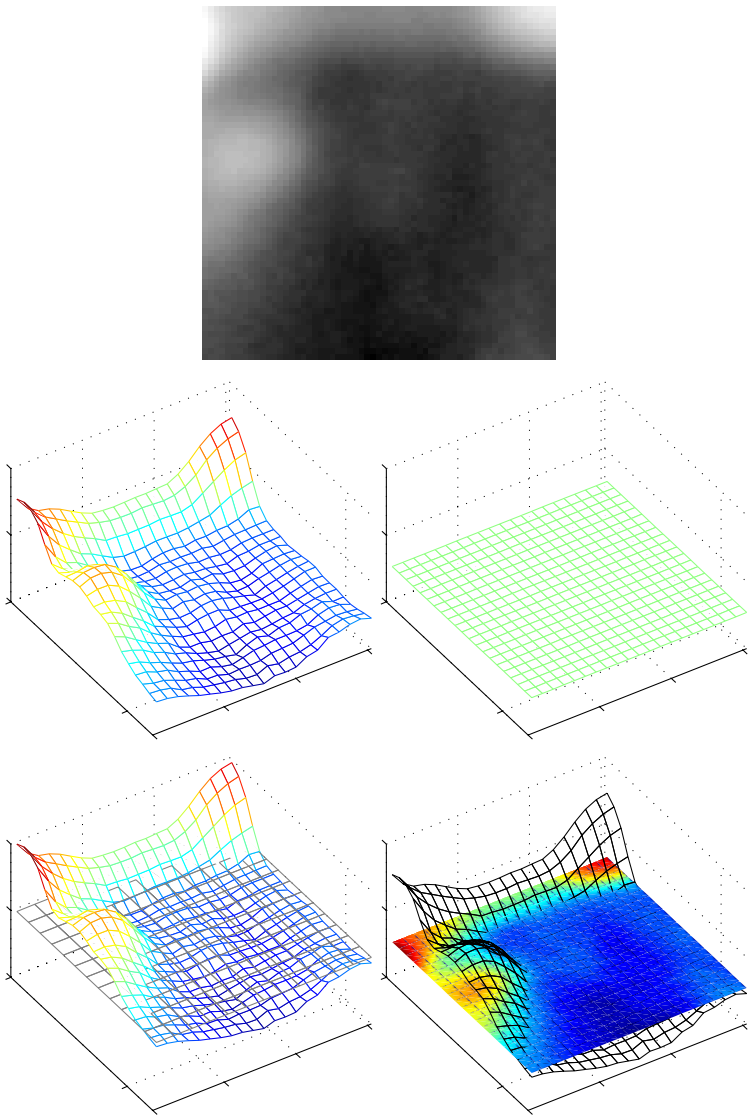


Figure 3.23: Parametric fit of diffusion spot model to spot 18. **Top**: Spot in question shown as grey-scale image. **Second row:** *left* – spot data in question shown as wire-frame, *right* – fitted model. **Bottom row:** *left* – residual shown with zero-level plane, *right* – spot data in question shown as wire-frame and the fitted model shown as a solid surface, coloured according to the residual.

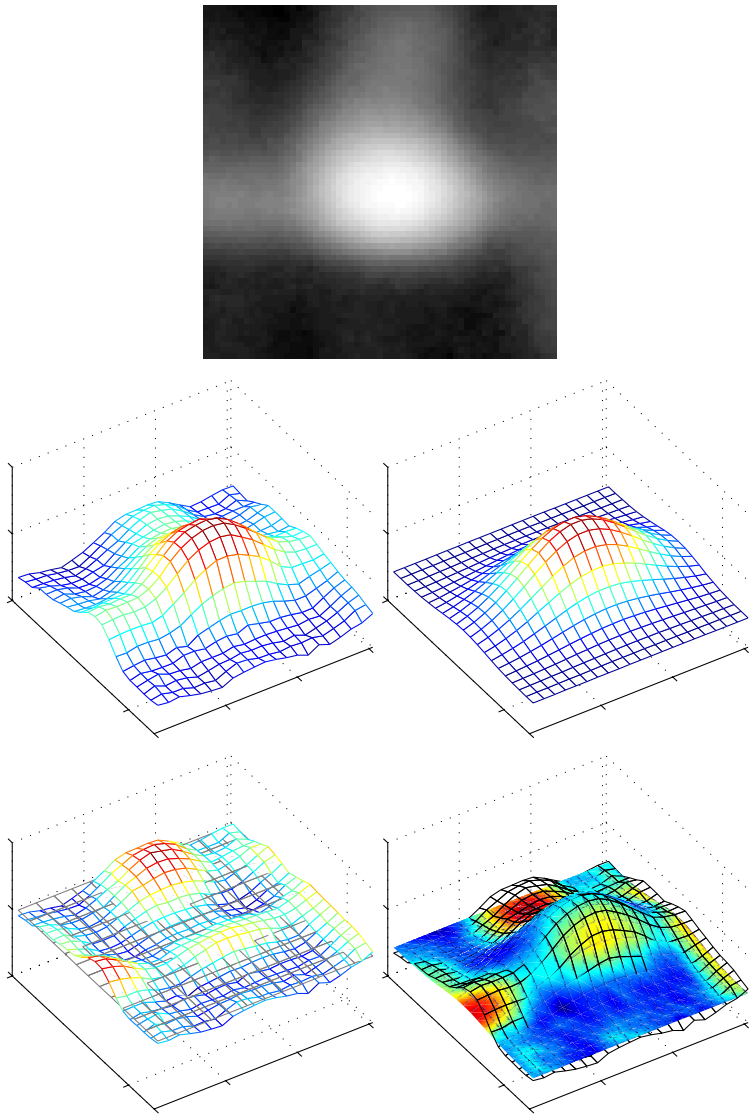


Figure 3.24: Parametric fit of Gaussian spot model to spot 20. **Top**: Spot in question shown as grey-scale image. **Second row:** *left* – spot data in question shown as wire-frame, *right* – fitted model. **Bottom row:** *left* – residual shown with zero-level plane, *right* – spot data in question shown as wire-frame and the fitted model shown as a solid surface, coloured according to the residual.

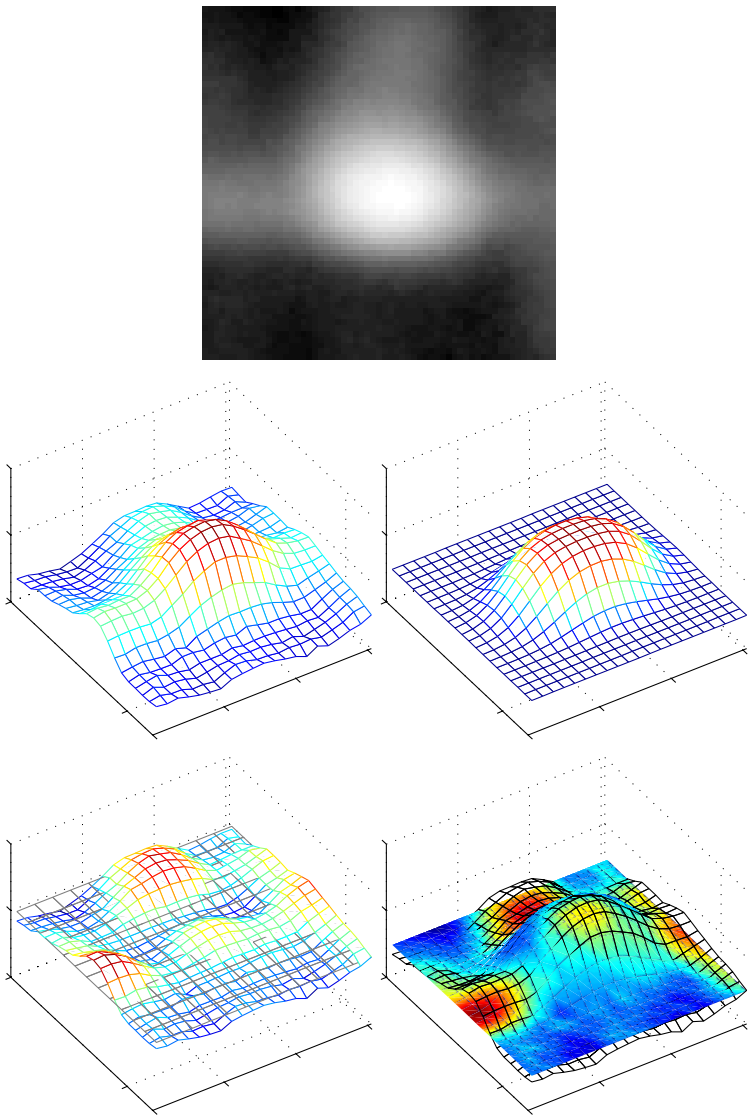


Figure 3.25: Parametric fit of diffusion spot model to spot 20. **Top**: Spot in question shown as grey-scale image. **Second row:** *left* – spot data in question shown as wire-frame, *right* – fitted model. **Bottom row:** *left* – residual shown with zero-level plane, *right* – spot data in question shown as wire-frame and the fitted model shown as a solid surface, coloured according to the residual.

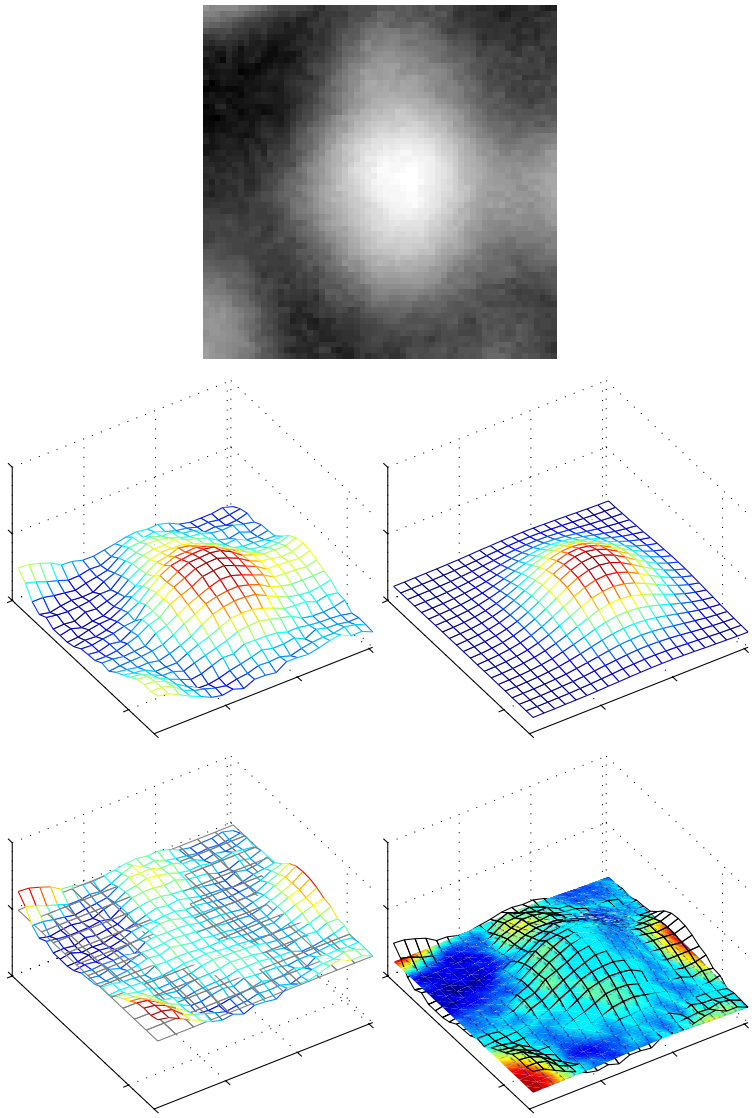


Figure 3.26: Parametric fit of Gaussian spot model to spot 36. **Top**: Spot in question shown as grey-scale image. **Second row:** *left* – spot data in question shown as wire-frame, *right* – fitted model. **Bottom row:** *left* – residual shown with zero-level plane, *right* – spot data in question shown as wire-frame and the fitted model shown as a solid surface, coloured according to the residual.

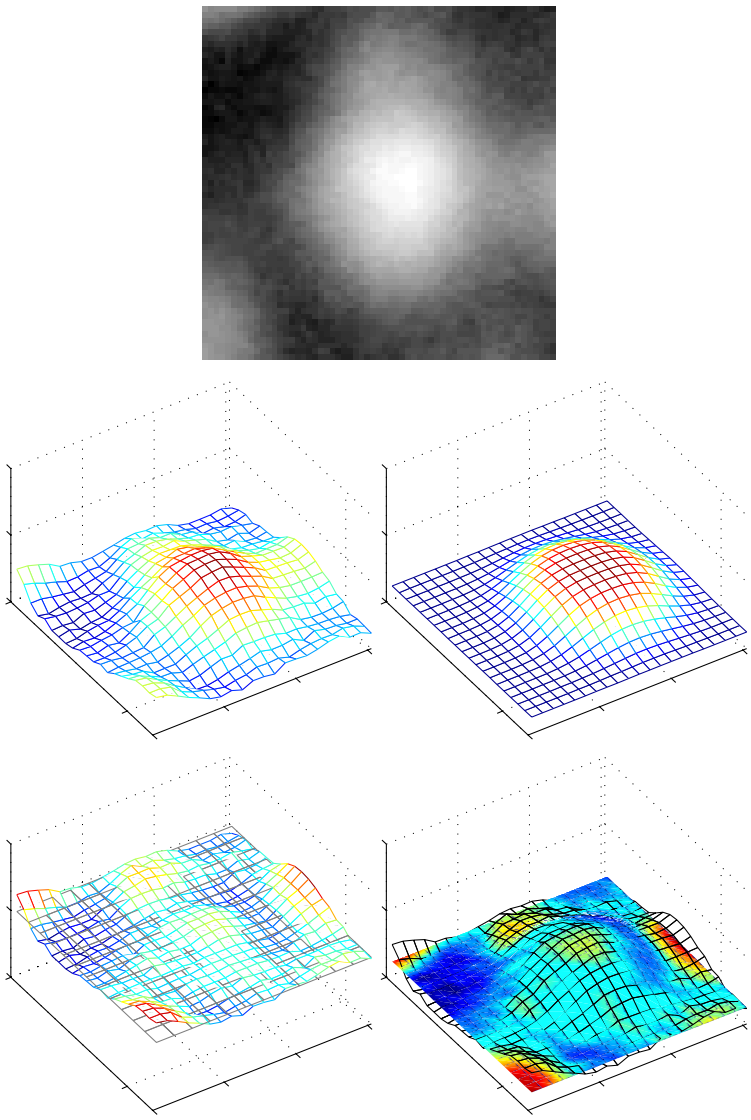


Figure 3.27: Parametric fit of diffusion spot model to spot 36. **Top**: Spot in question shown as grey-scale image. **Second row:** *left* – spot data in question shown as wire-frame, *right* – fitted model. **Bottom row:** *left* – residual shown with zero-level plane, *right* – spot data in question shown as wire-frame and the fitted model shown as a solid surface, coloured according to the residual.

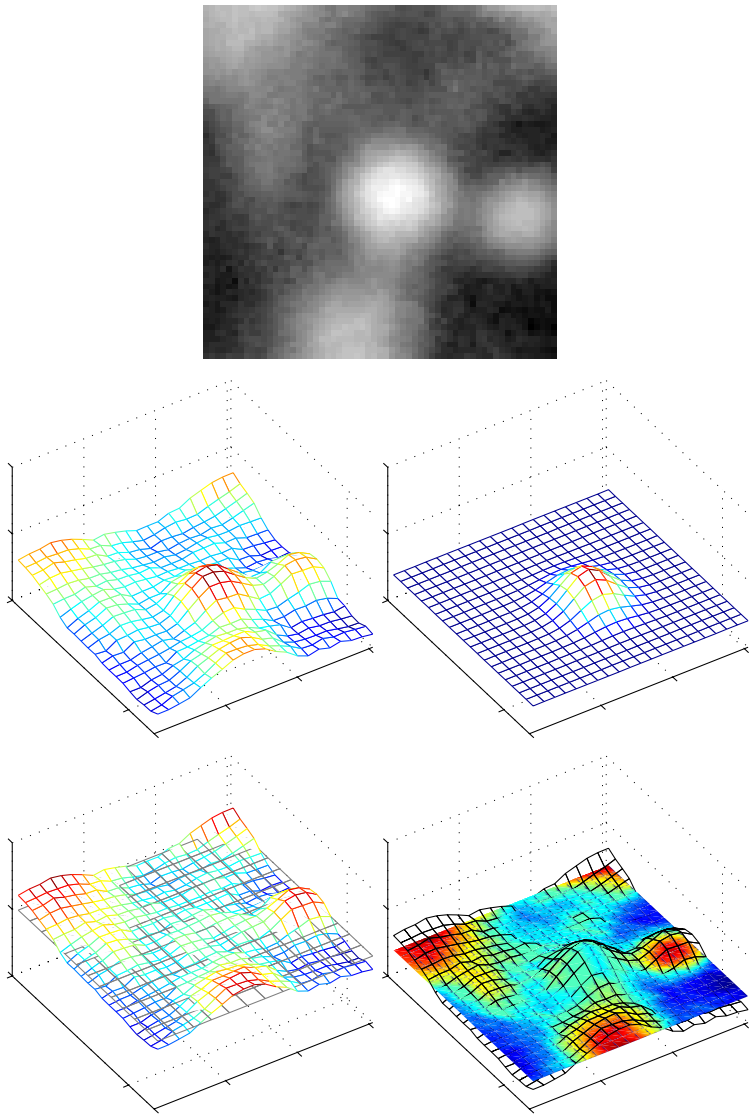


Figure 3.28: Parametric fit of Gaussian spot model to spot 53. **Top**: Spot in question shown as grey-scale image. **Second row:** *left* – spot data in question shown as wire-frame, *right* – fitted model. **Bottom row:** *left* – residual shown with zero-level plane, *right* – spot data in question shown as wire-frame and the fitted model shown as a solid surface, coloured according to the residual.

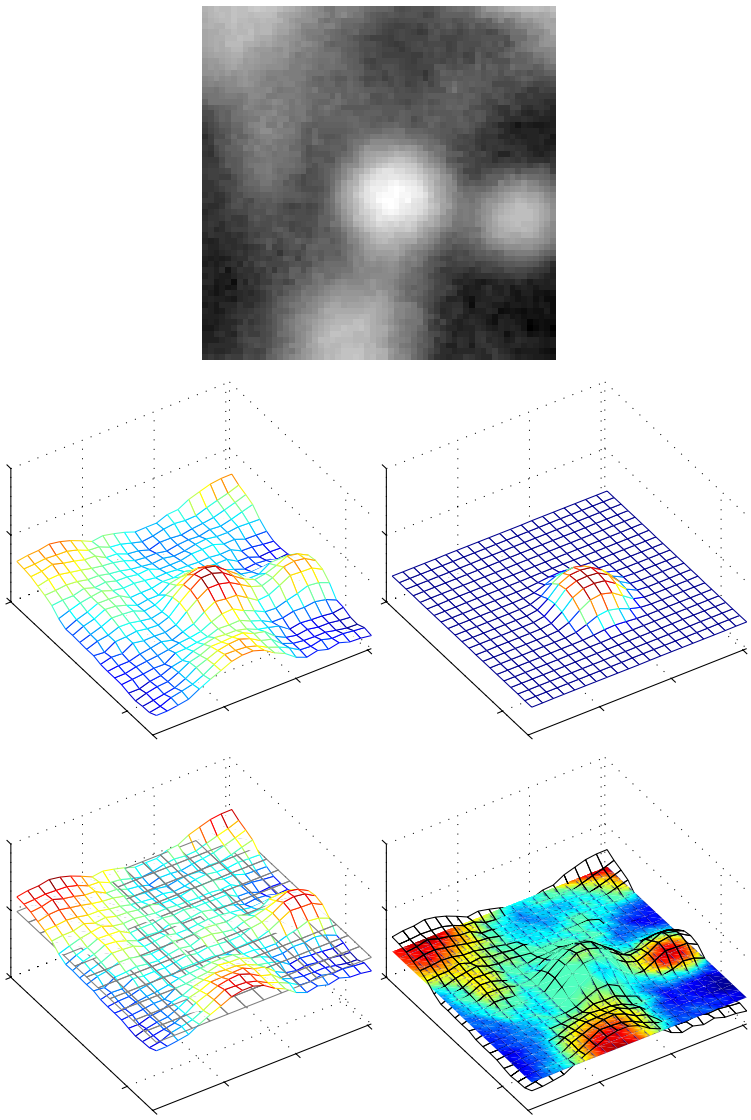


Figure 3.29: Parametric fit of diffusion spot model to spot 53. **Top**: Spot in question shown as grey-scale image. **Second row:** *left* – spot data in question shown as wire-frame, *right* – fitted model. **Bottom row:** *left* – residual shown with zero-level plane, *right* – spot data in question shown as wire-frame and the fitted model shown as a solid surface, coloured according to the residual.

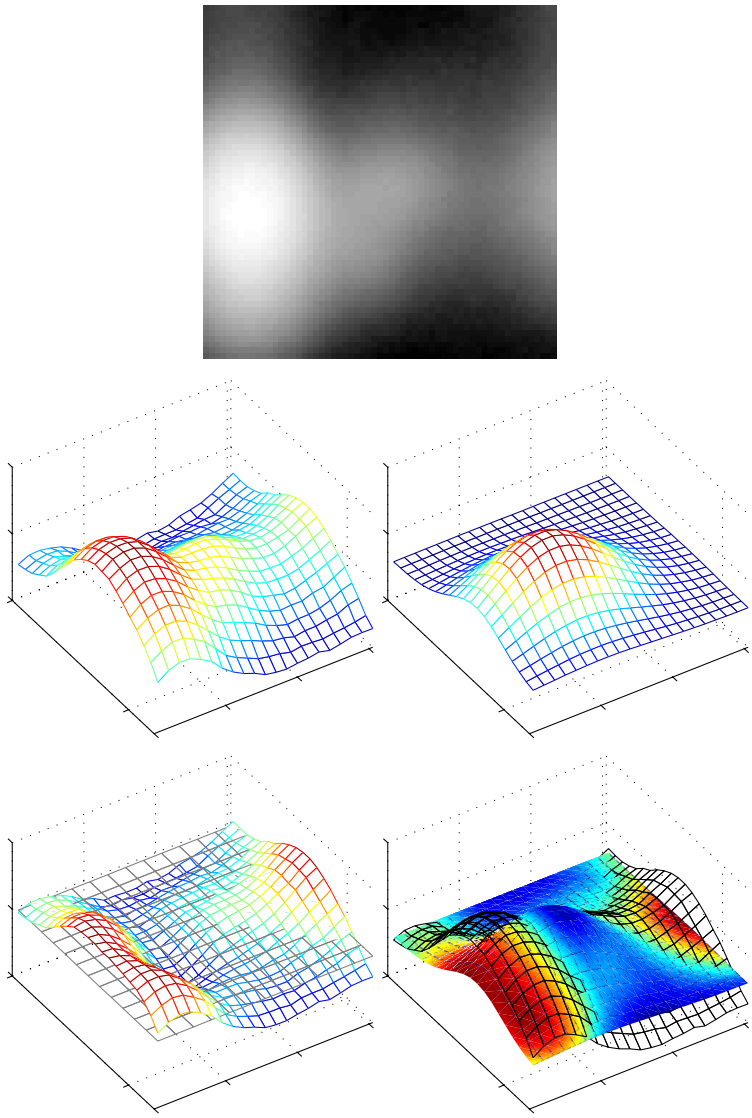


Figure 3.30: Parametric fit of Gaussian spot model to spot 54. **Top**: Spot in question shown as grey-scale image. **Second row:** *left* – spot data in question shown as wire-frame, *right* – fitted model. **Bottom row:** *left* – residual shown with zero-level plane, *right* – spot data in question shown as wire-frame and the fitted model shown as a solid surface, coloured according to the residual.

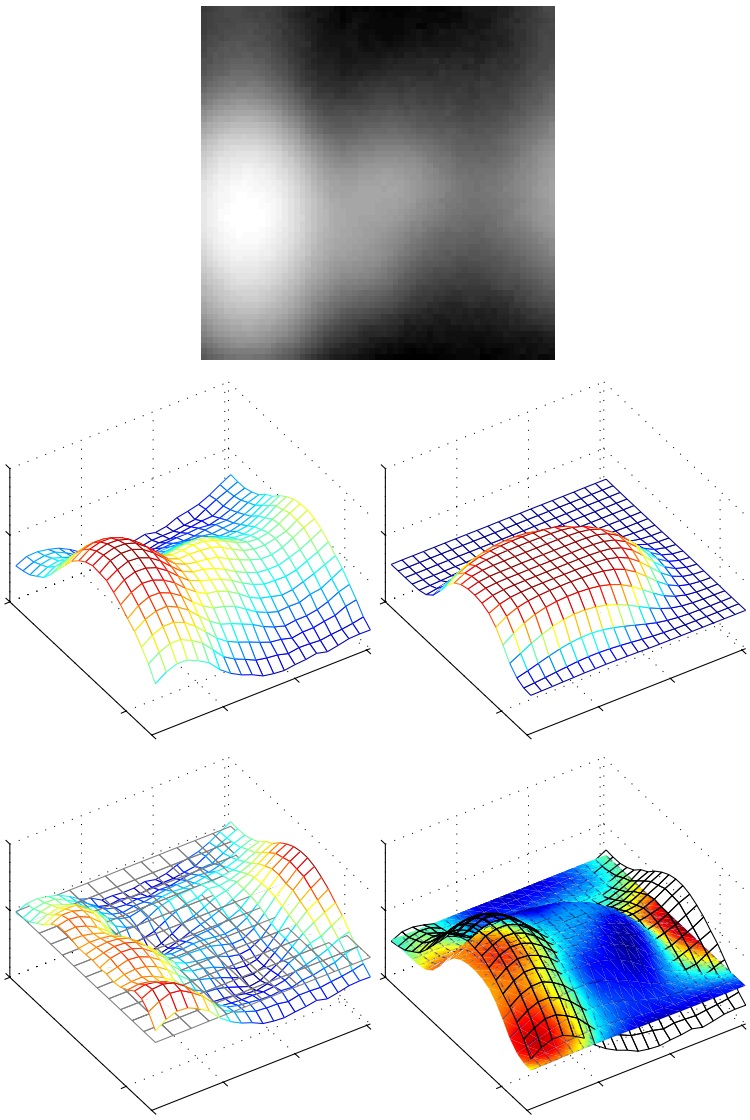


Figure 3.31: Parametric fit of diffusion spot model to spot 54. **Top**: Spot in question shown as grey-scale image. **Second row:** *left* – spot data in question shown as wire-frame, *right* – fitted model. **Bottom row:** *left* – residual shown with zero-level plane, *right* – spot data in question shown as wire-frame and the fitted model shown as a solid surface, coloured according to the residual.

3.5 Summary

Although not the primary focus of this work, various approaches to segmentation of two-dimensional electrophoresis gel images into protein spots and background have been presented. This is motivated by the fact that a good segmentation is fundamental to the point matching methods described in § 4.

Summarising the methods described:

- Mathematical morphology based methods.
  - The watershed by immersion method and marker based watersheds are presented and applied to a nanoparticle image.
  - H-domes method is explained and applied to a two-dimensional electrophoresis image.
- Scale space based blob detection. Lindeberg's scale space blob detector is described and applied to the nanoparticle image.
- Parametric protein spot models. Two models, a Gaussian and a diffusion based model are presented and their properties explored.
- The methods presented are all tested on real gel image data.

The watershed method is known to result in over-segmentation if not restricted by markers. Together with the scale space blob detector blobs can be segmented, but the choice of scale remains non-trivial. It has been demonstrated that even if the scale is chosen so that the *number* of detected blobs is close to the correct number of proteins, still quite a few spots are not detected and in return other blobs, that are not proteins are detect as blobs. This is probably due to the noisy nature of the images. Some of the very weak spots and spots located close to other spots are detected, only at a very low scale where far too many local minima are found. In other words, there exist minima in the image that are stronger than the weakest (known) spot. This fact makes the spot detection based on the image data alone challenging.

The parametric spot models seem to have most success at modelling relatively isolated spots. As expected, the diffusion model is better at modelling highly intense (saturated) protein spots, where for smaller spots the two models perform almost equally well.

The main deficiency of the spot models is, that they are not able to model *multiple* and *overlapping* spots within a small area. The choice of support region size also proved difficult without prior knowledge of the spot size.

A mixture model was proposed as a solution to the problems of the parametric models but it was not implemented.

Chapter 4

Point Pattern Matching

This chapter describes a wide range of point matching methods for general point pattern matching as well as a number of methods applied to spot matching in electrophoresis gel images. The motivation for investigating point pattern matching methods is its application in proteome analysis. Since the process of matching the protein spots is a problem of solving a correspondence problem, the field of point pattern matching methods seems appropriate to investigate. This chapter is divided into three main parts. First, *general methods* for point pattern matching are described, later, methods developed *especially* for the matching of point patterns from *electrophoresis gel images* are investigated, and finally a *comparison* of selected methods tested on real point patterns from electrophoresis gel data is presented.

Point pattern matching is a fundamental step in many tasks in the fields of image registration and warping [34, 38, 45], motion estimation, matching of shapes [6], shape analysis [15, 26, 28], structure from motion estimation [39, 80], and many more. Other terms to describe the process of point pattern matching are to solve the *correspondence* or solving the *optimal assignment* problem.

The area of applications for point matching includes: biology, medicine, astrophysics, robotics, and computer vision, to mention a few.

In the application of point pattern matching on protein spot centres there is a number of requirements for the matching method used (as identified in § 2.3). The method must:

- exactly and robustly match protein pairs,
- allow for non-linear distortions/transformations,
- robustly handle outliers in both sets,
- be able to handle point sets of stochastic/amorphous nature, and
- robustly match dense point sets.

4.1 Notation

Notions of point patterns, point correspondence, match matrices and motion are introduced.

Definition 1 (Point pattern) Define a two-dimensional point pattern $\mathcal{P} = \{p_i \,|\, i = 1, 2, \ldots, J\}$, where $p_i = (x_i, y_i)$ are the coordinates of the point in the x-y plane. The pattern, also referred to as a *point set* is said to have *cardinality* J. □

4.1.1 Correspondence and match matrices

The comparison of two point patterns necessitates a way to define the *correspondence* between homologous points in the two patterns:

Definition 2 (Point correspondence) Given two point sets $\mathcal{P}$ and $\mathcal{Q}$, two homologous points p_j and q_k are said to correspond:

$$p_j \sim q_k, \tag{4.1}$$

i.e., the *j'th* point in $\mathcal{P}$ corresponds to the *k'th* point in $\mathcal{Q}$. The symbol $\sim$ is normally used to describe directly similar figures, i.e., when all corresponding angles are equal and described in the same rotational sense. □

For the entire both patterns $\mathcal{P}$ and $\mathcal{Q}$ the correspondence between their points can be described in the *binary match matrix*:

Definition 3 (Binary match matrix) Consider two point patterns $\mathcal{P}$ and $\mathcal{Q}$ of equal cardinalities J. The $J \times J$ binary *match matrix* m defines the correspondence between homologous points in the two point patterns:

$$\forall j \leq J, \forall k \leq J \quad m_{jk} = \begin{cases} 1 & \text{if point } p_j \sim q_k \\ 0 & \text{otherwise} \end{cases} \tag{4.2}$$

subject to the constraint that there is a one-to-one mapping from $\mathcal{P}$ to $\mathcal{Q}$:

$$\forall j \leq J \quad \sum_{k=1}^{J} m_{jk} = 1 \tag{4.3}$$

and

$$\forall k \leq J \quad \sum_{j=1}^{J} m_{jk} = 1. \tag{4.4}$$

In other words, the points in both sets correspond to *exactly* one other point in the other set, i.e., all rows and all columns of m must sum to one. In this definition, it is not necessary to allow for missing points or for multiple-to-one correspondences. □

In many applications, points present in one set may be missing in the other and vice versa. The cardinalities may be intrinsically different or points may be missing. In such cases the one-to-one correspondence does not hold and other ways of describing the correspondence between homologous points must be defined. Denote points that do not have a corresponding partner point in the other pattern, *singles*. Singles may be due to noise or that the two point sets may be only partially overlapping.

Definition 4 (Single Point) Consider two point patterns $\mathcal{P}$ and $\mathcal{Q}$. A *single point* is a point in a point pattern that does not have a homologous point in the other pattern. This type of points is often also referred to as singles, outliers, contaminating points or dropouts. □

In order to represent correspondences and singles in the same data structure the *augmented* binary match matrix, $\hat{m}$ is defined.

Definition 5 (Augmented binary match matrix) Consider now two point patterns $\mathcal{P}$ and $\mathcal{Q}$ of possibly different cardinality J and K. The augmented match matrix $\hat{m}$ is a $(J+1) \times (K+1)$ matrix defining the state of each point in two point patterns $\mathcal{P}$ and $\mathcal{Q}$. $\hat{m}$ is similar to m in Def. 3 except that $\hat{m}$ has an *extra row* and an *extra column* to hold the information of single points. Again,

$$\forall j \leq J, \forall k \leq K \quad \hat{m}_{jk} = \begin{cases} 1 & \text{if point } p_j \sim q_k \\ 0 & \text{otherwise} \end{cases} \tag{4.5}$$

and for the last column (singles in $\mathcal{P}$)

$$\forall j \leq J \quad \hat{m}_{j(K+1)} = \begin{cases} 1 & \text{if point } p_j \text{ is a single} \\ 0 & \text{otherwise} \end{cases} \tag{4.6}$$

and similarly for the last row (singles in $\mathcal{Q}$)

$$\forall k \leq K \quad \hat{m}_{(J+1)k} = \begin{cases} 1 & \text{if point } q_k \text{ is a single} \\ 0 & \text{otherwise} \end{cases} \tag{4.7}$$

subject to the constraint that *either* a point has a corresponding point in the other set *or* the point is a single:

$$\forall j \leq J \quad \sum_{k=1}^{K+1} m_{jk} = 1 \tag{4.8}$$

and

$$\forall k \leq K \quad \sum_{j=1}^{J+1} m_{jk} = 1. \tag{4.9}$$

Note, that there are no constraints on the $(J+1)$'th row or the $(K+1)$'th column because there is no restriction on the number of singles in either point set. Also, $\hat{m}_{(J+1)(K+1)}$ is insignificant (it is not used) and can take any value. □

Example 1 (Two small point patterns) Fig. 4.1 shows two small point patterns $\mathcal{P}$ and $\mathcal{Q}$ with 5 and 7 points respectively. The two sets have 4 points in common, $\mathcal{P}$ has 1 single and $\mathcal{Q}$ has 3 singles. Eq. (4.10) shows the corresponding augmented match matrix □

$$\hat{m} = \left[\begin{array}{ccccccc:c} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ \hdashline 0 & 1 & 0 & 0 & 1 & 0 & 1 & \cdot \end{array}\right]. \tag{4.10}$$

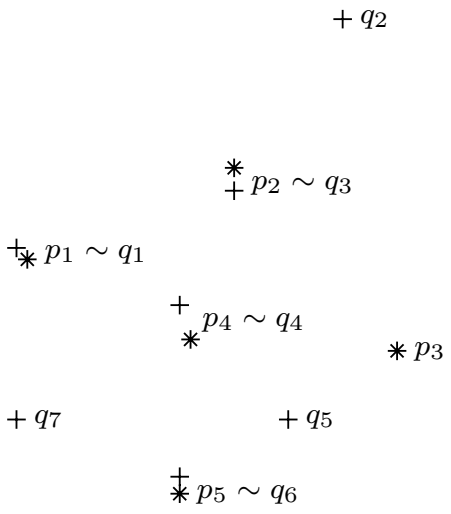


Figure 4.1: Two small point patterns $\mathcal{P}$ ($*$) and $\mathcal{Q}$ ($+$) with 5 and 7 points respectively. The two sets have 4 points in common, $\mathcal{P}$ has 1 single and $\mathcal{Q}$ has 3 singles. The correspondence is defined by $\hat{m}$ in eq. (4.10).

Solving the correspondence problem is a matter of establishing the correct correspondence between points in two point sets, i.e., to estimate the augmented match matrix. This process is referred to as *matching*.

When computing correspondences between point patterns, it may be beneficial to assign non-binary values to the correspondence estimates, i.e., to define a *fuzzy* correspondence matrix. In order to be able to handle partial correspondences the *augmented fuzzy match matrix*, in which entries are real numbers between 0 and 1, is defined.

Definition 6 (Augmented fuzzy match matrix) Given two point patterns $\mathcal{P}$ and $\mathcal{Q}$ of possibly different cardinality J and K. The elements of the augmented fuzzy match matrix $\tilde{m}$ equal the probabilities[1] of correspondence or single status. For the "inner" part of $\tilde{m}$

$$\forall j \leq J,\ \forall k \leq K \qquad \tilde{m}_{jk} = P\{p_j \sim q_k\}$$

for the last column (singles in $\mathcal{P}$)

$$\forall j \leq J,\ k = (K+1) \qquad \tilde{m}_{jk} = P\{p_j \text{ is single}\}$$

and for the last row (singles in $\mathcal{Q}$)

$$\forall k \leq K,\ j = (J+1) \qquad \tilde{m}_{jk} = P\{q_k \text{ is single}\}$$

where $\tilde{m}_{jk} \in [0,1]$, $\forall j \leq (J+1)$, $\forall k \leq (K+1)$. $P\{p_j \sim q_k\}$ means the probability that the j'th point in $\mathcal{P}$ corresponds to the k'th point in $\mathcal{Q}$. Furthermore the entries in the last row and last column of $\tilde{m}$ describe the probability for a point to be single. Naturally the probabilities for each point must sum to one so the row and column constraints still apply:

$$\forall j \leq J \qquad \sum_{k=1}^{K+1} m_{jk} = 1 \quad \text{and}$$

$$\forall k \leq K \qquad \sum_{j=1}^{J+1} m_{jk} = 1. \tag{4.11}$$

As for the binary match matrix there is no constraints on the $(J+1)$'th row or the $(K+1)$'th column because there is no restriction on the number of singles in either point set. □

[1] The entries in the augmented fuzzy match matrix are termed *probabilities* here because of the close analogy to probabilities although, strictly speaking they are not. Other authors use the term *goodness* instead.

To estimate the binary correspondence, $\tilde{m}$ can be converted to a binary augmented match matrix $\hat{m}$. A heuristic method for the binarization is given in BINARIZATION($\tilde{m}$, τ) (Alg. 8). τ is a threshold value. It is ensured, that no points match to several points in the other set.

In general, for a matrix m the *j'th* row and the *k'th* column is referred to as:

$$m_{j\cdot} = \text{j'th row in } m$$

$$m_{\cdot k} = \text{k'th column in } m.$$

4.1.2 Motion estimation

The points in two point patterns to be matched can be related by a (possibly non-linear) transformation $f(\cdot, \psi)$ that brings corresponding points into register.

Definition 7 (Motion) Given two point sets $\mathcal{P}$ and $\mathcal{Q}$ of *equal* cardinality J, and assume the correspondence to be one-to-one and $\forall j, p_j \sim q_j$. Find $f(\cdot, \psi)$ such that

$$\sum_{j=1}^{J} \|p_j - f(q_j, \psi)\|^2 \tag{4.12}$$

is minimal. For this example, the sum of squared Euclidean distances has been chosen as the measure of similarity. The transformation $f(\cdot, \psi)$ is defined as the *motion*, and for a given type of transformation, the parameters to be estimated are ψ. □

Depending on the application f could be a simple rigid transformation (translation and rotation), an affine (translation, rotation, scale and shear), a polynomial or even a non-rigid thin-plate spline transformation. For Euclidean and affine transformation the term *pose* [36] is also used for the motion.

If the cardinality of the point sets is not equal, but the correspondence ($\hat{m}$) is known the optimal transformation can still be found. The optimisation problem (4.12) can then be written

$$\sum_{j=1}^{J} \sum_{k=1}^{K} \hat{m}_{jk} \|p_j - f(q_k, \psi)\|^2, \tag{4.13}$$

i.e., the known correspondences in $\hat{m}$ are used to find the optimal motion, $f(\cdot, \psi)$.

4.1.3 The classical chicken and egg problem

There exists a cyclic dilemma inherently in the point pattern matching problem. Given the correspondence it would be easy to determine a transformation f and on the other hand given an optimal transformation f, it would be easy to establish the correspondence (e.g., by nearest neighbour points).

4.1.4 Graph based methods

The point pattern matching problem is often cast as a graph matching problem. In the field of graph theory, matching is an important subject and the problem of matching point patterns can be formulated in various terms related to graph theory. Examples are *minimum weighted bipartite graph matching* [50] and *relational graph matching* [27]. The notation and description of the graph representation will not be discussed in detail.

4.2 General Point Pattern Matching Methods

This section describes a range of general methods for point pattern matching. Tab. 4.1 gives an overview of the most important methods in the literature. Some of these methods are described in detail in the following sections. The field of point matching is vast and there definitely exist methods not mentioned here. For a number of ways to attack the point matching problem, the table shows references to relevant work, the method used for correspondence calculation, what transformations (motion) is considered, and in what type of application the method has been used (if any). Of the methods in Tab. 4.1 the following are described in further detail:

- iterative closest point (ICP),
- dual step expectation maximisation (EM),
- bipartite graph matching of shape context, and
- robust point matching (RPM)

with emphasis on the RPM methods (and extensions) due to their suitability to match 2DGE spot patterns.

Reference	Correspondence	Transformation (motion)	Applications
Gold and Rangarajan [35].	Robust Point Matching (RPM): deterministic annealing, softassign.	-	Synthetic data.
Rangarajan et al. [69].	Robust Point Matching.	Similarity transform.	Primate cortex autoradiographs.
Rangarajan et al. [67].	Softassign Procrustes matching.	Similarity transform.	Primate cortex autoradiographs.
Chang et al. [20].	Cluster based matching, maximum pairs support.	Affine.	Finger prints.
Gold et al. [36].	Robust Point Matching	Affine.	Hand-written characters, autoradiograph cortex slices.
Besl and McKay [9], Zhang [91].	Iterative Closest Point.	3D Affine.	Synthetic data. Static indoor scene from two views.
Ogawa [62].	Consistency graphs, Delaunay triangulation, maximal cliques.	Affine.	Star constellations.
Cheng [21]	Relaxation, Select-match-pair process.	Affine.	Synthetic data.
Cross and Hancock [27].	Dual step EM algorithm, relational graph matching.	Affine, perspective.	Road networks, aerial imagery.
Carcassoni and Hancock [17].	Spectral graph theory, point proximity matrix.	Affine.	
Guest et al. [38].	Correspondence by sensitivity to movement, modified ICP.	2D elastic, 3D rigid.	Warping of 2D serial histological sections of mouse embryo.
Chui and Rangarajan [23], Rangarajan et al. [68].	RPM.	Non-rigid, thin-plate spline.	synthetic data, anatomical sulcal brain MRI data
Sclaroff and Pentland [72].	Modal matching.	Non-rigid.	Hand tools, airplane silhouettes.
Kumar et al. [50], Kumar et al. [49].	Bipartite weighted graph matching. Hybrid of greedy and Hungarian algorithm.	Non-rigid.	Features in mammogram images, hurricane images.
Belongie and Malik [5], Belongie et al. [7].	Bipartite graph matching of shape contexts.	Thin-plate spline.	Recognition of trademarks, handwritten digits.
Murtagh [61].	"World view" vector similarity.	Non-rigid.	Star constellations.

Table 4.1: Overview of general point pattern matching methods.

4.2.1 Iterative closest point

Iterative closest points (ICP), Besl and McKay [9] and Zhang [91] is one of the more well-known methods for point pattern matching. The general idea is to iteratively match points in one set to closest points in another set. [9] proposes the ICP algorithm as a 3D shape *rigid* registration algorithm for point sets, curves and surfaces. Given two point sets $\mathcal{P}$ and $\mathcal{Q}$ for each point in $\mathcal{P}$ the closest point in $\mathcal{Q}$ is computed and from that a transformation (translation and rotation) is calculated. This transformation is applied to the points in $\mathcal{P}$. These steps are repeated in a iterative manner until the change in mean square error is below a given threshold. The algorithm is guaranteed to converge to a local minimum. One of the major disadvantages of this algorithm is its inability to handle a large proportion of outliers (singles) [9], [22].

4.2.2 Dual step EM

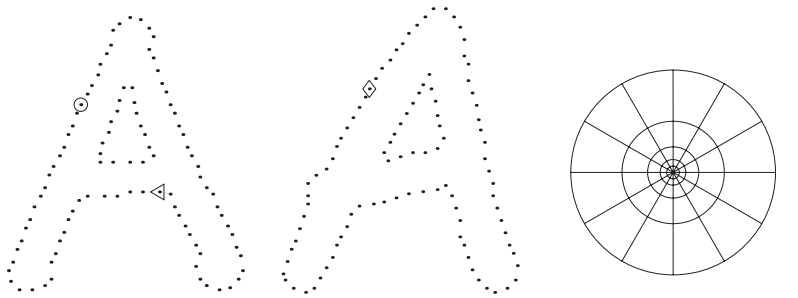
Cross and Hancock [27] and Carcassoni and Hancock [17] describe an approach very similar to the basic idea in RPM, namely alternately estimation of transformation parameters and correspondence matches, referred to as the dual step expectation maximisation algorithm (dual step EM). The transformation is confined to the affine or perspective geometry and relational graph matching of separate triangulation of the point sets constitutes the matching.

4.2.3 Bipartite graph matching of shape context

For the purpose of matching shapes Belongie and Malik [5], [6, 7] have developed a point matching method based on the matching of shape context descriptors. For each point, a shape context descriptor is calculated and the correspondence problem is solved by using bipartite graph matching of the descriptors. Fig. 4.2 is from [6].

The shape context descriptor of a point is a two-dimensional log-polar histogram. For the point in question, the mask in Fig. 4.2(c) is placed with the centre on the point and the count of points in each bin is assigned to the corresponding cell in the log-polar histogram. The polar coordinate is divided into 12 bins and the $\log r$ is divided into 5 bins.

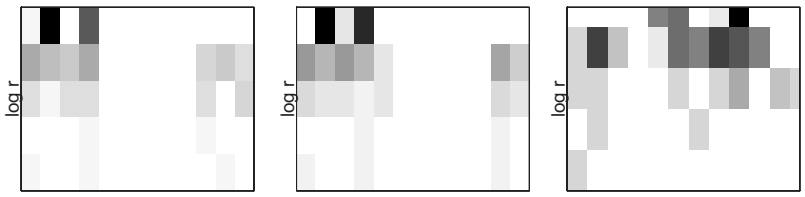
The histograms are used to calculate a cost matrix C, $C_{ij} = C(p_i, q_j)$ is the cost of matching point p_i with q_j and the matching is solved as a square assignment problem using an efficient implementation of the Hungarian method.



(a) Pattern 1.

(b) Pattern 2.

(c) Log-polar histogram bins.



(d) Shape context descriptor for point marked with ○ in Fig. 4.2(a).

(e) Shape context descriptor for point marked with ◇ in Fig. 4.2(b).

(f) Shape context descriptor for point marked with ◁ in Fig. 4.2(a).

Figure 4.2: Shape context and matching. From Belongie et al. [6].

If the two sets do not contain the same number of points, extra "dummy" points (with a sufficiently large cost) can be added to either of the point sets. This will ensure a square C_{ij}. The same trick is used to make the method more robust to singles. This does however require an estimate of the maximum number of singles in either set.

Heavily contaminated point sets can be suspected to cause problems, and [7] provides no solution for this case.

Ideas very similar to the shape context descriptor has previously been suggested by [64] (see § 4.3.1).

The complete matching method employs alternately estimation of the correspondence and a thin-plate-spline transformation inspired by the approach in [23] (see also § 4.2.4).

4.2.4 Robust point matching

Robust point matching (RPM) is a method based on both the estimation of the point correspondence and a the transformation (motion, pose) in a *deterministic annealing* setting. Alternately, the correspondence and the transformation is estimated in an iterative manner and by use of the *augmented fuzzy match matrix* (Def. 6) and *Sinkhorn's method* for matrix normalisation, RPM handles singles in a robust manner. The original method proposed by Gold and Rangarajan [35] has been extended to handle various types of transformations: similarity transform [69], affine transformation [36], Procrustes alignment [67], [70], and the non-rigid thin-plate spline transformation [68], [23].

In the following, the RPM method in general and some extensions to the *Sinkhorn's method* will be described. Two types of transformations, the affine and thin-plate spline, will be discussed in detail. The level of detail is greater in the discussion of the RPM methods because their ability to match protein spot patterns from electrophoresis gels has been investigated thoroughly (§4.5).

Given two point sets $\mathcal{P}$ and $\mathcal{Q}$ of possibly different cardinality and assuming that some knowledge about the *type* of transformation (motion), that relates two point sets $\mathcal{P}$ and $\mathcal{Q}$, is available, (e.g., that the motion is affine), it remains to estimate the optimal parameters ψ in the transformation $f(\cdot, \psi)$ *and* the correspondence $\hat{m}$.

The assignment problem and estimating the transformation parameters is formulated as an energy minimisation problem on the form:

$$\begin{aligned} \min_{\hat{m},\psi} E &= \sum_{j=1}^{J}\sum_{k=1}^{K} \hat{m}_{jk}\|p_j - f(q_k,\psi)\|^2 \\ &- \zeta\sum_{j=1}^{J}\sum_{k=1}^{K}\hat{m}_{jk} + T\sum_{j=1}^{J}\sum_{k=1}^{K}\hat{m}_{jk}(\log\hat{m}_{jk}-1) \\ &+ \lambda J(f(\psi)). \end{aligned} \tag{4.14}$$

The first term is the sum of squared distances between the points in the two sets. The weighting of the distances by $\hat{m}_{jk}$ penalises distances between points with high probability to match. The second term provides a regularisation on the match matrix $\hat{m}$ ensuring that not all points are classified as singles, which would cancel out the first term of the energy function. The third term is an entropy barrier function to ensure positivity of $\hat{m}$. To ensure reasonable behaviour of the transformation, the fourth term serves as a regularisation on the transformation $f(\cdot,\psi)$. ζ and λ serve as regularisation parameters.

This is a general form of the energy function and later more specific forms (in cases of f being an affine and a thin-plate spline transformation) are presented.

Algorithm 1 ROBUST-POINT-MATCHING($\mathcal{P}$, $\mathcal{Q}$, T_0)

1: Initialise $\psi = 0$. *All parameters are initialised to zero*
2: Initialise $T = T_0$.
3: Initialise $\tilde{m}_{jk} = 1 + \epsilon$.
4: **repeat** *Deterministic Annealing*
5: **repeat** *at each temperature level*
6: $Q_{jk} \leftarrow -\dfrac{\partial E}{\partial m_{jk}}$
7: $\tilde{m}_{jk} \leftarrow \exp(Q_{jk}/T)$ *Ensure positivity of $\tilde{m}$*
8: $\tilde{m} \leftarrow$ SINKHORN($\tilde{m}$) *Sinkhorn's normalisation*
9: Estimate transformation parameters ψ
10: **until** # of iterations $> I_0$
11: Decrease T, decrease λ
12: **until** $T < T_{end}$
13: **return** $\tilde{m}, f(\cdot,\psi)$

The RPM algorithm outlined in Alg. 1 is explained in further detail.

The outer loop is the Deterministic Annealing (DA). It basically consists of slowly decreasing the temperature parameter T. T plays an important role in

Algorithm 2 SINKHORN($\tilde{m}$)

1: **repeat**
2: Normalise all rows:
3: $\forall j \leq J+1, \forall k \leq K+1 \quad \tilde{m}^1_{jk} \leftarrow \frac{\tilde{m}_{jk}}{\sum_{k'=1}^{K+1} \tilde{m}_{jk'}}$
4: Normalise all columns:
5: $\forall j \leq J+1, \forall k \leq K+1 \quad \tilde{m}_{jk} \leftarrow \frac{\tilde{m}^1_{jk}}{\sum_{j'=1}^{J+1} \tilde{m}^1_{j'k}}$
6: **until** $\tilde{m}$ converges or # of iterations $> I_1$

the estimation of the point correspondences and is used to weight the importance of the inter-point distances. Note how the current estimate of the transformation, $f(\cdot, \psi)$ is used to calculate the next estimate of the match matrix, $\hat{m}$. In the first iterations (at large T) the distance between points is less important for the correspondence, but as the temperature decreases, small distances are rewarded more and large distances are penalised more.

In the inner loop the matrix $Q_{jk} \leftarrow -\frac{\partial E}{\partial m_{jk}}$ is calculated. For fixed parameters ψ this is an assignment problem [35], reversing the sign because this is a minimisation instead of a maximisation as in the canonical form of the assignment problem. With the transformation $\tilde{m} \leftarrow exp(Q_{jk}/T)$ (line 7) it is ensured that a small Euclidean distance between a point pair corresponds to a large probability of match for this pair.

For the elements in $\tilde{m}$ to be probabilities the column and row constraints (4.11) must hold. This is ensured (line 8) by the *Sinkhorn's* matrix normalisation (Alg. 2)[1]. The convergence criterion of $\tilde{m}$ in Alg. 2 (line 6) is a limit on the sum of absolute differences of matrix elements between two consecutive iterations of the loop:

$$\sum_{j=1}^{J+1} \sum_{k=1}^{K+1} |\tilde{m}_{jk} - \tilde{m}^0_{jk}| < \epsilon_2.$$

After this normalisation the newly updated match matrix $\tilde{m}$ is used to update the transformation parameters ψ (line 9). The knowledge of the fuzzy correspondences allows to put more emphasis on high probability matches in the parameter estimation. Furthermore there will often be some regularisation of the parameters ψ to avoid inappropriate behaviour of the transformation. This will be demonstrated in the specific examples of the transformation being affine and a thin-plate spline.

[1] SINKHORN($\tilde{m}$) is described in Alg. 2, however, it is suggested to use the EXTENDED-SINKHORN(,) Alg. 3 instead.

After a fixed number of iterations I_0 at the current temperature level the inner loop is done and the temperature is decreased (line 11), e.g., by an exponential temperature schedule. Simultaneously, the regularisation weight λ for the regularisation of the transformation parameters is relaxed (decreased). This is because the transformation parameters need only to be bounded in the beginning of the iterations when corresponding points may be far apart. As the algorithm progresses the estimates of the transformation parameters become better and there is less need to bound them.

In the first iterations (large T) all pairs will be assigned a small probability of match but as T decreases point pairs close to each other will be rewarded with relatively larger probability for match and vice versa. Line 7 and 8 in Alg. 1 is termed *softassign* [35] because the inter-distances between points are converted to fuzzy (soft) correspondences (assignment probabilities).

For equal point sets, $\mathcal{P}$ and $\mathcal{Q}$, the energy function in (4.14) is a linear assignment problem [23] or a bipartite matching problem with respect to the correspondences. The RPM algorithm minimises the energy function because *softassign* used within *deterministic annealing* has been shown to find the optimal solution to linear assignment [47]. Although not strictly an assigment problem, Gold et al. [36] imply this result is valid for the case of unequal point sets by introducing slack variables. However, no proof is given. Furthermore, [23] state that while softassign and the transformation estimation are independently optimal, the combination is not necessarily so. This approach is only guaranteed to find a local minimum for both the mapping and the correspondence.

Application of the RPM algorithm to 2DGE data has proven to work well in practice (§ 4.5).

RPM algorithm initialisation

All parameters of the transformation are initialised to $\psi = 0$ or so that the point set is undergoing no transformation in the beginning because no knowledge of this transformation is available. T_0 is the starting temperature and should be chosen suitably large. [23] proposes to set T_0 slightly larger than the greatest squared distance of all point pairs to allow for all possible matchings at first. In cases with almost aligned point sets and small deformations (e.g., for 2DGE spot patterns) T_0 can be lowered to save computational effort.

The ζ parameter serves as a threshold error distance and indicates how far two points can be apart before being treated as outliers. This parameter needs to be adjusted depending of the problem at hand.

The λ parameter can be adjusted to control the penalty on the transformation parameters and must be chosen dependent upon the problem at hand.

Every element in the match matrix $\hat{m}_{jk}$ is initialised to $1 + \epsilon$, where ϵ is a small number, e.g. 10^{-6}. This only has effect on the last row and the last column (for the outliers) because the "inner" part of $\hat{m}$ is immediately overwritten. Alternatively, [23] suggests to initialise $\hat{m}$ such that entries in the $(J \times K)$ inner part of the matrix are all set to $\frac{1}{J}$ (which seems strange, because these elements are again immediately overwritten) and the elements in the outlier row and the outlier column are initialised to $\frac{1}{100J}$. The initialisation first suggested has been used in the experiments conducted later (§4.5).

Extended Sinkhorn's method

There are problems using the Sinkhorn method to normalise $\tilde{m}$. First, $\tilde{m}$ is often a non-square matrix, but the Sinkhorn result is valid for *square* matrices only. Second, in the definition of the fuzzy match matrix (Def. 6) there are *no* constraints on the last row and the last column of $\tilde{m}$, so these should *not* be normalised as it happens in Alg. 2. In other words, the elements in the last column should not be normalised with respect to the sum of that column, but its elements should still enter in the normalisation of the corresponding matrix rows and vice versa for the last row. This idea has been implemented in Alg. 3.

The changes from Alg. 2 (as defined in Gold et al. [36]) inside the convergence criterion has merely to do with not normalising the last row and the last column of $\tilde{m}$ according to the reasons given above.

One may easily think of a "pathological" case where Alg. 3 (and also Alg. 2) will fail.

Example 2 ("pathological case") Consider the augmented match matrix $\tilde{m}$ representing the correspondence between two points in one set ($\mathcal{P}$) and three points in the other set ($\mathcal{Q}$). There is a tie between q_1 and q_2 to match p_1.

$$\tilde{m} = \left[\begin{array}{ccc:c} 1 & 1 & 0 & 0 \\ \hdashline 0 & 0 & 1 & 1 \\ \hdashline 0 & 0 & 1 & 0 \end{array}\right]$$

The row and column normalisations will result in the unfortunate result:

$$\tilde{m} = \left[\begin{array}{ccc:c} 1 & 1 & 0 & 0 \\ \hdashline 0 & 0 & .38 & .62 \\ \hdashline 0 & 0 & .62 & 0 \end{array}\right]$$

and the last two lines in Alg. 3 adjusting the outlier row and column result in:

$$\tilde{m} = \left[\begin{array}{ccc:c} 1 & 1 & 0 & -1 \\ 0 & 0 & .38 & .62 \\ \hdashline 0 & 0 & .62 & 0 \end{array}\right]$$

which is not an acceptable result. □

On the other hand $\tilde{m}$ was not valid from the beginning because all elements must be *strictly positive*.

Example 3 By adding a small constant, e.g. 10^{-5} to all elements in $\tilde{m}$,

$$\tilde{m}' = \tilde{m} + 10^{-5}$$

a valid input matrix to Extended-Sinkhorn is obtained and the result, valid to two decimal places after 79 iterations of Alg. 3 becomes:

$$\tilde{m} = \left[\begin{array}{ccc:c} 0.50 & 0.50 & 0.00 & 0.00 \\ 0.15 & 0.15 & 0.29 & 0.41 \\ \hdashline 0.35 & 0.35 & 0.71 & 0 \end{array}\right]$$

which is a valid match matrix with all rows and columns summing to one, save the outlier row and column. The tie has not been resolved (of course), and the binarization as proposed in § B, Alg. 8 of the result with $\tau = 0.5$ will result in:

$$\hat{m} = \left[\begin{array}{ccc:c} 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ \hdashline 1 & 1 & 1 & 0 \end{array}\right]$$

i.e., all points are declared outliers which seems reasonable considering the input. A less conservative setting of the binarization threshold $\tau = 0.2$ results in

$$\hat{m} = \left[\begin{array}{ccc:c} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ \hdashline 1 & 1 & 0 & 0 \end{array}\right].$$ □

The examples above illustrate the reason for the initialisation of $\tilde{m}$ as shown in Alg. 1. It is important, that the outlier row and column do *not* contain zero elements.

Sinkhorn proves that any square matrix with all *positive* elements will converge to a doubly stochastic matrix by alternately normalising rows and columns in

an iterative manner. A doubly stochastic matrix is a square matrix in which elements are all positive and sum to one both along rows and along columns. The problem of $\tilde{m}$ often being non-square is solved by a simple heuristic. At the end of the iterations the following rule has been added: if the "inner" columns and rows do not sum to one this is forced to happen by appropriately adjusting the outlier values (line 8-9, Alg. 3).

Algorithm 3 EXTENDED-SINKHORN($\check{m}$)

1: Initialise $\tilde{m} \leftarrow \check{m}$
2: **repeat**
3: Normalise the J top rows:
4: $\forall j \leq J, \forall k \leq K+1 \qquad \tilde{m}^1_{jk} \leftarrow \dfrac{\tilde{m}_{jk}}{\sum_{k'=1}^{K+1} \tilde{m}_{jk'}}$
5: Normalise the K leftmost columns
6: $\forall j \leq J+1, \forall k \leq K \qquad \tilde{m}_{jk} \leftarrow \dfrac{\tilde{m}^1_{jk}}{\sum_{j'=1}^{J+1} \tilde{m}^1_{j'k}}$
7: **until** $\tilde{m}$ converges or # of iterations $> I_1$
8: $\forall j \leq J \quad \tilde{m}_{j(K+1)} \leftarrow 1 - \sum_{k=1}^{K} \tilde{m}_{jk}$ *force top J rows to sum to 1*
9: $\forall k \leq K \quad \tilde{m}_{(J+1)k} \leftarrow 1 - \sum_{j=1}^{J} \tilde{m}_{jk}$ *force K leftmost columns to sum to 1*

Although no proof for convergence is given, these changes have proven very useful in making the point matching algorithm more robust.

A simulation with 1000 matrices was carried out in order to study the convergence properties of Alg. 3. The dimensions of the matrices, $J \times K$ were random integer numbers drawn from $\{2, 3, ..., 100\}$, i.e., the matrices were *non-square* in most cases, but also square matrices could occur. The matrix entries were random real numbers in $]0, 1]$. The stop criteria was chosen so that the limit for change in $\tilde{m}$, $\epsilon_2 = 10^{-6}$ and the maximum number of iterations $I_1 = 1000$. All simulations converged to match matrices, fulfilling the constraints in (4.11) in less than 105 iterations.

In this thesis the EXTENDED-SINKHORN($\hat{m}$) method substitutes the SINKHORN($\hat{m}$) method in all experiments.

RPM with affine transformation

One of the important transformations implemented in the Robust Point Matching scheme is the affine pose [36].

Here, the affine transformation of a point in $\mathcal{Q}$ is

$$f(q_k, \psi) = t + Aq_k$$

where t is a translation vector and A is a 2×2 matrix defined as

$$A = s(a)R(\Theta)Sh_1(b)Sh_2(c)$$

where

$$s(a) = \begin{pmatrix} e^a & 0 \\ 0 & e^a \end{pmatrix}, \qquad Sh_1(b) = \begin{pmatrix} e^b & 0 \\ 0 & e^{-b} \end{pmatrix}, \qquad Sh_2(c) = \begin{pmatrix} \cosh(c) & \sinh(c) \\ \sinh(c) & \cosh(c) \end{pmatrix}$$

and $R(\Theta)$ is the well known rotation matrix:

$$R(\Theta) = \begin{pmatrix} \cos(\Theta) & -\sin(\Theta) \\ \sin(\Theta) & \cos(\Theta) \end{pmatrix}$$

The pose parameters can be summarised in

$$\psi = (t, \Theta, a, b, c).$$

In the case of affine transformation the energy function in Eq. 4.14 becomes:

$$\begin{aligned} E &= \sum_{j=1}^{J}\sum_{k=1}^{K} \hat{m}_{jk}\|p_j - t - Aq_k\|^2 \\ &- \alpha\sum_{j=1}^{J}\sum_{k=1}^{K} m_{jk} + T\sum_{j=1}^{J}\sum_{k=1}^{K} \hat{m}_{jk}(\log \hat{m}_{jk} - 1) \\ &+ \gamma g(A) \end{aligned} \tag{4.15}$$

composed of four terms. The first term is the sum of squared weighted distances between points in $\mathcal{P}$ and the transformed points in $\mathcal{Q}$.

The second term ensures that not all points are classified as singles. The α parameter is a threshold error distance that determines how distant two points can be before their match is unlikely.

The fourth term $\gamma g(A)$ ($= \lambda J(f(\psi))$ in Eq. (4.14)) provides a regularisation of the affine transformation by penalising large values of the scale and shear parameters:

$$\gamma g(A) = \gamma(a^2 + b^2 + c^2),$$

where γ allows for adjustment of the penalty.

By differentiation of the energy function in (4.15) with respect to the match matrix and setting equal to zero [36]:

$$\begin{aligned}
\frac{\partial E}{\partial \hat{m}_{jk}} &= \|p_j - t - Aq_k\|^2 - \alpha + T(\log(\hat{m}_{jk}) - 1) + T\frac{\hat{m_{jk}}}{\hat{m}_{jk}} \\
&= \|p_j - t - Aq_k\|^2 - \alpha + T\log(\hat{m}_{jk}) \;=\; 0 \qquad \Rightarrow \\
\hat{m}_{jk} &= \exp\big(-\frac{\|p_j - t - Aq_k\|^2 - \alpha}{T}\big)
\end{aligned}$$

Gold et al. [36] also provides analytical solutions to update Θ and t and suggests using Newton's method to update a, b, and c.

In the case of affine transformation the Alg. 1 specialises to Alg. 4.

Algorithm 4 RPM-AFFINE($\mathcal{P}$, $\mathcal{Q}$, T_0)

```
Initialise ψ = (t, Θ, a, b, c) = 0.          All parameters are initialised to zero
Initialise T = T_0.
Initialise m̃_jk = 1 + ε.
repeat                                        Deterministic Annealing
   repeat                                     at each temperature level
      Q_jk ← −(‖p_j − t − Aq_k‖² − α)
      m̃_jk ← exp(Q_jk/T)                      Ensure positivity of m̃
      m̃ ← SINKHORN(m̃)                         Sinkhorn's normalisation
      Update t and Θ using analytic solutions  Update pose parameters
      Update a, b, and c using Newton's method
   until # of iterations > I_0
   Decrease T and γ
until T < T_end
return m̃, f(·, ψ)
```

Pedersen and Ersbøll [65] suggested to gradually reduce α so that in the beginning, when the match matrix and the affine pose parameters are uncertain, larger distances between points is allowed for. Later (after more iterations), when the correspondence and the pose are more certain, a smaller distance between points is required for them to be likely matches.

§4.3.4 embeds this method in a regionalised setting and results of application to matching of 2D electrophoresis spot patterns are given in §4.5.

For the experiments in §4.5 the following values were used: $T_0 = 110$, $\alpha = 100$, $\gamma = 1$, $\epsilon = 10^{-5}$, $\epsilon_2 = 0.5$, $I_0 = 20$, $I_1 = 5$. The annealing rate (the rate at which T and γ was decreased) was 0.93 for both parameters.

RPM with thin-plate spline transformation

Another transformation, the non-rigid thin-plate spline is extremely flexible and well suited for modelling the complex disparity field between 2DGE point sets. Chui and Rangarajan [23] and Rangarajan et al. [68] have developed a powerful method for using the thin-plate spline transformation in conjunction within the Robust Point Matching framework. The thin-plate spline transformation is described in §A.

The thin-plate spline functional on the form (A.2)

$$f(q_k, \psi) = q_k \cdot d + \phi(q_k) \cdot c \tag{4.16}$$

is the sum of an affine part and a non-affine part. The parameters are $\psi = (d, c)$. d is a (3×3) affine parameter matrix (using homogeneous coordinates, refer to (§A)) and the non-affine parameters c is a $(K \times 3)$ matrix.

The energy function minimised by this algorithm is

$$\begin{aligned} E(\hat{m}, d, w) &= \sum_{j=1}^{J}\sum_{k=1}^{K} \hat{m}_{jk}\|p_j - q_k d - \phi(q_k)c\|^2 \\ &- \zeta\sum_{j=1}^{J}\sum_{k=1}^{K}\hat{m}_{jk} + T\sum_{j=1}^{J}\sum_{k=1}^{K}\hat{m}_{jk}(\log\hat{m}_{jk} - 1) \\ &+ \lambda_1 \operatorname{tr}(c^T\Phi c) + \lambda_2 \operatorname{tr}[(d-I)^T(d-I)]. \end{aligned} \tag{4.17}$$

The first term is again the sum of squared distances between the points in $\mathcal{P}$ and the transformed points in $\mathcal{Q}$. The second term also again ensures that not all points are classified as singles. The third term is an entropy barrier function to ensure positivity of $\hat{m}$. The last two terms are regularisation terms on the affine (d) and non-affine (c) parameters respectively. The penalty on the behaviour of these parameters is split into two terms, conveniently providing separate control (λ_1 and λ_2) over the affine and non-affine behaviour of the transformation. The parameters ζ, λ_1, and λ_2 must be chosen according to the problem at hand.

Differentiation of (4.17) with respect to $\hat{m}$ and setting the result to zero leads to [23]:

$$Q_{jk} = -\big(\|p_j - q_k d - \phi(q_k)c\|^2 - \zeta\big) \quad \text{and}$$

$$\tilde{m}_{jk} = \exp(Q_{jk}/T),$$

i.e., an estimate of the correspondence $\tilde{m}$. After normalisation of the fuzzy correspondence estimates the thin-plate spline parameters can be estimated.

§ A explains how, using QR decomposition, a least squares estimate of the thin-plate spline parameters (d, c) can be obtained when correspondences are known (at least assumed known). Alg. 5 shows the RPM-TPS scheme.

Algorithm 5 RPM-TPS($\mathcal{P}$, $\mathcal{Q}$, T_0)

```
Initialise ψ = 0.                                  All parameters are initialised to zero
Initialise T = T_0.
Initialise m̃_jk = 1 + ε.
repeat                                             Deterministic Annealing
   repeat                                          at each temperature level
      Q_jk ← −(‖p_j − q_k d − φ(q_k)c‖² − ζ)
      m̃_jk ← exp(Q_jk/T)                           Ensure positivity of m̃
      m̃ ← SINKHORN(m̃)                              Sinkhorn's normalisation
      Update ψ = (d, c)                            Update TPS parameters
   until # of iterations > I_0
   Decrease T
   Decrease λ_1 and λ_2
until T < T_end
return m̃, f(·, ψ)
```

Note, that the regularisation parameters are relaxed as the iterations proceed. This way the transformations are allowed more freedom when the points are brought closer. In the beginning of the iterations, the correspondence is uncertain and the points may be far apart, but later when the match matrix is more certain, the transformation is allowed to behave more freely.

For the experiments in §4.5 the following values were used: $T_0 = 300$, $\zeta = 0$, $\lambda_1 = 10^6$, $\lambda_2 = 10^6$, $\epsilon = 10^{-5}$, $\epsilon_2 = 0.5$, $I_0 = 20$, $I_1 = 5$. The annealing rate (the rate at which T was decreased) was 0.93. λ_1 and λ_2 were kept constants.

Extended RPM TPS

The RPM-TPS method uses the point *locations* in the estimation of correspondence and transformation. Often additional point information other than the

spatial location of the points is available. The RPM-TPS method can easily be extended to include the point attribute information in the point matching. Provided that there is some correlation between the corresponding points and their attribute information, the attribute information has the ability to reduce the number of possible matches for each point, resulting in a faster and more reliable algorithm.

Given D scalar attributes to each point in $\mathcal{P}$ and $\mathcal{Q}$. The D attributes to point p_j in $\mathcal{P}$ are the aligned in the $(D \times 1)$ attribute information vector a_j^p and similarly for $\mathcal{Q}$.

With offset in the energy function (4.17) a sixth term is added to include the attribute information in the matching process:

$$\gamma \sum_{j=J}^{J} \sum_{k=1}^{K} \hat{m}_{jk} \delta(a_j^p, a_k^q)$$

where $\delta(\cdot, \cdot)$ is some distance measure. γ is a parameter to control the influence of the attribute information. The final energy function gathers to:

$$\begin{aligned} E(\hat{m}, d, w) &= \sum_{j=1}^{J} \sum_{k=1}^{K} \hat{m}_{jk} \|p_j - q_k d - \phi(q_k) c\|^2 \\ &- \zeta \sum_{j=1}^{J} \sum_{k=1}^{K} \hat{m}_{jk} + T \sum_{j=1}^{J} \sum_{k=1}^{K} \hat{m}_{jk} (\log \hat{m}_{jk} - 1) \\ &+ \lambda_1 \operatorname{tr}(c^T \Phi c) + \lambda_2 \operatorname{tr}[(d - I)^T (d - I)] \\ &+ \gamma \sum_{j=J}^{J} \sum_{k=1}^{K} \hat{m}_{jk} \delta(a_j^p, a_k^q). \end{aligned} \tag{4.18}$$

Setting $\frac{\partial E}{\hat{m}_{jk}} = 0$ results in

$$Q_{jk} = -\big(\|p_j - q_k d - \phi(q_k) c\|^2 - \zeta + \gamma \delta(a_j^p, a_k^q)\big),$$

which will require a corresponding change of Alg. 5, line 6.

4.3 Point Matching of Protein Spot Patterns

To remind of the protein spot pattern nature, Fig. 4.3 repeats Fig. 2.13 from §2. The known correspondences in Fig. 4.3 show the very complex nature of

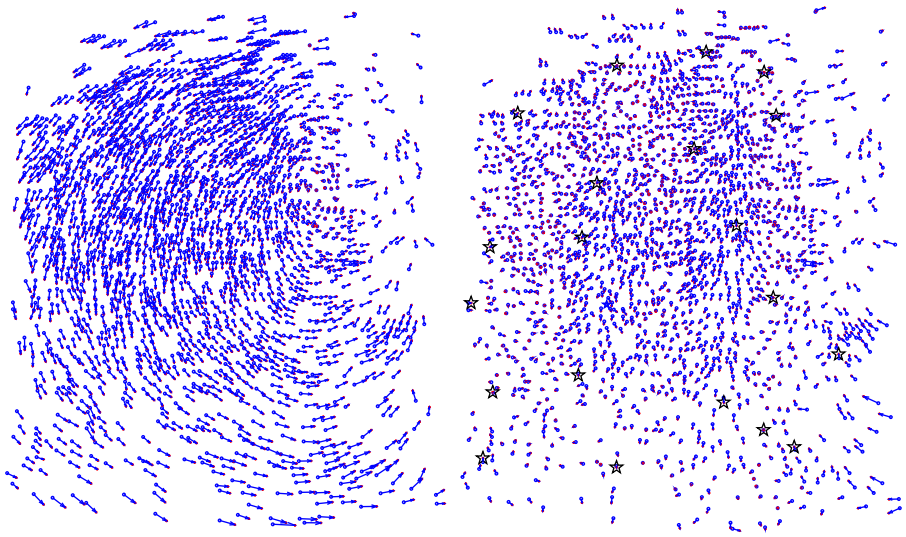


Figure 4.3: Known correspondence between spots in gel A and gel B. Corresponding spots are connected with small arrows. Left: Before initial alignment. Right: Residual after correction for 1st order polynomial transformation using landmarks. Manually defined landmarks are marked with stars.

the protein spot patterns. The plot of two point sets with corresponding points connected with arrows or lines is referred to as the *disparity field* for the two point sets. The left plot in Fig. 4.3 shows the disparity field of two point sets $\mathcal{P}$ and $\mathcal{Q}$. The dominant rotation and translation suggests that a large part of the disparity field can be recovered by a global alignment of the point patterns. Therefore, a disparity field model δ on the following form is suggested:

$$\delta = \delta_g + \delta_l, \tag{4.19}$$

where the total disparity δ is a sum of two contributions, a global field and a local field. The first term, δ_g is the *global disparity* field and can be described by a fairly simple parametric transformation that accounts for the different location of the gels during the image capture process. The second term, δ_l models the *local* non-linear deformations of the gels due to non-uniform diffusion constants in the gel, non-homogeneity of the chemical solutions and the electrical potential fields involved in the protein separation process.

The methods presented in the rest of this section are developed to estimate point correspondences between two sets of protein spot centres. The input to these methods have all been corrected for the global disparity field δ_g.

The matching of points instead of protein spots is based on successful segmentation of images into spots and non-spots. Please refer to § 3 for details on the gel segmentations.

Comparison of matching results in the literature is difficult because of the variety of visualisation methods used to produce the gel images. Some methods generate images with only a small number of spots while methods more sensitive (as [^{35}S]-methionine labelling detected with a phosphor imager) are able to detect far more proteins from the same sample (see Fig. 2.5). The sensitive methods detect more proteins in the same area, hence more dense point sets. Intuitively, the risk of making a mistake in the matching is higher when matching dense point sets.

Another factor impeding the comparison of match results within the literature is that, even using the same visualisation method, inter-laboratory differences (procedure, detergents, equipment, etc.) influence the gel images and therefore also the density of the spots and the reproducibility of the gels in general.

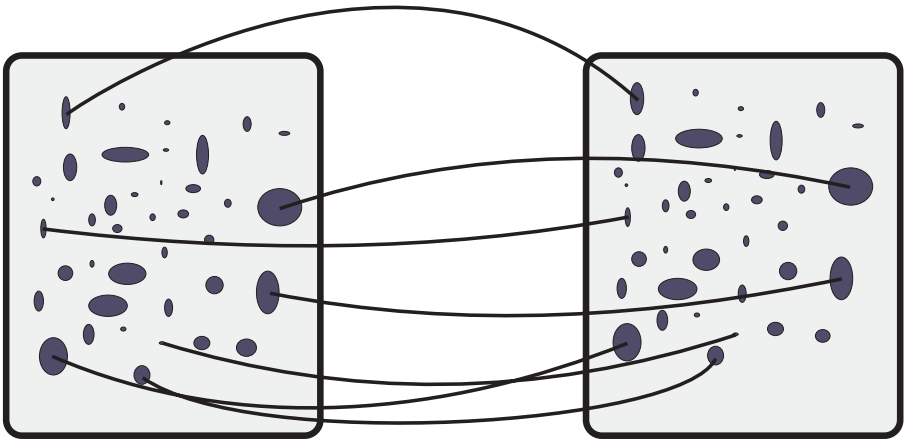


Figure 4.4: Principal sketch of (partial) correspondences between protein spots in two gel images. In order to compare protein expression levels between two gels the correct correspondence between matching spots is necessary. Repetition of Fig. 2.11.

A literature overview of important attempts to match protein spots is given in Tab. 4.2. Selected methods in the areas of:

- neighbourhood based matching,

Reference	Correspondence	Transformation (motion)	Outlier handling	Commercial software
Watanabe et al. [88], Watanabe and Takahashi [87], Watanabe and Takahashi [86].	Graph matching of Delaunay and relative neighbourhood graphs. Breadth first search.	-	Yes.	DNAinsight.
Pánek and Vohradský [64].	Maximal similarity of neighbourhood descriptors and relative position.	-	Yes.	
Akutsu et al. [2].	Neighbourhood based, dynamic programming, practical heuristic algorithm.	-	Yes.	
Appel et al. [3], Miller et al. [60].	Heuristic cluster matching.	-	Yes.	Melanie II.
Hoffmann et al. [40], Hoffmann et al. [41], Pleissner et al. [66].	Incremental Delaunay triangulation, hashing. Uses spot intensities.	-	Yes.	CAROL
Horgan et al. [42].	Visual comparison by colour super-imposition.	Affine, thin-plate spline warp of entire gel image using manually selected landmarks.	No.	
Menard and Soulie [59]	Graph matching of topological maps	-	?	

Table 4.2: Point pattern matching methods applied to 2D electrophoresis protein spot patterns.

- graph based matching,
- successive point matching, and
- regionalised robust point matching (regionalised RPM)

will be discussed with emphasis on the regionalised RPM methods.

4.3.1 Neighbourhood based matching

Comparison of spot neighbourhoods [64], [2], [3] is a popular approach in matching of electrophoretic protein spot patterns. Given the usual matching problem: two gels, a *reference* gel Ω_p and a *match* gel Ω_q with point patterns $\mathcal{P}$ and $\mathcal{Q}$. Find the best match matrix $\hat{m}$.

The common idea behind these methods is: For a given point p_j in $\mathcal{P}$, find its homologous point q_k in $\mathcal{Q}$. It is likely that the neighbourhood of q_k "looks like" the neighbourhood of p_j. A neighbourhood descriptor is defined for each point in both sets and these neighbourhood descriptors are then compared in order to find the best match for each point.

This idea attempts to mimic the method of an investigator performing the matching of spot patterns by eye.

Pánek and Vohradský

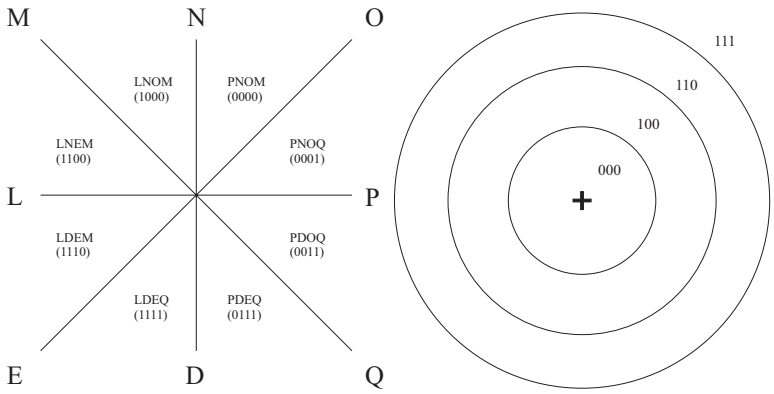
One of the more promising of the existing approaches to matching protein spot patterns is the one by Pánek and Vohradský [64]. They attempt to match the spot patterns by comparing spot neighbourhoods. A syntactic descriptor of the neighbourhood is defined and together with a metric definition of position similarity a simple comparison is done with five possible candidate points. For a point p_j the candidate points in $\mathcal{Q}$ are the five closest points to the position of p_j in $\mathcal{Q}$.

The syntactic neighbourhood descriptor is related to the shape context descriptor in [5]. Pánek and Vohradský [64] also divides the neighbourhood around the spot into segments (concentric rings and wedges) each with a binary code. See Fig. 4.5.

The segment codes of neighbouring segments differ by only one binary digit. Note the similarity to the binary Gray-code. Furthermore, codes for opposite segments have no digits in common.

Spots in the neighbourhood of a spot are assigned the binary code corresponding to their relative position in the neighbourhood. These segment codes are arranged in a descriptor matrix. The similarity between two descriptor matrices (one from $\mathcal{P}$ and one from $\mathcal{Q}$) can then be used to find the best match for a spot.

The paper also describes a measure of "goodness of match" by comparing so called *directional vectors*. Given a number of landmarks in both $\mathcal{P}$ and $\mathcal{Q}$ the difference in length and absolute angles of the directional vectors from the point to the nearest landmark is used to determine the goodness of match. In other words, for a match ($p_j \sim q_k$) to be good p_j and q_k must be located in a similar way relatively to the same landmarks.



(a) Wedges. **(b)** Rings.

Figure 4.5: Panek and Vohradsky neighbourhood segment description. From [64].

The method has been implemented and tested in §4.5.

Miller, Appel et al.

The work by Miller et al. [60] and Appel et al. [3] is another example of neighbourhood based matching. Spots are matched one by one and the spot in question is called the primary spot. The neighbour spots within a certain distance from the primary spot is termed the secondary spots or *cluster* of the primary spot. A probabilistic criterion is used to determine the correct match. The probability of two clusters to mismatch (match by random) is approximated and if this probability is less than 0.01, the two clusters are considered

a match. To further reduce the risk of mismatch, the *secondary* clusters (i.e., the clusters of the secondary spots) are compared.

4.3.2 Graph based matching

Among the approaches using graph matching, Watanabe et al. [88] has reported good results on matching spots in restriction landmark genome scanning (RLGS) electrophoresis gels for the purpose of high-speed genome scanning. Although these are not protein spots, the images have some similarities to the 2DGE images discussed here. They use an elliptic operator to detect the spots in X-ray images. For the matching, a relative neighbourhood graph (RNG) of the reference pattern is matched to a Delaunay net (DN) of the match pattern using a breadth first search progressing from initially given corresponding points.

4.3.3 Successive point matching

In situations where more information than the spot centre location (x, y) is available, it seems natural to exploit that extra information in the matching process. It seems likely, that other spot attributes as percent integrated intensity (%II) or spot area can be used to reduce the matching effort and increase the robustness of the matching. This assumes, of course, that corresponding spots, besides having similar (x, y)-coordinates ((pI, MWt) properties), also have similar intensity or area values. Although this should be true for the majority of spot pairs, this is not always the case in particular not for the interesting marker proteins, and therefore caution must be taken when attribute information is used. The above considerations also apply to the approach in § 4.2.4.

Hoffmann et al. [40] present the idea of ranking spots from both sets according to their integrated intensity and then match spots with similar rank only. The idea is to match most important (intense) spots first and then, in an iterative manner, proceed with the matching of less important spots. Please refer to Tab. 4.2 for more references to work by this group.

Discretisation

[40] propose a heuristic method to assign a *discrete* intensity level to each spot based on its continuous integrated intensity value, i.e., a binning of the continuous integrated intensities into a number (L_{max}) of discrete intensity levels, $[1, 2, \ldots, L_{max}]$.

Given a set of J spots $\hat{\mathcal{P}} = (\mathcal{P}, a_p)$. $\mathcal{P}$ is the well-known set of spot centre coordinates and a_p is a vector of J real spot intensity values, one for each spot. Assign a discrete intensity level to each spot so that:

- half of the spots[1] ($J/2$) are assigned one of the top $L_{max}/2$ discrete levels,
- the top $L_{max}/2$ discrete levels contain equally many spots, and
- the sum of spot intensities in the bottom half discrete levels is the same.

The algorithm

The idea behind successive spot matching is, in a successive way to match spots from $\hat{\mathcal{P}}$ to spots in $\hat{Q} = (\mathcal{Q}, a_q)$ that differs no more than *two* discrete intensities from spots in $\hat{\mathcal{P}}$. Let $\hat{\mathcal{P}}_c = \hat{\mathcal{P}}(l = l_c)$ denote the spots in $\hat{\mathcal{P}}$ with discrete intensity level $l = l_c$ (centre level) and $\hat{\mathcal{Q}}_c = \hat{\mathcal{Q}}(l_{min} \leq l \leq l_{max})$ denote the spots in $\hat{\mathcal{Q}}$ with discrete intensity level l in the range l_{min} to l_{max}. The successive matching as proposed in [40] is outlined in Alg. 6.

Algorithm 6 SUCCESSIVE SPOT MATCHING($\hat{\mathcal{P}}_c$, $\hat{\mathcal{Q}}_c$)

1: Initialise $L_{min} = 1$, $L_{max} = 10$.
2: **for** $l_c = L_{max}$ downto L_{min} **do**
3: $l_{min} = \max(L_{min}, l_c - 2)$ and $l_{max} = \min(L_{max}, l_c + 2)$
4: $\hat{\mathcal{P}}_c = \hat{\mathcal{P}}(l = l_c)$ and $\hat{\mathcal{Q}}_c = \hat{\mathcal{Q}}(l_{min} \leq l \leq l_{max})$
5: match $\hat{\mathcal{P}}_c$ to $\hat{\mathcal{Q}}_c$
6: update global match matrix
7: **end for**

Alg. 6 is explained in further detail. Line 1 is initialisation of the maximum and minimum discrete levels. Line 2 to 7 is the main loop where the centre level l_c is decreased from L_{max} to L_{min}. For the current centre level l_c the lower and upper limits for the discrete levels (used for selection from $\hat{\mathcal{Q}}$) is calculated in line 3. Line 4: $\hat{\mathcal{P}}_c$ is the subset of $\hat{\mathcal{P}}$ with discrete intensity level $l = l_c$ and $\hat{\mathcal{Q}}_c$

[1] Hoffmann et al.[40] set this number to 500 but does not mention the total number of spots. Assigning a fraction, e.g., half, of the spots to the top levels is a more general rule.

is the subset of $\hat{\mathcal{Q}}$ with discrete intensity levels in the range l_{min} to l_{max}. The cardinality of $\hat{\mathcal{Q}}_c$ is usually much larger than the cardinality of $\hat{\mathcal{P}}_c$ because $\hat{\mathcal{P}}_c$ contains spots from 1 level only and $\hat{\mathcal{Q}}_c$ contains spots from 3 up to 5 levels.

In line 5 the two subsets P_c and Q_c are matched using the some matching method robust to a large number of outliers in at least one of the sets, e.g., from the family of Robust Point Matching methods (§ 4.2.4).

Instead of an RPM method [40] use incremental Delaunay triangulation and a 2-step hashing technique to match the spots.

In line 6 the global match matrix is updated with the pairs found from the matching of $\hat{\mathcal{P}}_c$ and $\hat{\mathcal{Q}}_c$.

Results of discretisation

The discretisation of the point sets has an undesired side effect, namely a number of irrecoverable singles that will never be matched. As mentioned, for each centre level l_c, the two point subsets $\hat{\mathcal{P}}_c$ and $\hat{\mathcal{Q}}_c$ are of quite different size. $\hat{\mathcal{Q}}_c$ is chosen to contain points with centre level $l_c \pm 2$ levels in order to ensure that all points in $\hat{\mathcal{P}}_c$ have a "partner" in $\hat{\mathcal{Q}}_c$. This is, however, not always the case. $\hat{\mathcal{P}}_c$ can have quite a few singles despite the efforts to avoid this. These singles, introduced by the discretisation, can inherently *never* be matched to their corresponding points in $\hat{\mathcal{Q}}$ by Alg. 6, at least not without some post-processing matching step.

A small experiment using point and attribute data from gel 1A and gel 2A as $\hat{\mathcal{P}}$ and $\hat{\mathcal{Q}}$ respectively (see § C.1.1) shows the number of irrecoverable singles in $\hat{\mathcal{P}}_c$ for each centre level l_c. Tab. 4.3 shows the discretisation effects of two spot sets $\mathcal{P}$ and $\mathcal{Q}$ with 1919 and 1918 spots respectively. Note, that the number of singles in $\hat{\mathcal{P}}_c$ exceeds 25 at several discretisation levels. The total number of singles in $\hat{\mathcal{P}}_c$ for all levels is 183 or almost 10% of the total number of spots.

The results above show that caution should be exercised when matching discretised point sets. The spots representing the extra singles, introduced by the discretisation, are often *particularly interesting* from a biological point of view because they vary in %II. Such variations in protein levels across different samples can be used to monitoring of diseases development (§ 2.1.1) and should not be overlooked.

Extensions to the successive point matching approach could include

l_c	10	9	8	7	6	5	4	3	2	1
l_{min}	8	7	6	5	4	3	2	1	1	1
l_{max}	10	10	10	9	8	7	6	5	4	3
$\hat{\mathcal{P}}_c$	192	192	192	192	192	98	120	145	189	407
$\hat{\mathcal{Q}}_c$	576	768	960	864	785	731	724	958	862	749
# pairs	190	188	186	169	161	80	95	128	161	378
# singles in $\hat{\mathcal{P}}_c$	2	4	6	23	31	18	25	17	28	29
# singles in $\hat{\mathcal{Q}}_c$	386	580	774	695	624	651	629	830	701	371

Table 4.3: Results of protein spot set discretisation. Top row: centre discrete intensity level l_c. 2nd and 3rd row: lower and upper limit for discrete intensity level range. 4th row: number of spots in $\hat{\mathcal{P}}_c = \hat{\mathcal{P}}(l_c)$. 5th row: number of spots in $\hat{\mathcal{Q}}_c = \hat{\mathcal{Q}}(l_{min} : l_{max})$. 6th row: number of pairs at discrete intensity level l_c. 7th row: number of singles (outliers) in $\hat{\mathcal{P}}_c$. Bottom row: number of singles (outliers) in $\hat{\mathcal{Q}}_c$.

- a post-processing step to catch up on singles introduced by the discretisation,
- exclusion of already matched spots from $\hat{\mathcal{Q}}$. If a spot in $\hat{\mathcal{Q}}$ has been matched to a spot in $\hat{\mathcal{P}}$ at an earlier discretisation level, there is no need to attempt matching of this spot again, and
- use of even more attribute information. It would be desirable to be able to make the attributes discrete in a manner that would leave as few outliers as possible in $\hat{\mathcal{P}}_c$ at every level. Often more attributes than the integrated intensity (actually %II, as used here) are available and. Discretisation of a (linear) *combination* of the attributes available could perhaps reduce the number of singles in the $\hat{\mathcal{P}}_c$'s. Attributes available could be, spot area, spot peak intensity, background intensity, percentage integrated intensity, and percentage Gaussian integrated intensity.

4.3.4 Regionalised robust point matching

The transformation to be recovered when matching two point sets can be complex, especially if the point sets are large. Most methods for general point pattern matching described earlier have difficulties matching large point sets with complex distortion.

A regionalised approach, as proposed here, divides a large matching problem into several smaller (local) problems and then successively solves the sub problems. Finally, the individual sub results are merged to form the result of the large problem.

The regionalisation is a general approach that can be used with *any* point matching method, robust to a large number of outliers. Here, the RPM methods from § 4.2.4 will be applied. The regionalised point matching with RPM-affine was proposed in Pedersen and Ersbøll [65].

Gel regions

Definition 8 (Region) Given a gel Ω_p containing protein spots represented by their spatial locations $\mathcal{P}$. A region ω_p is a square area inside the borders of the gel so that $\omega_p \subset \Omega_p$. The size of ω_p is $(s_p \times s_p)$ and its centre coordinates are (x_r, y_r) relative to the origin of Ω_p. The J^r points inside the borders of ω_p are denoted $\mathcal{P}^r$ so that $\mathcal{P}^r \subset \mathcal{P}$. □

Definition 9 (Region Pair) Given two gels Ω_p and Ω_q containing point sets $\mathcal{P}$ and $\mathcal{Q}$ respectively. Define a *pair* of square regions (ω_p, ω_q) both with centre coordinates (x_r, y_r) relative to their gel origin. ω_p is of size $(s_p \times s_p)$ and ω_q is of size $(s_q \times s_q)$. The number of points in the two sub point sets $\mathcal{P}^r$and $\mathcal{Q}^r$are of size J^r and K^r respectively.

For a region pair (ω_p, ω_q), furthermore denote the regional $(J^r + 1) \times (K^r + 1)$ match matrix $\hat{m}^r$ and the regional transformation f^r. □

Note that ω_p and ω_q, from two different gels, can be of different sizes although they are still centred around the *same* coordinates.

Region sizes

The following contains some general considerations on the choice of region sizes s_p and s_q.

The size of the square regions should be sufficiently large to ensure a "suitable" number of points inside the region (enough points to avoid ill-posed problems in the computation of f^r). Also, if the region becomes too large the regional transformation f^r will maybe not to be able to "capture" all the motion in the area covered by the region. For all point matching algorithms the regional computation time depends on the number of points to match so this may also be a reason to not have too large regions.

As mentioned earlier, the regions in the region pair $(\omega_p^{rc}, \omega_q^{rc})$ are centred around the same spatial coordinates. Provided that the deformations are not too large,

this ensures that the two sub sets $\mathcal{P}^r$ and $\mathcal{Q}^r$ contain a large number of homologous points.

With *equal* region sizes ($s_p = s_q$) both $\mathcal{P}^r$ and $\mathcal{Q}^r$ can have singles due to the local deformations. These *false* singles are not real singles actually in $\mathcal{P}$ or $\mathcal{Q}$ but they have been introduced because of the regionalisation.

By choosing *different* region sizes for ω_p and ω_q so that ω_q is always larger than ω_p ($s_q > s_p$), it can be guaranteed that the smaller point set $\mathcal{P}^r$ contains no false singles. This will on the other hand introduce a number of extra *false* singles in $\mathcal{Q}^r$. In other words let

$$s_q = s_p + s_b \tag{4.20}$$

where the buffer size $s_b > 0$. The choice of letting ω_q be larger than ω_p is arbitrary. If s_b is selected larger than the largest distance between any two homologous points it is guaranteed that there are no so called "false" singles in $\mathcal{P}^r$. The largest distance between any two homologous points can not be known since the correspondence is unknown, so it must be estimated to select a suitable value of s_b.

The number of false singles in $\mathcal{Q}^r$ can be large depending on the buffer size s_b, but a point matching method that is robust to a large number of singles can handle this.

Fig. 4.6 shows the principle of the regional point matching. To the top left is Ω_p with $\mathcal{P}$ and a small square marking ω_p centred around (x^r, y^r) and similarly, to the top right is Ω_q with $\mathcal{Q}$ and a *little larger* square marking the corresponding region ω_q, also centred at (x^r, y^r). A zoom of the region in question is shown at the bottom with spots from both gels together. The solid line shows the outline of the small region, ω_p with $\mathcal{P}^r$ (the grey spots) inside and the dashed line shows the border of the large region, ω_q with $\mathcal{Q}^r$ (the white spots) inside. Homologous spots in $\mathcal{P}^r$ and $\mathcal{Q}^r$are connected with lines.

Overlapping regions

Having specified how to define a region *pair* (ω_p, ω_q) one question remaining is how to apply the proposed regional point matching to the *entire* gel areas, Ω_p and Ω_q. The ideal, but computationally unsuitable way would be to slide the region pair around the gel area and then solve the matching problem for all possible positions of the region centres. This would reduce possible side effects of dividing the problem into sub problems but also it would result in an immense and often redundant number of regional matches.

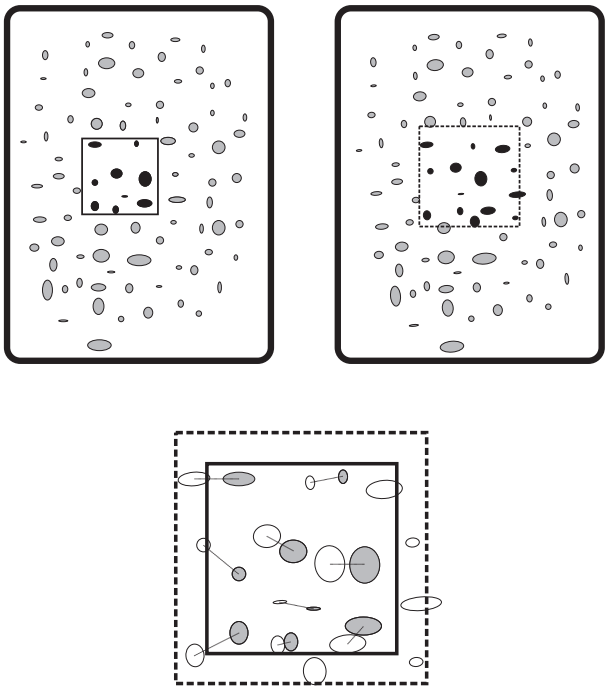


Figure 4.6: Principle of the regional point/spot matching. To the top left is Ω_p with $\mathcal{P}$ and a $(s_p \times s_p)$ square marking ω_p centred at (x^r, y^r). To the top right is Ω_q with $\mathcal{Q}$ and a $((s_p + s_b) \times (s_p + s_b))$ square marking the corresponding region ω_q also centred at (x^r, y^r). In the bottom is shown a zoom of the region in question. The solid line shows the small region, ω_p with $\mathcal{P}^r$ (the grey spots) inside and the dashed line shows the border of the large region, ω_q with $\mathcal{Q}^r$ (the white spots) inside. The lines connect homologous spots.

It is therefore proposed to position the region centres on a *regular equidistant grid* so that neighbour regions overlap with a certain ratio. Now, for a moment, consider a single gel.

Definition 10 (Regionalisation) The division of an area Ω into a set of square (possibly overlapping) regions $\{\omega^{rc}\}$ with the horizontal and vertical distance between neighbour region centres d_c is a *regionalisation* of the area. ω^{rc} is the specific region at the grid position (r, c), i.e., in row r and column c of the grid (see also Fig. 4.9). □

Specification of the region overlap

For $d_c \geq 2s_p$, where s_p is the region size, there is no overlap between neighbouring regions. This is of course not valid if all points (the entire gel area) need to be processed. For $d_c \to \epsilon, \epsilon > 0$ the situation with a sliding region is approached, but this extreme is computationally unsuitable. Fig. 4.7 defines the distances.

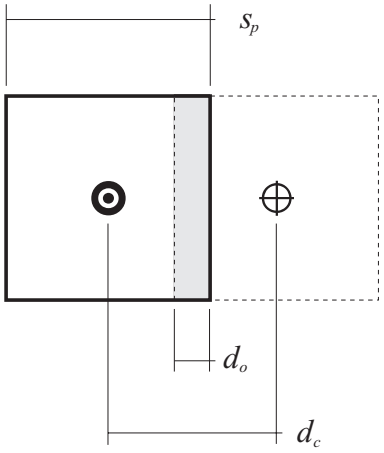


Figure 4.7: Neighbouring regions. Main region (bold quadratic) with right nearest neighbour region (dashed quadratic). Region size s_p, neighbour region distance d_c, and overlap distance d_o. The region overlap is shaded.

With the region overlap distance

$$d_o = \begin{cases} s_p - d_c & \text{if } d_c \in [0; s_p] \\ 0 & \text{otherwise} \end{cases} \tag{4.21}$$

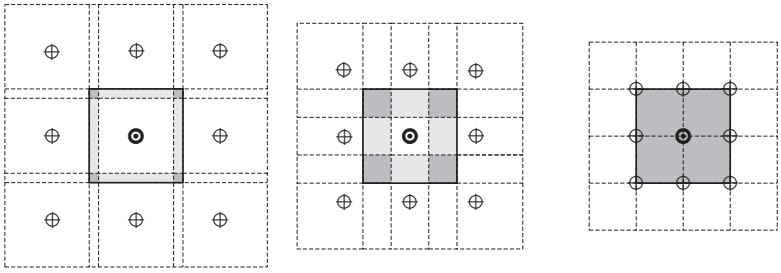
the neighbour region overlap is defined:

Definition 11 (Region overlap ratio) Given a quadratic region of region size s_p. The region overlap ratio is the amount of the region area overlapped by each of the *four* nearest neighbour regions. □

The region overlap ratio can be expressed in simple terms of the region overlap distance as:

$$o \quad = \quad d_o s_p/(s_p s_p) \cdot 100\% \quad = \quad d_o/s_p \cdot 100\%. \tag{4.22}$$

Fig. 4.8 shows the degree of overlap for different values of o. The grey areas are the parts of the centre region overlapped by its four nearest neighbour regions. To the right is seen the case of $o = 50\%$. For this particular choice of overlap each point will be processed 4 times.



(a) 10% **(b)** 30% **(c)** 50%

Figure 4.8: Overlapping neighbour regions. Centre region (bold line square) shown with eight nearest neighbour regions (dashed lines). For (**a**), (**b**), and (**c**) the region size is the same, but the region overlap ratio is 10%, 30%, and 50% respectively. Areas inside the centre region are colour-coded. White area: no neighbour regions overlap so only the centre region covers this area. Light grey: one neighbour region overlaps this area, i.e., a total of two regions cover this area. Dark grey: three neighbour regions overlap this area. Including the centre region, four regions cover this area. For the rightmost example where $o = 50\%$ the entire centre region is covered by four regions.

The algorithm

The regionalisation is a general approach that can be used with *any* point matching method, robust to a large number of outliers. The robust point matching (RPM) methods presented earlier (§4.2.4) fulfil this requirement.

The basic idea is, that equipped with the RPM methods, the matching problem can be solved for each of the region pairs $(\omega_p^{rc}, \omega_q^{rc})$ in the grid and then gather

the regional correspondence information (the match matrices, $\tilde{m}^{rc}$) into a global match matrix, $\tilde{m}$.

In the following a regular grid with equidistant (d_c) nodes (region centres) and overlap ratio $o = 50\%$ is assumed. Let R be the number of rows in the grid and let C denote the number of columns. Fig. 4.9 shows the regionalisation grid for this configuration. $(\omega_p^{rc}, \omega_q^{rc})$ denotes the region pair at position (r, c) in the grid.

Furthermore the regions size for ω_p is s_p. The region size for ω_q is somewhat larger: $s_q = s_p + s_b$.

The REGIONALISED RPM(Ω_p, Ω_q, $\mathcal{P}$, $\mathcal{Q}$, T_0, s_p, s_q, o, R, C) algorithm is outlined in Alg. 7.

The global match matrix $\tilde{m}$ is initialised to the zero matrix and the two gels Ω_p and Ω_q are divided into regions. For all positions on the grid the points in the region pairs $(\omega_p^{rc}, \omega_q^{rc})$ are matched using a Robust Point Matching method.

A simple voting strategy is used to transfer the regional matches ($\tilde{m}_{rc}$) to the global match matrix $\tilde{m}$. This is done by adding the entries in $\tilde{m}_{rc}$ to the corresponding entries in $\tilde{m}$. Because of the chosen overlap ratio $o = 50\%$ each point enters the matching process *four* times in total. $\tilde{m}_{rc}$ has real entries in $[0, 1]$ so the global match matrix, $\tilde{m}$ will, after matching of all region pairs, contain entries in $[0, 4]$.

Algorithm 7 REGIONALISED RPM(Ω_p, Ω_q, $\mathcal{P}$, $\mathcal{Q}$, T_0, s_p, s_q, o, R, C)

```
Initialise m̃ ← 0
Regionalise Ω_p                          (R × C) grid of (s_p × s_p) square regions
Regionalise Ω_q                          (R × C) grid of (s_q × s_q) square regions
for all r such that 1 ≤ R do
  for all c such that 1 ≤ C do                    match region pair (ω_p^rc, ω_q^rc)
    Extract P^rc and Q^rc
    [m̃^rc, f^rc] ← ROBUST-POINT-MATCHING(P^rc, Q^rc, T_0)
    update m̃ with m̃^rc
  end for
end for
return m̃
```

Choice of transformation and region size

The choice of transformation in the RPM method and the choice of region size is related. In general, the affine method is faster, but of course not as flexible as

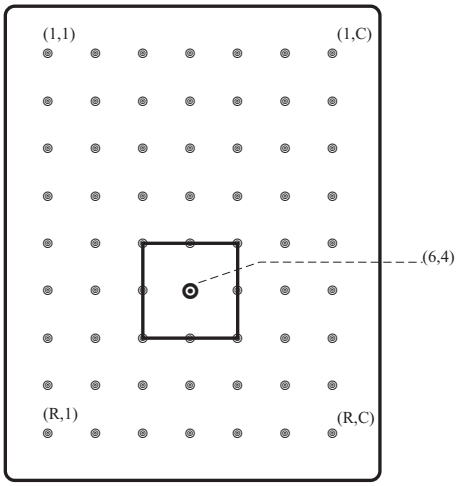


Figure 4.9: Regionalisation grid, $o = 50\%$. The region in the 6th row and the 4th column, ω^{64} is marked with bold lines. The gel Ω_p has been regionalised into $R \cdot C$ regions. For most regions, only region centres are marked and for the region ω^{64} the region borders are shown.

the thin-plate spline based method. The affine transformation (§ 4.2.4) cannot capture complex local deformations and therefore the region size should not be too large. On a 2048×2048 gel pair with approximately 1900 points in each point set, experiments with real spot location data (§ C.2) have shown good results setting $s_p \leq 128$.

The thin-plate spline transformation (§ 4.2.4) can handle almost arbitrarily complex local deformations so the region size does not matter in this case. However, the computational cost can become very high for large point sets and this limits the region size. Matching results with the same data as mentioned above have indicated that $s_p \leq 256$ is reasonable for this method.

The regional robust point matching approach has been applied to matching of 2D gel electrophoresis spot location centres using the two different transformations for the motion, namely the affine transformation. The results of these experiments can be found in § 4.5.

4.4 Match Evaluation

The matching result, i.e., the correspondence estimate of two point sets $\mathcal{P}$ and $\mathcal{Q}$ is defined by the match matrix $\hat{m}$. This section describes an evaluation scheme

of the match matrix, given the *ground truth* correspondence. The ground truth correspondence is defined by the matrix $\hat{M}$, usually established by an operator.

For the match methods returning a fuzzy match matrix $\tilde{m}$ initial binarization of this matrix to obtain $\hat{m}$ is necessary. Given $\tilde{m}$ and a threshold value τ Alg. 8 provides a method to do the binarization so that draws are resolved. Values in $\tilde{m}$ equal to or above the threshold, τ will result in a "1" in $\hat{m}$ and values below τ will result in a "0" in $\hat{m}$. Clearly, the threshold value τ is important. High values of τ results in a conservative binarization where, in cases of doubt, it is preferred to mark a point as an outlier instead of risking a false match. For low values of τ a number of draws can occur in $\hat{m}$. If more than one element in a row or a column is "1" a *draw* has occurred. Alg. 8 resolves the draws in a conservative manner to avoid false matches, i.e., if in doubt, it is preferred to classify a point as an outlier.

The evaluation of the match results is a comparison of the two binary match matrices $\hat{m}$ and $\hat{M}$. For a point 4 natural types (or classes) of the match result can be identified.

- $\mathbf{T_1}$ A pair point *correctly* matched to another pair point.
- $\mathbf{T_2}$ A single point *correctly* detected as single.
- $\mathbf{T_3}$ A pair point or a single *incorrectly* matched with another. pair point or with a single.
- $\mathbf{T_4}$ A pair point *incorrectly* detected as single.

A point belongs to exactly one of these types.

$\mathbf{T}_1$ and $\mathbf{T}_2$ are successful results and $\mathbf{T}_3$ and $\mathbf{T}_4$ are faults. For both point sets $\mathcal{P}$ and $\mathcal{Q}$, the number of points classified as each type is counted, T_1, T_2, etc. and these are used to define 4 scores to evaluate the results:

$$S_1 = \frac{T_1 + T_2}{T_1 + T_2 + T_3 + T_4} \tag{4.23}$$

$$S_2 = \frac{T_1 + T_2 - T_3}{T_1 + T_2 + T_3 + T_4} \tag{4.24}$$

$$S_3 = \frac{T_3}{T_1 + T_2 + T_3 + T_4} \tag{4.25}$$

$$S_4 = \frac{T_4}{T_1 + T_2 + T_3 + T_4} \tag{4.26}$$

S_1 is the fraction correct matches *and* singles and S_2 is the same, but adjusted by subtracting the number of serious faults, T_3 in the nominator. S_3 is fraction incorrect pairs and S_4 is fraction incorrect singles.

For a matching to be successful, S_1 and S_2 should be as high as possible, and S_3 and S_4 should be as small as possible.

4.5 Experiments and Results

This section presents experiments and results of different point pattern matching methods applied to a number of electrophoresis gel data sets. The methods are:

$\mathbf{M_1}$ Pánek and Vohradský's method [64],

$\mathbf{M_2}$ regionalised RPM (affine transformation). Region size $s_p = 128$, and buffer size $s_b = 32$ (for RPM parameters, see § 4.2.4, p.92), and

$\mathbf{M_3}$ regionalised RPM (thin-plate spline transformation). Region size $s_p = 256$, and buffer size $s_b = 32$ (for RPM parameters, see § 4.2.4, p.94)

all methods previously described in this chapter. The data set is described in § C and the 15 experiment combinations (gel pairs) specified in Tab. 4.4 have been used. The correspondence has been established by a skilled operator for all combinations of gels. This correspondence is denoted the *ground truth* correspondence although there might be a few errors due to human mistakes.

		Group 1			Group 2			Group 3		
		A	B	C	A	B	C	A	B	C
Group 1	A		x		x	x		x	x	
	B				x	x		x	x	
	C									
Group 2	A					x		x	x	
	B							x	x	
	C									
Group 3	A								x	
	B									
	C									

Table 4.4: Experiment specification. Group 1, 2, and 3. Matching experiments marked with 'x'. E.g., gel 1A vs. gel 1B, gel 1A vs gel 2A, etc.

The gel pair, gel 1A vs. gel 2A will be used as example in the following and Fig. 4.10 shows the gel images with known spot centres ($\mathcal{P}$ and $\mathcal{Q}$) overlaid. The known correspondence (the expert established correspondence) is shown in Fig. 4.11. This correspondence is aimed to be recovered using the point pattern matching methods $\mathbf{M_1}$, $\mathbf{M_2}$, and $\mathbf{M_3}$.

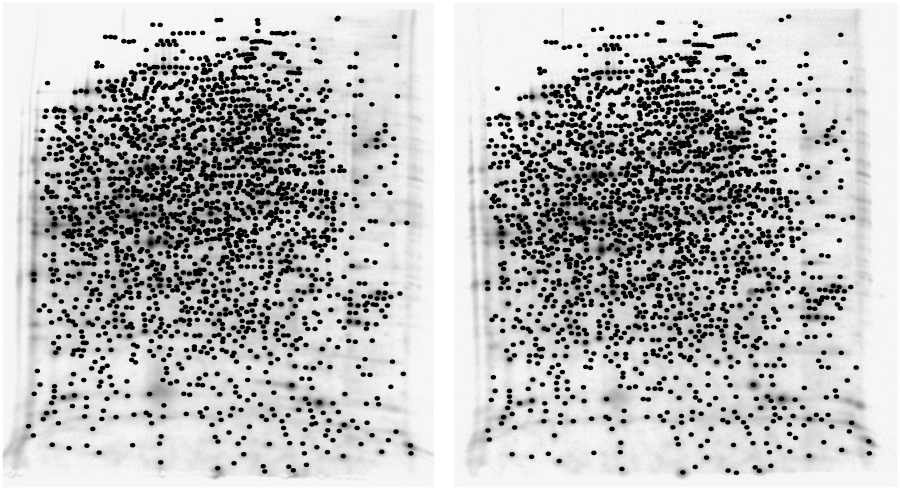


Figure 4.10: Gel images with known spot centres overlaid as points. Left: gel A (1919 spots), right: gel B (1918 spots). 1918 spots in common, which means that one spot present in gel A is missing in gel B.

4.5.1 The trade-off resulting from binarization

The match matrix binarization introduces an important trade-off in the evaluation which will be studied in the following.

For $\mathbf{M_1}$, the match matrix is already binary and no binarization is needed.

For the regionalised matching methods, $\mathbf{M_2}$ and $\mathbf{M_3}$ an overlap ratio, $o = 50\%$ has been used in the following experiments. In § 4.3.4 it is shown that with this choice of overlap ratio $\mathbf{M_2}$ and $\mathbf{M_3}$ return match matrices with real elements in $[0, 4]$. For these methods, it is necessary to choose τ.

Fig. 4.12 shows how the $\mathbf{M_2}$ match scores vary for different values of the binarization threshold τ. This is the results of matching gel 1A with gel 2A using the *regionalised robust point matching with affine transformation*, $\mathbf{M_2}$. The higher the value of τ, the more trustworthy the matches of the point pairs. As a consequence the number of correct matches decreases as the threshold is increased but so will the number of false (incorrect) matches. In other words, there is a *trade-off* between having many correct matches or few false matches. If no (or very few) incorrect matches can be tolerated τ has to be chosen large. This behaviour is common for matching methods returning fuzzy match matrices.

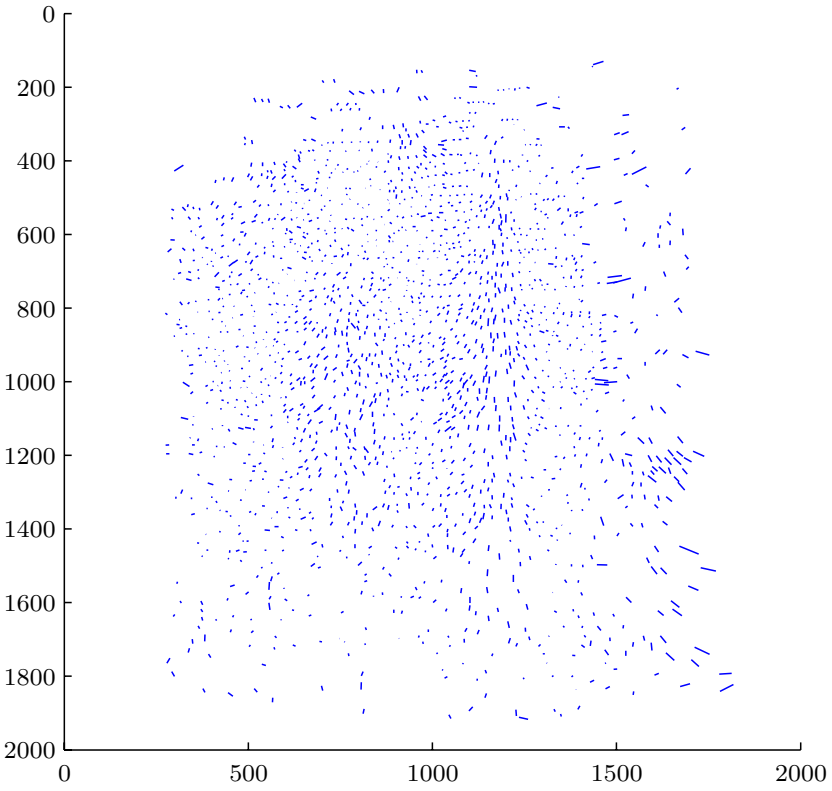


Figure 4.11: Known correspondence between $\mathcal{P}$ and $\mathcal{Q}$ (for the gels shown in Fig. 4.10).

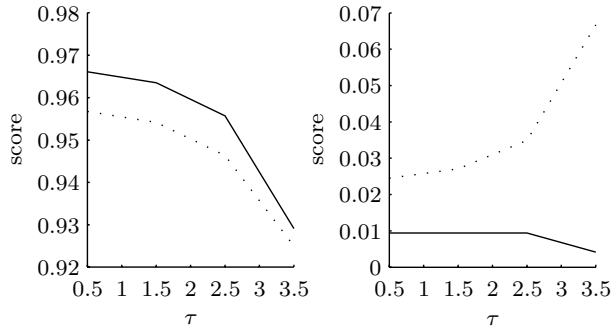


Figure 4.12: Binarization effect. $\mathbf{M_2}$ scores for point set $\mathcal{P}$ at different levels of the binarization threshold τ. Left: Solid line: S_1, dotted line: S_2. Right: Solid line: S_3, dotted line: S_4. The scores S_1 and S_2 declines as τ increases. S_3 (false pair points) becomes very small as τ increases at the expense of more false singles (S_4).

Tab. 4.5 shows, together with the ground truth (GT), the type counts and the scores for matching gel 1A with gel 2A using $\mathbf{M_1}$, $\mathbf{M_2}$, and $\mathbf{M_3}$. The binarization threshold, τ has been set very liberally to 0.5. In the similar Tab. 4.6 a more conservative choice of binarization levels is shown. For $\mathbf{M_2}$, $\tau = 3.5$ and for $\mathbf{M_3}$, $\tau = 2.5$. The counts and scores for $\mathbf{M_1}$ are the same in both tables because this method needs no binarization of the match matrix.

	$\mathcal{P}$				$\mathcal{Q}$			
	GT	$\mathbf{M_1}$	$\mathbf{M_2}$	$\mathbf{M_3}$	GT	$\mathbf{M_1}$	$\mathbf{M_2}$	$\mathbf{M_3}$
T_1	1918	1827	1853	1892	1918	1827	1853	1892
T_2	1	1	1	1	0	0	0	0
T_3	-	4	18	11	-	4	18	11
T_4	-	87	47	15	-	87	47	15
check	1919	1919	1919	1919	1918	1918	1918	1918
S_1	1	0.9526	0.9661	0.9865	1	0.9526	0.9661	0.9864
S_2	1	0.9505	0.9567	0.9807	1	0.9505	0.9567	0.9807
S_3	0	0.0021	0.0094	0.0057	0	0.0021	0.0094	0.0057
S_4	0	0.0453	0.0245	0.0078	0	0.0454	0.0245	0.0078

Table 4.5: Evaluation of match result. Gel pair 1A vs. 2A. For $\mathbf{M_2}$ and $\mathbf{M_3}$, the binarization threshold, $\tau = 0.5$. GT is the ground truth.

	$\mathcal{P}$				$\mathcal{Q}$			
	GT	$\mathbf{M_1}$	$\mathbf{M_2}$	$\mathbf{M_3}$	GT	$\mathbf{M_1}$	$\mathbf{M_2}$	$\mathbf{M_3}$
T_1	1918	1827	1782	1859	1918	1827	1782	1859
T_2	1	1	1	1	0	0	0	0
T_3	-	4	8	5	-	4	8	5
T_4	-	87	128	54	-	87	128	54
check	1919	1919	1919	1919	1918	1918	1918	1918
S_1	1	0.9526	0.9291	0.9693	1	0.9526	0.9291	0.9692
S_2	1	0.9505	0.9250	0.9666	1	0.9505	0.9249	0.9666
S_3	0	0.0021	0.0044	0.0026	0	0.0021	0.0042	0.0026
S_4	0	0.0453	0.0667	0.0281	0	0.0454	0.0667	0.0282

Table 4.6: Evaluation of match result. Gel pair 1A vs. 2A. For $\mathbf{M_2}$ $\tau = 3.5$ and for $\mathbf{M_3}$, $\tau = 2.5$. GT is the ground truth.

4.5.2 Method comparison

To compare the performance of the three point matching methods, Fig. 4.13 presents the matching scores for each method in the matching of all 15 experiment pairs. The binarization threshold has been chosen individually for the two RPM methods. For $\mathbf{M_2}$, $\tau = 3.5$ and for $\mathbf{M_3}$, $\tau = 2.7$. Tab. 4.7 shows the average scores across all 15 experiment pairs.

	$\mathbf{M_1}$	$\mathbf{M_2}$	$\mathbf{M_3}$
S_1	0.9502	0.9197	0.9610
S_2	0.9470	0.9134	0.9589
S_3	0.0032	0.0063	0.0021
S_4	0.0467	0.0740	0.0369

Table 4.7: Average scores across 15 experiment pairs.

In general, $\mathbf{M_3}$ seems to outperform both other methods. It has the highest S_1 score in 11 of 15 experiments *and* the lowest S_3 score in 12 of 15 experiments. Considering the false matches (S_3), $\mathbf{M_1}$ is lowest in 3 of 15 experiments, and the method results in a large number of wrong singles yielding weaker S_1 and S_4 scores. $\mathbf{M_2}$ exhibits the poorest performance in all of these experiments. Note however, that except for $\mathbf{M_2}$ in two cases, none of the methods have more than 1% false matches (S_3) and in the main part of the experiments with $\mathbf{M_1}$ and $\mathbf{M_3}$ have less than 0.4% false matches.

4.5.3 Error locations

The three figures (4.14-4.16) show the spatial location of the errors corresponding to the example results in Tab. 4.5. Correct matches are marked with cyan (grey) lines ($\mathbf{T_1}$) and missing matches are shown with blue (black) lines. Wrongly matched points drawn with black (black) lines.

Accumulated errors location

The spatial locations of the errors across all 15 experiments have been recorded for each method. The plot in Fig. 4.17 shows the location of the false matches from all 15 experiments in the same plot. The Figs. 4.18-4.19 show the similar plots for the $\mathbf{M_2}$ and $\mathbf{M_3}$ methods respectively. The general trend is, that the methods seem to fail more often in the border areas of the gels, where geometrical distortions are largest.

The three plots have been plotted together in Fig. 4.20. All methods seem to have problems with the same five groups of spots near the right edge of the gels.

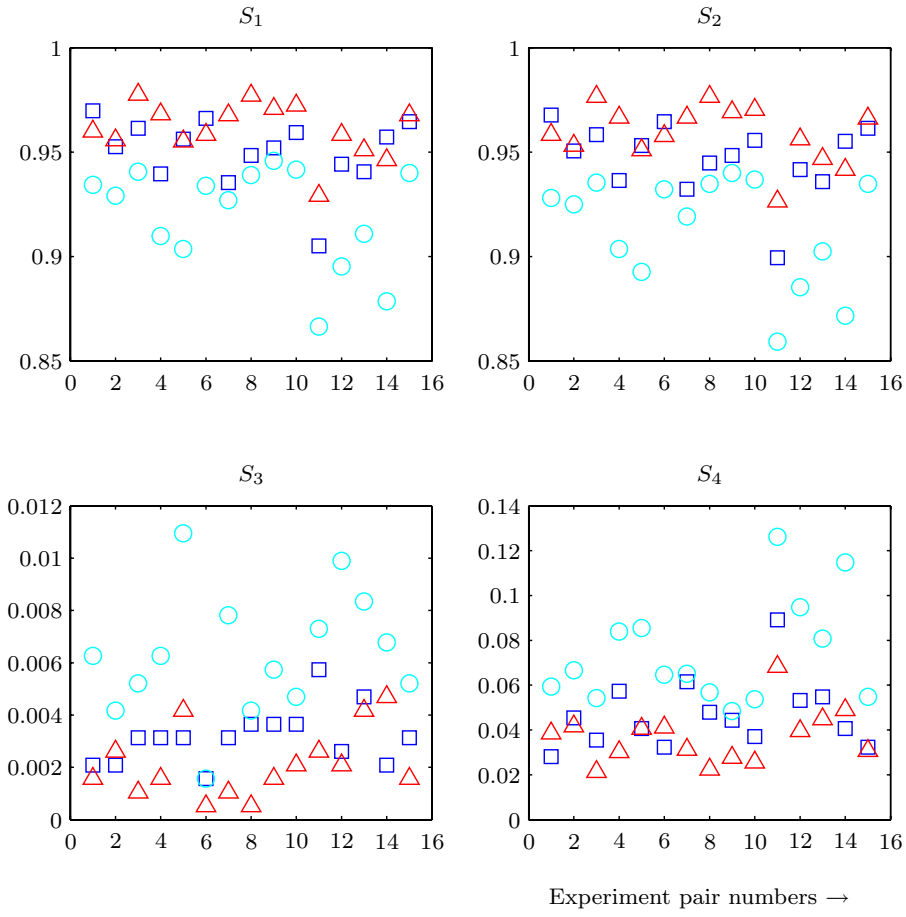


Figure 4.13: Test scores for $\mathbf{M_1}$ (□), $\mathbf{M_2}$ (◯), $\tau = 3.5$, and $\mathbf{M_3}$ (△), $\tau = 2.7$ on all 15 experiment pairs.

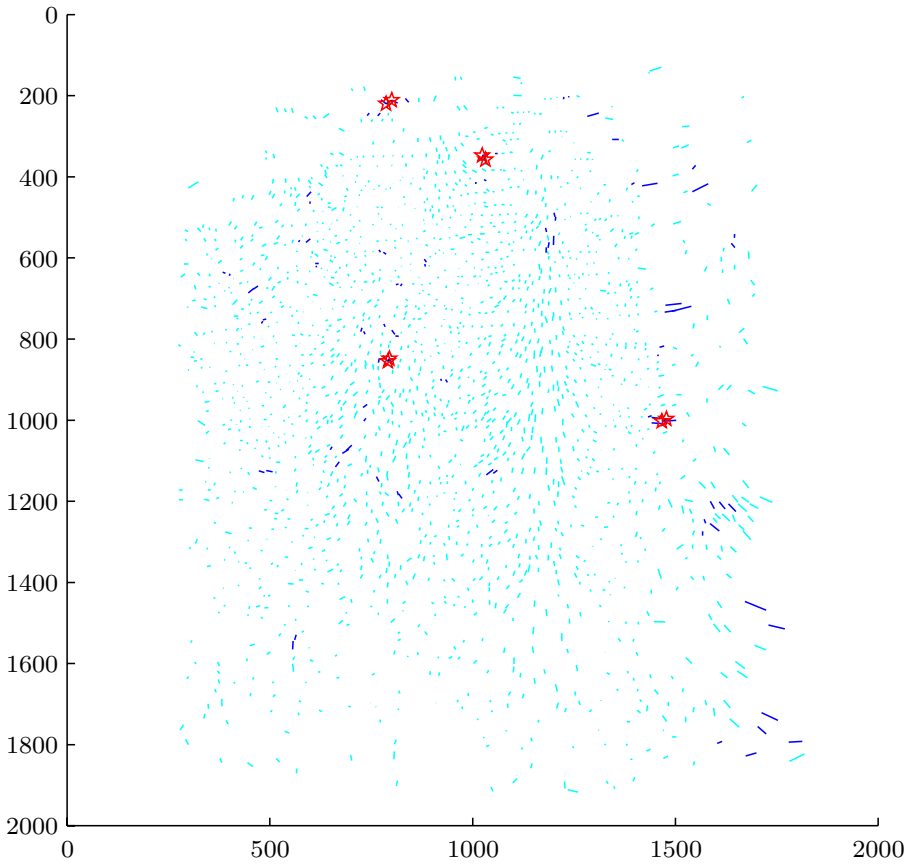


Figure 4.14: Error locations. Method $\mathbf{M_1}$. The cyan (grey) lines with no end markers show the *correct* matches ($\mathbf{T_1}$). The blue (black) lines with no end markers are the *missing* matches ($\mathbf{T_4}$), i.e., where the algorithm should have found a correspondence. The black (black) lines with markers (stars) at the ends show the *wrong* matches ($\mathbf{T_3}$).

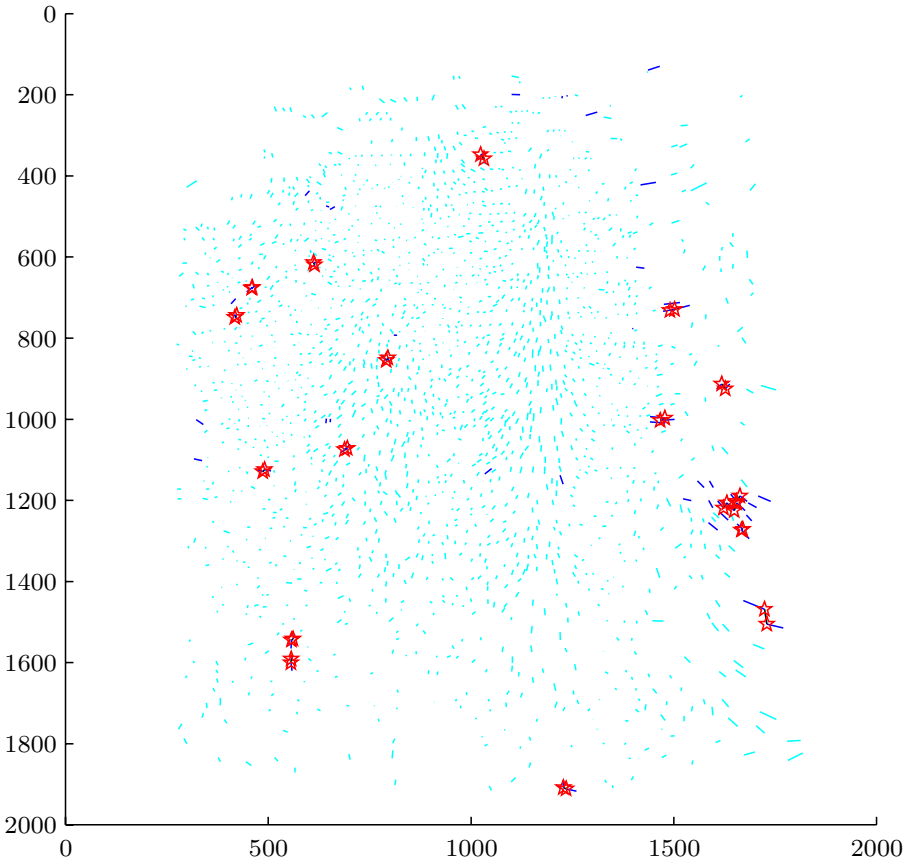


Figure 4.15: Error locations. Method $\mathbf{M_2}$. $\tau = 0.5$. The cyan (grey) lines with no end markers show the *correct* matches ($\mathbf{T_1}$). The blue (black) lines with no end markers are the *missing* matches ($\mathbf{T_4}$), i.e., where the algorithm should have found a correspondence. The black (black) lines with markers (stars) at the ends show the *wrong* matches ($\mathbf{T_3}$).

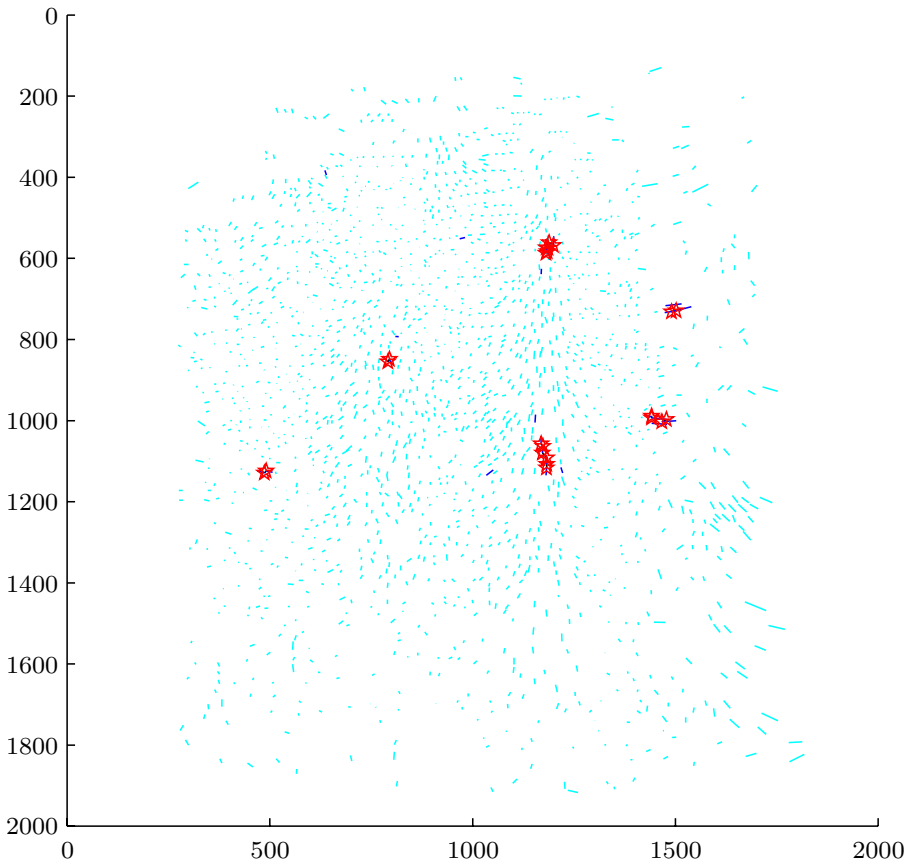


Figure 4.16: Error locations. Method $\mathbf{M_3}$. $\tau = 0.5$. The cyan (grey) lines with no end markers show the *correct* matches ($\mathbf{T_1}$). The blue (black) lines with no end markers are the *missing* matches ($\mathbf{T_4}$), i.e., where the algorithm should have found a correspondence. The black (black) lines with markers (stars) at the ends show the *wrong* matches ($\mathbf{T_3}$).

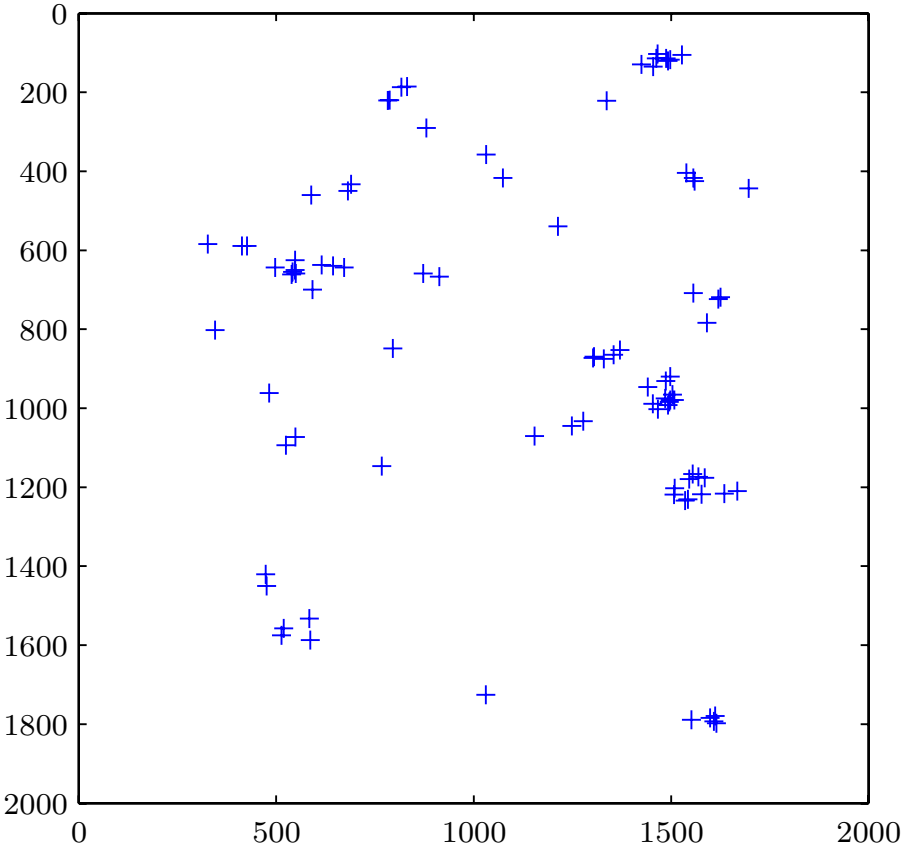


Figure 4.17: Spatial location of errors in all experiments, $\mathbf{M_1}$.

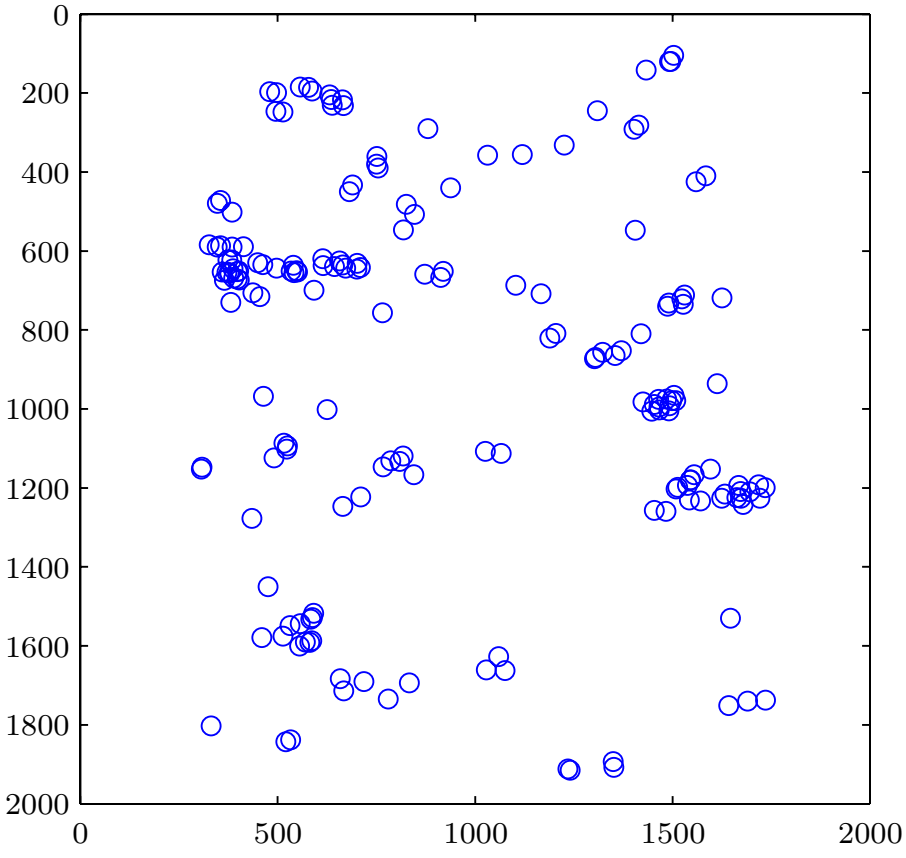


Figure 4.18: Spatial location of errors in all experiments, $\mathbf{M_2}$ ($\tau = 3.5$).

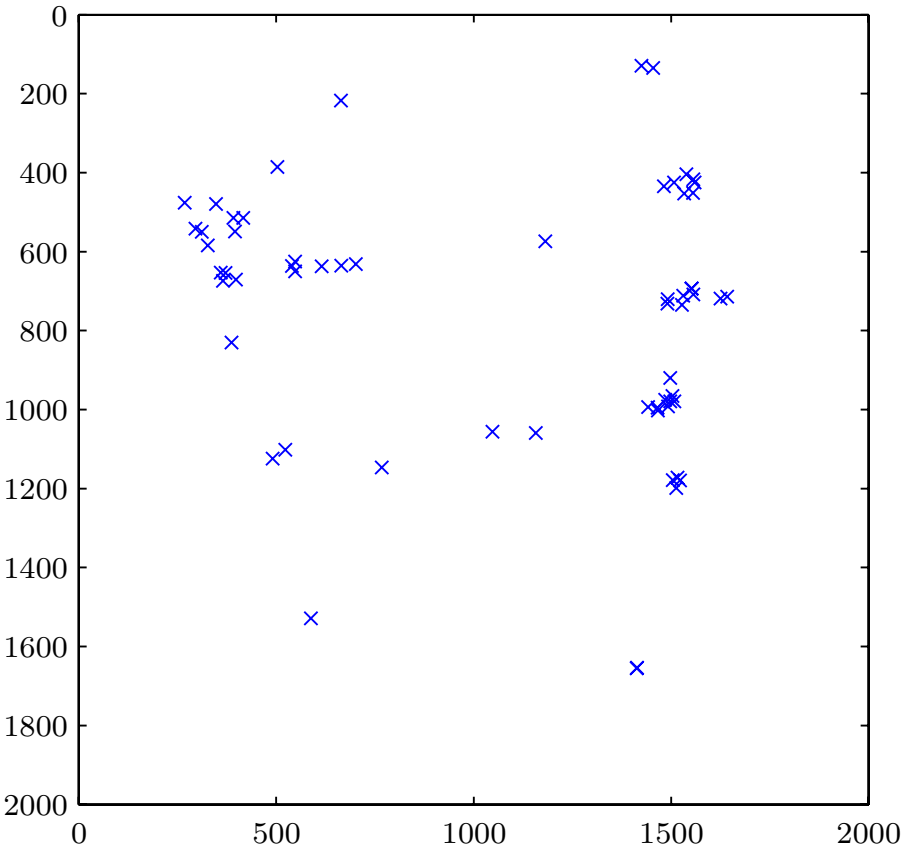


Figure 4.19: Spatial location of errors in all experiments, $\mathbf{M_3}$ ($\tau = 2.7$).

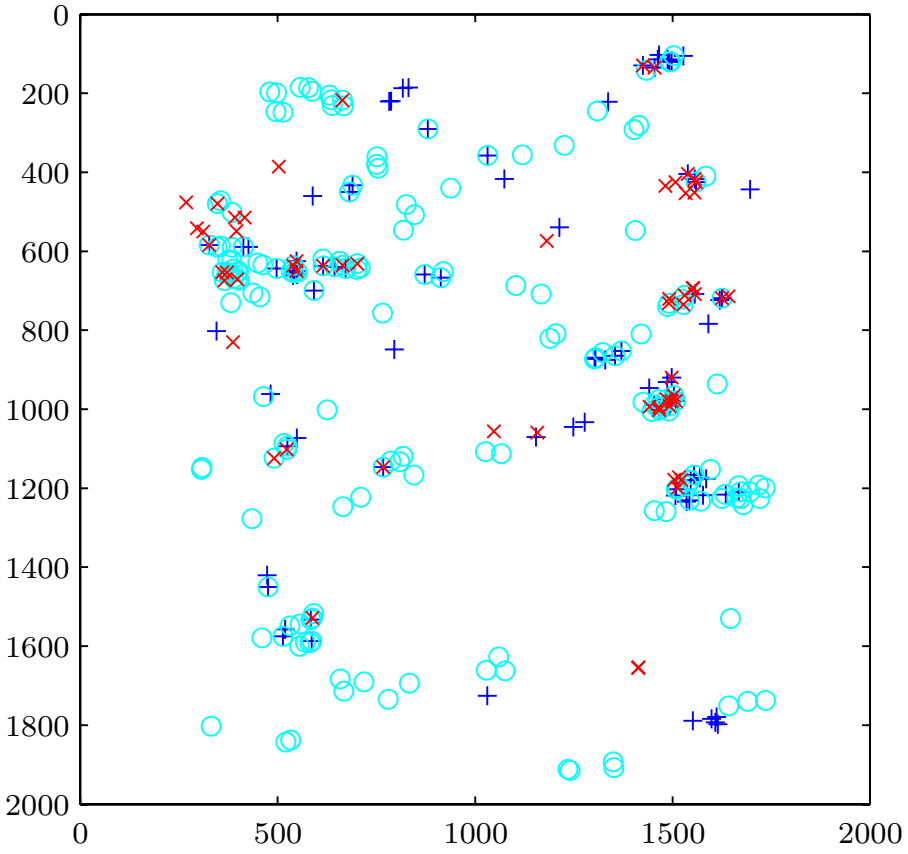


Figure 4.20: Spatial location of errors in all experiments. $\mathbf{M_1}$(+), $\mathbf{M_2}$(o), and $\mathbf{M_3}$(x).

4.6 Summary

After having provided the notation necessary to describe correspondence, match matrices, and motion estimation a range of general point pattern matching methods was described. A group of these, namely the family of robust point matching (RPM) methods was explained in detail.

The general RPM method was extended by a modification of Sinkhorn's matrix normalisation algorithm to improve robustness of the method.

After the detailed description of the special cases of affine and thin-plate spline transformations, the energy function of the latter was extended to take extra point attribute information into account.

The nature of protein spot patterns requires special properties of the point matching algorithm (§ 2.3), but not many of the methods in the literature fulfil these requirements. The robust point matching method with thin-plate spline transformation has been identified as capable of fulfilling all requirements.

The most important, existing methods for point matching of 2DGE spot patterns were presented and a new *regionalised* approach was proposed. This new approach was based on the RPM matching methods, but any point matching method, robust to a large number of outliers/singles could be used.

It was shown, that although the idea is appealing care should be taken if using successive matching of the spots, where most intense spots are matched together first and so on. This is due to the quite dynamic behaviour of the protein expression levels between samples. The variability of a protein's expression level can be so large that too much weight should not be put on the spot intensities when matching the spots.

A method proposed by Pánek and Vohradský [64] and two variants of the regionalised RPM methods were applied to 15 combinations of real 2DGE protein spot data. The matching results of the experiments demonstrated the superiority of the regional RPM method using the thin-plate spline (TPS) transformation. The average S_1 score (correct matches) for this method was 0.9610 (96.1%) with an average S_3 score (false matches) of only 0.0021 (0.21%). An average of 3.7% of the spots were declared to difficult to match and falsely classified as singles.

The second method tested, $\mathbf{M_2}$ was the poorest performing of the three methods. This method used regionalised RPM with affine transformation and since the same method using TPS transformation had quite good performance, the

results indicate, that the affine transformation was not able to properly capture the motion (deformation) within each region. Even though the region size s_p was only half the region size for the M_3 method. By further lowering the region size, the local deformation inside each region should approach the affine transformation and could maybe improve the performance of this method.

Dependent on an appropriate choice of binarization threshold for the fuzzy match matrix, the regionalised RPM-TPS ($\mathbf{M_3}$) was able to match 98.7% of the points correctly while making only 0.6% false matches in the matching of two typical sets of points with 1900+ points. The remaining fraction was false singles. By lowering the binarization threshold for $\mathbf{M_3}$, the fraction of correctly matched pairs could be increased at expense of more false matches, or by raising the threshold false matches could be avoided at the expense of fewer correct matches and more false singles.

From a practical viewpoint it is better to have a few extra false singles which are easier to find by eye than a few incorrect matches which are "hidden" by all the other matches.

The spatial location of errors was investigated and it was shown, that the methods failed more often in the edge areas of the gel, where distortions are known to be large and more complex. For the regional methods, this indicate that the buffer size $s_b = 32$ pixels was maybe chosen too small.

Chapter 5

Elastic Graph Matching

This chapter is confidential and is omitted from this edition of the thesis.

Chapter 6

Conclusion

This thesis addresses the main issues in pattern analysis of two-dimensional gel electrophoresis data. The two-dimensional electrophoresis (2DGE) technique, used for protein separation in proteome analysis, results in grey level intensity images showing thousands of proteins more or less focused in spots.

In the analysis of the 2DGE images, a number of issues have been identified and they divide naturally into two groups:

- issues in image segmentation and
- issues in protein spot matching.

Segmentation of the 2DGE images was shown to be a non-trivial task because of the relatively large number of weak spots, the high spatial density of spots, the large number of overlapping spots, and the varying background across the images.

A number of different approaches to the segmentation of 2DGE images was demonstrated.

H-domes and *marker based watershed segmentation*, both methods from the mathematical morphology, were applied to the segmentation of the gel images. The h-domes technique experienced difficulties segmenting overlapping spots.

When a scale space blob detector was used to define the markers for the marker based watershed segmentation, reasonable results were obtained without the notorious over-segmentation from the watershed segmentation.

Regarding the noisy nature of the 2DGE images, an experiment on a small gel region with *known* locations of protein spots yielded a remarkable result: The image contained non-spot blobs, stronger than the weakest known spots, and also, at the (known) location of some very weak spots, no local minima could be detected.

Parametric modelling of protein spots using a diffusion based spot model gave promising results for relatively isolated protein spots. However, for the large number of spots with close neighbours, there is a need for a *mixture* spot model taking several spots into account. The outlines of such a model was proposed.

The main focus of the thesis was the protein spot matching. For this purpose, a number of point matching methods for *general point matching* and methods designed *specifically for matching patterns of protein spots* were investigated.

A successful matching of two point patterns consists of establishing the correct point correspondence between all point pairs in the sets and identify points that have no homologous point in the other set.

Based on the complex nature of the protein spot patterns, five requirements for point matching methods to solve the correspondence problem were identified. An ideal point matching method must:

- exactly and robustly match protein pairs,
- allow for non-linear distortions/transformations,
- robustly handle outliers in both sets,
- be able to handle point sets of stochastic/amorphous nature, and
- robustly match dense point sets.

The only methods found in the literature of general point pattern matching methods, capable of satisfying all requirements is the family of robust point matching (RPM) methods. These methods were extended and embedded in a regionalised setting to improve robustness, speed and flexibility.

Two variants of regionalised robust point matching (affine and thin-plate-spline) were tested on real 2DGE spot data and among the matching methods for

2DGE protein spot patterns, the most promising method [64] was implemented to provide a level of comparison.

In 15 experiments with matching different pairs of 2DGE spot sets, the regionalised RPM-TPS method proposed here showed to be superior to the other methods tested with average correct matching of 96.1% of the protein pairs and 0.2% wrong matches. By lowering the binarization threshold, the fraction of correctly matched pairs could be increased at expense of more false matches or by raising the threshold false matches could be avoided at the expense of fewer correct matches and more false singles.

By examination of the errors' spatial location it was shown, that the methods tested failed more often in the edge areas of the gels. This could be due to the larger distortion typically found near the gel edges.

As an alternative to the traditional sequential approach to analysis of the 2DGE gels, namely the image segmentation followed by spot matching, a new unified approach was suggested. The elastic graph matching (EGM) exploits prior knowledge of the spot configuration (from the *reference image*) *and* the image information available in the incoming gel (the *match* image). This method has not been extended to its full potential but experiments conducted showed promising results and a wide range of suggestions to improve on the method was given.

The main conclusion from this work is that the task of analysing 2DGE images in a robust and objective way is far from being trivial, but the methods developed here will most likely provide more robust and objective means in the analysis.

Also, the sequential approach based on image segmentation succeeded by spot matching is probably not the most suitable approach, although the most common in commercially available software packages. Mainly because both tasks are performed without using all information available. A unified approach, on the other hand, as the one suggested here (EGM) has the potential of significantly improve the image analysis / pattern recognition part in proteomics and thereby reduce time and labour costs in this part of the process.

Appendix A

Thin-Plate Spline Transformation

The flexibility of the thin-plate spline (TPS) transformation makes it suitable to model the complex disparity patterns of protein spots in two-dimensional electrophoresis gels (2DGE). The literature on the thin-plate spline transformation and its applications in image analysis is vast and the central references include [15, 84] with additional and alternative descriptions in [29, 37, 45, 71, 78].

We will show some properties of the thin-plate spline in two dimensions, but references will be given to extensions to higher dimensions.

The thin-plate spline functional is an extension of the cubic spline in 1 dimension.

Consider two point sets $\mathcal{P} = \{p_i\}$, $i = 1, \ldots, n$, $p_i = \left({}^{p}x_1(i), {}^{p}x_2(i)\right)$ and $\mathcal{Q} = \{q_j\}$, $j = 1, \ldots, n$, $q_j = \left({}^{q}x_1(j), {}^{q}x_2(j)\right)$ of n corresponding points in the two-dimensional space $\mathbb{R}^2$ ($D = 2$). By correspondence we mean that the point sets are pairwise homologous, i.e. $p_k \sim q_k \; \forall k \in 1 \ldots n$. In the literature these points have multiple names: landmarks, fiducial markers etc., dependent on the application. In medical imaging these points often correspond to some anatomical feature in the image, in aerial imagery geographical landmarks are common landmark points. The centre of distinct and significant protein spots can be used as landmarks in images of electrophoresis gels.

In homogeneous coordinates we denote

$$\mathbf{p} = \begin{bmatrix} 1 & {}^{p}x_1(1) & {}^{p}x_2(1) \\ 1 & {}^{p}x_1(2) & {}^{p}x_2(2) \\ \vdots & \vdots & \vdots \\ 1 & {}^{p}x_1(n) & {}^{p}x_2(n) \end{bmatrix} \quad \text{and} \quad \mathbf{q} = \begin{bmatrix} 1 & {}^{q}x_1(1) & {}^{q}x_2(1) \\ 1 & {}^{q}x_1(2) & {}^{q}x_2(2) \\ \vdots & \vdots & \vdots \\ 1 & {}^{q}x_1(n) & {}^{q}x_2(n) \end{bmatrix}. \tag{A.1}$$

In mapping of landmarks describing biological shapes the thin-plate spline (TPS) transformation has proven useful [15].

The mapping can be formulated as a variational problem where we wish to find a function $f : \mathbb{R}^2 \mapsto \mathbb{R}^2$ that minimises:

$$E_{TPS} = \frac{1}{n}\sum_{i=1}^{n}\big\|p_i - f(q_i)\big\|^2 + \lambda J_m^D(f), \tag{A.2}$$

where $f(q_i)$ is the thin-plate spline transformation (or warp) of the points in $\mathcal{Q}$. J_m^D is the general form of the thin-plate penalty functional as defined in [84]. In two dimensions ($D = 2$) and using degree of derivatives ($m = 2$)

$$J_2^2 = J = \iint \left(\left(\frac{\partial^2 f}{\partial x_1^2}\right)^2 + \left(\frac{\partial^2 f}{\partial x_1 \partial x_2}\right)^2 + \left(\frac{\partial^2 f}{\partial x_2^2}\right)^2 \right) dx_1 dx_2. \tag{A.3}$$

By minimising the first term in (A.2) the points in $\mathcal{Q}$ is mapped as closely as possible to their corresponding points in $\mathcal{P}$. Including the second term in the minimisation imposes a limit on the second partial derivatives of f, i.e., a smoothness constraint on the bending energy of the spline. $\lambda \in \mathbb{R}^+$ is introduced to regularise the problem. For $\lambda \to 0$ there is no penalty on the bending energy of the spline and f can behave freely, i.e., in the interpolating form. For increasing values of λ f becomes gradually more smooth.

Writing

$$t = (x_1, x_2), \quad t_i = (x_1(i), x_2(i)) \tag{A.4}$$

and in homogeneous coordinates

$$\phi_1(t) = 1, \quad \phi_2(t) = x_1, \quad \phi_3(t) = x_2 \tag{A.5}$$

it can be shown [84, 29] that if $t_1, \ldots, t_n$ are such that least squares regression on $\phi_1, \ldots, \phi_n$ is unique then the unique minimiser f_λ of (A.2) is

$$f_\lambda(t) = \sum_{\nu=1}^{3} d_\nu \phi_\nu(t) + \sum_{i=1}^{n} c_i E(t, t_i), \tag{A.6}$$

where $E(t, t_i) = E(\tau)$ is the Greens function or the thin-plate spline kernel. $\tau = |t - t_i|$ is the Euclidean distance between a point t and the landmark point t_i. In two dimensions

$$E(t_i, t_j) = E(\tau) = \frac{1}{16\pi} |\tau|^2 ln|\tau|. \tag{A.7}$$

In (A.6) the unknowns are $d = (d_1, \ldots d_3)$ and $c = (c_1, \ldots, c_n)$. Substitution of (A.6) into (A.2) results in minimisation of

$$E_{TPS2}(c, d) = \frac{1}{n} \left\| \mathbf{p} - \mathbf{q}d - Kc \right\|^2 + \lambda c' K c \tag{A.8}$$

subject to $\mathbf{q}'c = 0$. Here, the points in $\mathcal{P}$ and $\mathcal{Q}$ have been arranged in $n \times 3$ matrices $\mathbf{p}$ and $\mathbf{q}$ (homogeneous coordinates), d is a 3×3 matrix holding the affine parameters of the transformation, c is a $n \times 3$ matrix with the parameters specifying the non-affine part of the transformation. K is an $n \times n$ matrix of thin-plate spline kernel values, i.e. the ijth entry $K(i, j) = E(t_i, t_j)$.

c and d are the minimisers of (A.8) and in order to find them a QR-decomposition of $\mathbf{q}$ is used:

$$\mathbf{q} = \begin{pmatrix} \mathbf{q}_1 : \mathbf{q}_2 \end{pmatrix} \begin{pmatrix} \mathbf{r} \\ 0 \end{pmatrix}. \tag{A.9}$$

The QR-decomposition results in the two matrices $\begin{pmatrix} \mathbf{q}_1 : \mathbf{q}_2 \end{pmatrix}$ and $\mathbf{r}$. The first is orthogonal and $\mathbf{r}$ is a 3×3 upper triangular matrix. $\mathbf{q}_1$ is $n \times 3$ and $\mathbf{q}_2$ is $n \times (n - 3)$. Setting $\mathbf{q}_2\gamma = c$, substituting into (A.8) and using $\mathbf{q}_2'\mathbf{q} = 0$ yields:

$$E_{TPS} = \frac{1}{n} \left\| \mathbf{p} - \begin{pmatrix} \mathbf{q}_1 : \mathbf{q}_2 \end{pmatrix} \begin{pmatrix} \mathbf{r} \\ 0 \end{pmatrix} d - K\mathbf{q}_2\gamma \right\|^2 + \tag{A.10}$$

$$\lambda\gamma' \mathbf{q}_2' K \mathbf{q}_2 \gamma \tag{A.11}$$

$$= \frac{1}{n} \left\| \mathbf{q}_1'\mathbf{p} - \mathbf{r}d - \mathbf{q}_1' K \mathbf{q}_2 \gamma \right\|^2 + \tag{A.12}$$

$$\frac{1}{n} \left\| \mathbf{q}_2'\mathbf{p} - \mathbf{q}_2' K \mathbf{q}_2 \gamma \right\|^2 + \tag{A.13}$$

$$\lambda\gamma' \mathbf{q}_2' K \mathbf{q}_2 \gamma. \tag{A.14}$$

It can then be shown that

$$\mathbf{r}d = \mathbf{q}_1'\mathbf{p} + \mathbf{q}_1'K\mathbf{q}_2\gamma = \mathbf{q}_1'(\mathbf{p} - K\mathbf{q}_2\gamma), \quad \text{and} \tag{A.15}$$

$$\mathbf{q}_2'\mathbf{p} = \mathbf{q}_2'K\mathbf{q}_2\gamma + \lambda I\gamma = (\mathbf{q}_2'K\mathbf{q}_2 + \lambda I)\gamma \tag{A.16}$$

leading to expressions for the minimisers d and γ:

$$\gamma = \left(\mathbf{q}_2'K\mathbf{q}_2 + \lambda I\right)^{-1}\mathbf{q}_2'\mathbf{p} \tag{A.17}$$

$$d = \mathbf{r}^{-1}\left(\mathbf{q}_1'\mathbf{p} + \mathbf{q}_1 K\mathbf{q}_2\gamma\right) \tag{A.18}$$

Appendix B

Algorithms

B.1 Binarization of fuzzy match matrix

This section provides the structure of three algorithms:

- Binarization($\tilde{m}, \tau$),
- Select-best-match-in-row($\tilde{m}, {}^{r}\hat{m}, j, c$), and
- Select-best-match-in-column($\tilde{m}, {}^{c}\hat{m}, k, r$).

The dual functions Alg. 9 and 10 are called by Binarization($\tilde{m}, \tau$). This algorithm provides a heuristic method for the binarization of the *fuzzy* augmented match matrix $\tilde{m}$ given a threshold value, τ. The output of the algorithm is a *binary* augmented match matrix, $\hat{m}$. Definitions of fuzzy and binary match matrices are given in § 4.1.1.

A simple threshold of the input matrix, $\tilde{m}$, so that entries above or equal to the threshold value τ are set to 1, and entries below τ are set to 0 could, for low values of τ (≤ 0.5), result in ambiguities, i.e., in one-to-many or many-to-one correspondences.

It is ensured by heuristics in the algorithm, that the output matrix confines to the requirements in § 4.1.1, so that ambiguities are avoided.

The result of simple thresholding $\tilde{m}$ is processed row-wise and column-wise separately using Alg. 9 and 10, respectively and the two sub-results are combined using elementwise multiplication.

The two sub-functions, Alg. 9 and 10 are dual and work on rows and columns separately. In cases of more than one element in a row (or column) *above or equal to* the threshold value, these functions are used to select the best match in the row (column). The best match is the largest element in the original input matrix $\tilde{m}$. Other candidates above the threshold are set to zero. In cases of doubt (equal fuzzy match entries), the entire row (column) is set to zero and the point is marked as a single.

Algorithm 8 BINARIZATION($\tilde{m}$, τ)

1: ${}^{r}\hat{m} \leftarrow \tilde{m} \geq \tau$ *threshold each matrix element*
2: **for** $j = 1$ to J **do** *for all rows, resolve draws if any*
3: $c \leftarrow$ candidate columns *columns that candidate for the match*
4: **if** card $c > 1$ **then**
5: SELECT-BEST-MATCH-IN-ROW($\tilde{m}$,${}^{r}\hat{m}, j, c$) *see Alg. 9*
6: **else if** card $c == 0$ **then** *none over the treshold*
7: ${}^{r}\hat{m}_{j\cdot} = 0$ *zero j'th row*
8: ${}^{r}\hat{m}_{j(K+1)} = 1$ *mark as single*
9: **end if**
10: **end for**
11: ${}^{c}\hat{m} \leftarrow \tilde{m} \geq \tau$ *threshold each matrix element*
12: **for** $k = 1$ to K **do** *for all cols, resolve draws if any*
13: $r \leftarrow$ candidate rows *rows that candidate for the match*
14: **if** card $r > 1$ **then**
15: SELECT-BEST-MATCH-IN-COLUMN($\tilde{m}$,${}^{c}\hat{m}, k, r$) *see Alg. 10*
16: **else if** card $r == 0$ **then** *none over the treshold*
17: ${}^{c}\hat{m}_{\cdot k} = 0$ *zero k'th column*
18: ${}^{c}\hat{m}_{(J+1)k} = 1$ *mark as single*
19: **end if**
20: **end for**
21: $\hat{m} = \hat{m}r \odot \hat{m}c$ *elementwise multiplication*
22: $\forall j \leq J \quad \hat{m}_{j(K+1)} \leftarrow 1 - \sum_{k=1}^{K} \hat{m}_{jk}$ *force top J rows to sum to 1*
23: $\forall k \leq K \quad \hat{m}_{(J+1)k} \leftarrow 1 - \sum_{j=1}^{J} \hat{m}_{jk}$ *force K leftmost columns to sum to 1*

Algorithm 9 SELECT-BEST-MATCH-IN-ROW($\tilde{m}, {}^r\hat{m}, j, c$)

1: $\mu \leftarrow \max \tilde{m}_{j\cdot}$ *max value in row j*
2: $maxc \leftarrow$ elements in row j that equals μ *max candidates*
3: $\lambda \leftarrow \sum_{k=1}^{K}(\tilde{m} == \mu)$ *number of elements in row j that equals μ*
4: **if** card $maxc > 1$ **then** *more than one maximum*
5: ${}^r\hat{m}_{j\cdot} = 0$ *zero all elements in the j'th row*
6: ${}^r\hat{m}_{j(K+1)} = 1$ *mark this row as single*
7: **else** *exactly* one *maximum in the row*
8: ${}^r\hat{m}_{j\cdot} = 0$ *zero all elements in the j'th row*
9: ${}^r\hat{m}_{j\,maxc} = 1$ *except for the maximum in column maxc*
10: **end if**

Algorithm 10 SELECT-BEST-MATCH-IN-COLUMN($\tilde{m}, {}^c\hat{m}, k, r$)

1: $\mu \leftarrow \max \tilde{m}_{\cdot k}$ *max value in column k*
2: $maxr \leftarrow$ elements in column k that equals μ *max candidates*
3: $\lambda \leftarrow \sum_{j=1}^{J}(\tilde{m} == \mu)$ *number of elements in row k that equals μ*
4: **if** card $maxr > 1$ **then** *more than one maximum*
5: ${}^c\hat{m}_{\cdot k} = 0$ *zero all elements in the k'th column*
6: ${}^c\hat{m}_{(J+1)k} = 1$ *mark this column as single*
7: **else** *exactly* one *maximum in the column*
8: ${}^c\hat{m}_{\cdot k} = 0$ *zero all elements in the k'th row*
9: ${}^c\hat{m}_{maxr\,k} = 1$ *except for the maximum in row maxr*
10: **end if**

Appendix C

Data material

Centre for Proteome Analysis in Life Sciences (CPA) has kindly provided the data material. A data set of 15 16-bit grey level images with various attribute information available for every protein spot in every image. This information was extracted by a skilled technician from a manually guided segmentation and matching process using commercially available state-of-the-art software.

C.1 Gel Images

The radioactive marked proteins were separated in the first dimension on commercially available immobilised pH gradient gels spanning the pH interval from 4 to 7 using in total 50kVhrs. Subsequently the proteins were separated in the second dimension by their molecular weight on the 12.5% polyacrylamide gels. After electrophoresis the gels were vacuum dried on a paper support and exposed a phosphor plate for 5 days. The images were captured using a phosphor imager (Agfa) in a 12-bit format. See also § 2.1.2.

C.1.1 Data set

The dimensions of the 15 images are approximately 2000 by 1800 pixels. The protein samples used to produce the images are prepared from bakers yeast *Saccharomyces cerevisiae* strain `Fy1679-28C EC [pRS315]` and strain `Fy1679-28C EC [pRS315::PDR1]`.

The 15 images of data set 1 is divided in 5 groups (1 to 5) with each 3 images (A, B, C). Images from group 1, 2, and 3 are comparable, i.e., in a biological context it makes sense to compare the samples. Images from group 4 and 5 are comparable, but cannot be compared to samples images from group 1, 2 or 3 and vice versa. See Fig. C.1.

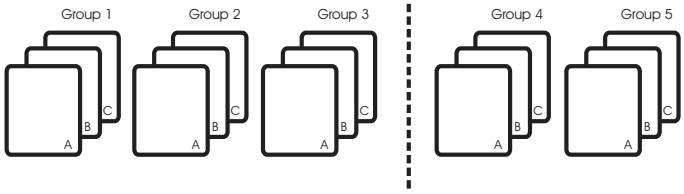


Figure C.1: Overview of images in Data Set 1.

C.2 Protein Spot Attribute Information

For each spot a range of attribute information is available:

- ID
- area
- x centre coordinate
- y centre coordinate
- integrated intensity (II)
- peak intensity
- background intensity
- % integrated intensity (%II)
- Gaussian integrated intensity

The attribute information has been generated and extracted from a commercial software product BioImage^{TM} for analysis of 2D electrophoresis images.

C.2.1 Match information

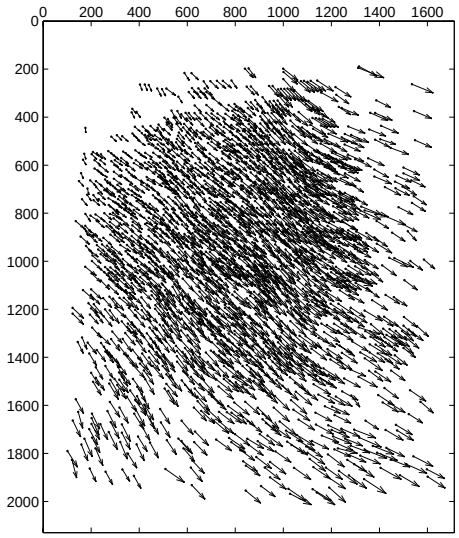
The correspondence between homologous protein spots is known across gel images within the same data set. Spots with same spot ID correspond. This information can be represented in a match matrix $\hat{m}$ (see §4.1.1).

C.3 Disparity Analysis

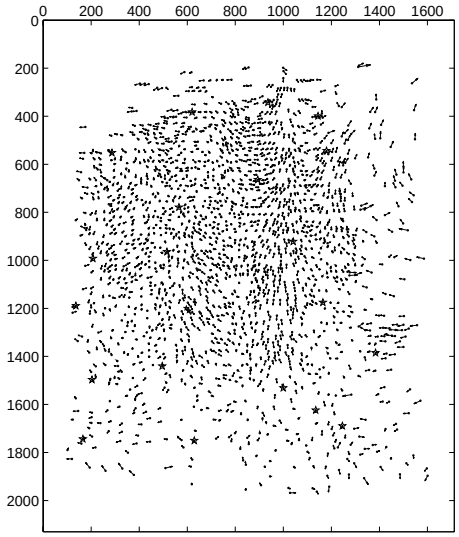
Fig. C.2 shows the spots stemming from *two* different gels plotted together as points. Homologous (corresponding) points are connected with arrows. Long arrows mean that corresponding points are far apart and vice versa. The arrows form a so called *disparity field*.

The top plot displays the initial or raw disparity field by simply superimposing the two point sets and drawing lines between corresponding points. The bottom plot shows the residual disparity field after correction for an initial alignment (1st order polynomial transformation) of the two point sets.

Fig. C.3 shows the corrected disparity field of another gel pair and the axes of the plot show "histograms" of the disparity lengths. It gives an impression of the average disparity in different areas of the gel.



(a) Initial disparity field.



(b) Disparity field after correction for global 1st order polynomial transformation.

Figure C.2: Group 1A vs. Group 1B.

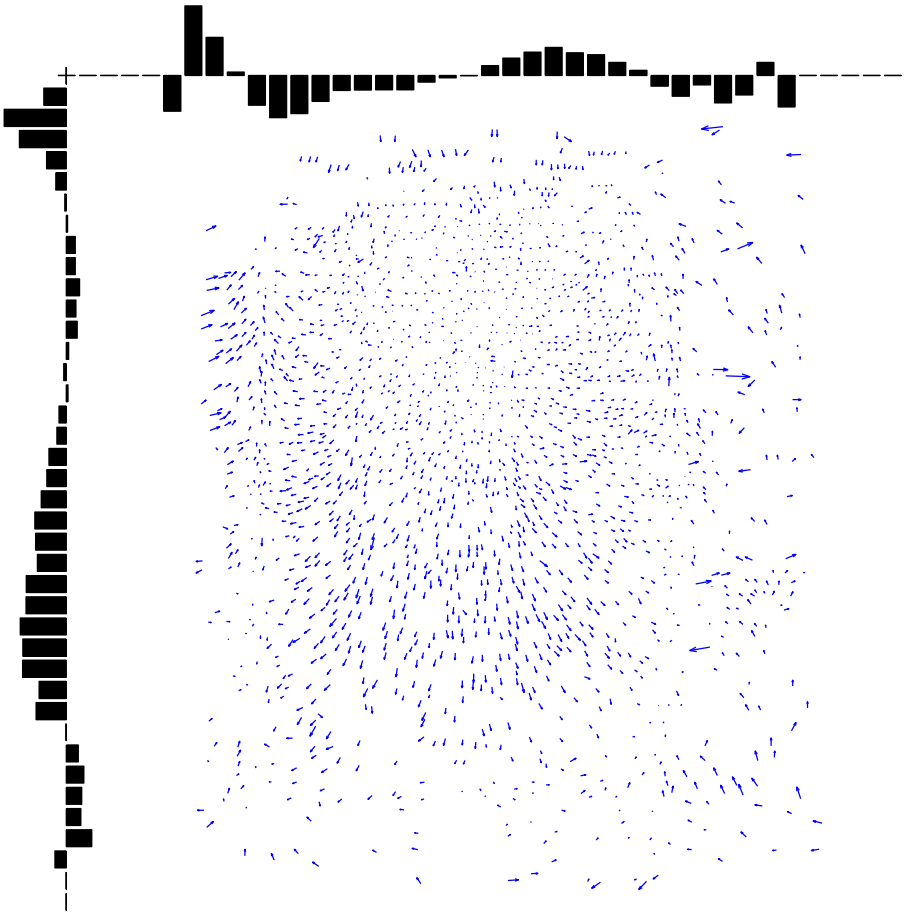


Figure C.3: Disparity field for gel 1A vs. gel 2A with average disparity histograms.

Appendix D

Grey level based warping

The method proposed by Conradsen and Pedersen [25] does not provide means for regularisation of the disparity maps and therefore produces quite rough mappings. Warping of the gel images according to the rough disparity maps results in transformed gel images which may be aligned well but also are deformed so that the spots do not resemble protein spots any longer.

At increasing levels of resolution, disparity maps are calculated in the line (horizontal) direction, δ_h and in the column (vertical) direction, δ_v. Starting at low resolution, the disparity maps are calculated by minimising sums of squared differences and this disparity information is propagated to the next (higher) resolution level.

A simple regularisation could consist of smoothing the disparity maps using Gaussian convolution.

Given an image I and an isotropic 2D Gaussian $g(\mathbf{x};\sigma)$,

$$g(\mathbf{x};\sigma) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}}.$$

σ is referred to as the standard deviation of the 2D Gaussian and L is defined as the convolution of I with $g(\mathbf{x};\sigma)$:

$$L(\mathbf{x};\sigma) = I * g(\mathbf{x};\sigma).$$

By Gaussian smoothing of the disparity maps δ_v and δ_h at each level of resolution before the *warping* step (see [25], Fig. 4) a regularisation is introduced. The standard deviation, σ of the Gaussian convolution kernel serves as regularisation parameter. The larger σ, the more the disparity maps are smoothed and the warping becomes less rough.

In the following the effect of regularisation will be demonstrated by showing an example of grey level registration of two 2DGE gel regions.

D.1 Experiments

Fig. D.1 shows corresponding 512×512 sub-regions of two gel images. These regions will be the subject in the following examples showing the results of the original [25] method without regularisation of the disparity maps as well as results of regularisation using Gaussian smoothing of the disparity maps.

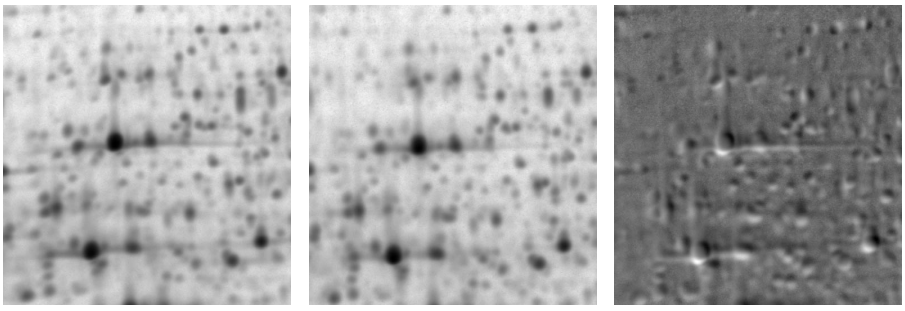


Figure D.1: Original 512×512 region of two gel images. Left: Reference image (A). Centre: Match image (B). Right: Pixel-wise difference A-B.

Following [25], the reference image (A) is warped according to the horizontal disparity map δ_h and the match image (B) is warped according to the vertical disparity map δ_v.

D.1.1 No regularisation

Fig. D.2 shows the disparity maps and a 512×512 pixel regular grid warped according to the disparity maps.

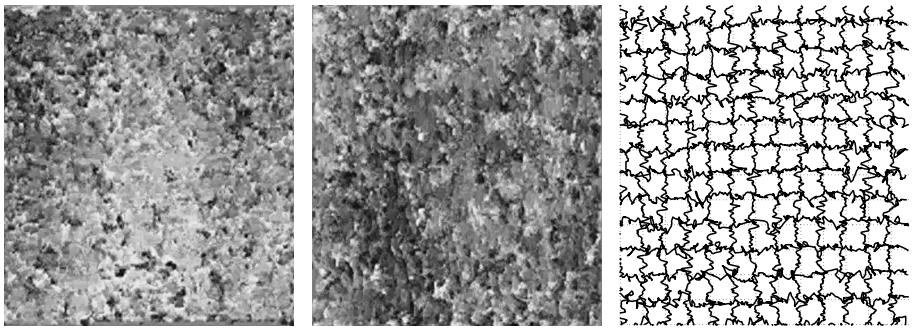


Figure D.2: No regularisation of disparity maps. Disparity maps and warped grid. Left: Horizontal disparity map δ_h. Centre: Vertical disparity map δ_v. Right: Regular grid warped according to disparity maps.

Fig. D.3 shows the result of the warping. Note by comparison with the original images in Fig. D.1 how the rough disparity maps have deformed the protein spots.

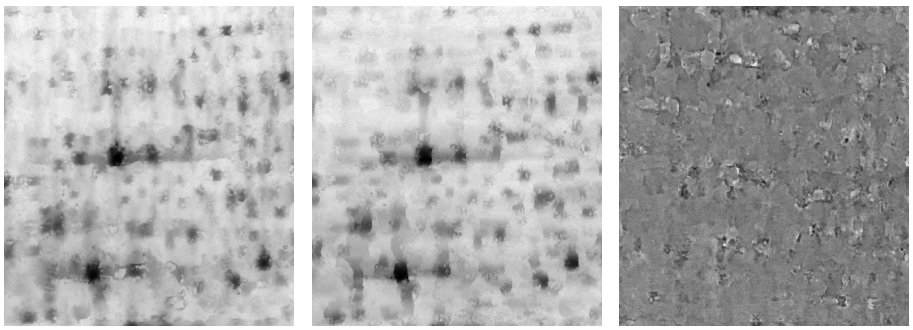


Figure D.3: No regularisation of disparity maps. Warped versions of original 512×512 images. Left: Reference image warped according to δ_h (Aw). Centre: Match image warped according to δ_v (Bw). Right: Pixel-wise difference Aw-Bw.

D.1.2 Gaussian smoothing

Fig. D.4 shows the disparity maps and a 512×512 pixel regular grid warped according to the *smoothed* disparity maps. In this experiment, the standard deviation of the 2D Gaussian was chosen to $\sigma = 3$.

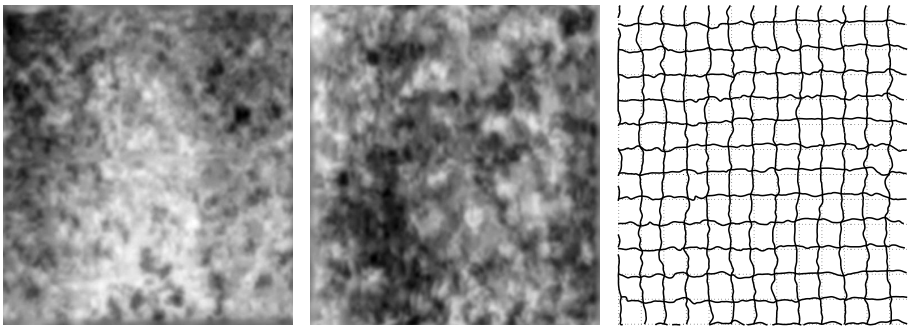


Figure D.4: Gaussian smoothing of disparity maps. Disparity maps and warped grid. Left: Horizontal disparity map δ_h. Centre: Vertical disparity map δ_v. Right: Regular grid warped according to disparity maps.

Fig. D.5 shows the result of the warping. Note by comparison with the original images in Fig. D.1 how the smooth disparity maps have preserved the shapes of the spots better than without regularisation.

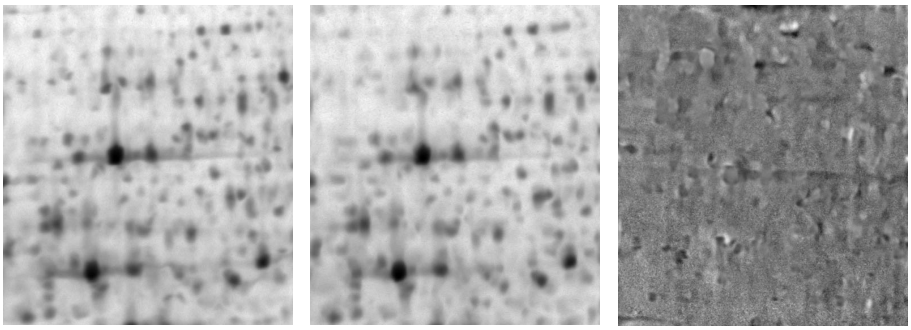


Figure D.5: Gaussian smoothing of disparity maps, $\sigma = 3$. Warped versions of original 512×512 images. Left: Reference image warped according to δ_h (Aw). Centre: Match image warped according to δ_v (Bw). Right: Pixel-wise difference Aw-Bw.

Fig. D.6 shows the difference images from Figs. D.1, D.3 and D.5 displayed in common grey level range. Note how the regularisation of the disparity maps also results in the smallest residuals after the warp.

Fig. D.7 displays the image pairs before warp and after both types of warp using a pseudo-colour display[75] where the reference gel is displayed as green and the match gel is magenta. Good alignment will appear as white, grey or black because of the overlay of magenta and green. Wherever there is differential

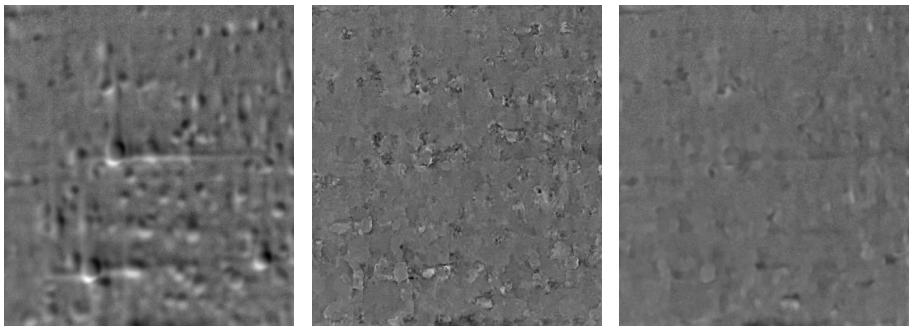


Figure D.6: Difference images from Figs. D.1, D.3 and D.5 displayed in common grey level range. Left: Difference image before warp. Centre: Difference image after warp *without* regularisation of the disparity maps. Right: Difference image after warp *with* regularisation of the disparity maps.

expression or the images are not well aligned, the image will appear in green or magenta tones.

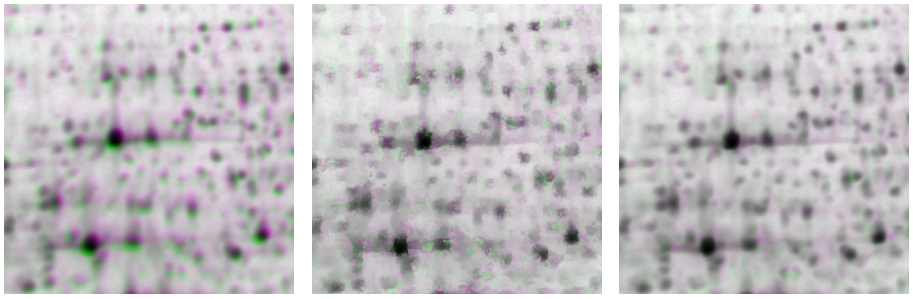


Figure D.7: Pseudo-colour display of image pairs. Left: Original images A (green) and B (magenta). Centre: Aw (green) and Bw (magenta) after warp *without* regularisation of the disparity maps. Right: Aw (green) and Bw (magenta) after warp *with* regularisation of the disparity maps.

D.2 Application in Point Matching

In matching of point sets (see § 4) the grey level matching technique described here may serve as a preprocessing step. By warping the point sets according to the Gaussian smoothed disparity maps, corresponding points are positioned closer to each other and the correspondence is expected to be easier to resolve.

Experiments with this approach using the regionalised RPM-affine method (§ 4.2.4, 4.3.4) did however not result in significantly better matching results. This may be explained by the fact, that the corresponding points are brought closer prior to the point matching, but the local deformation no longer is affine. Similar experiments have not been conducted with the regionalised RPM-TPS method but it is expected that for this method, the preprocessing step will be beneficial.

Bibliography

[1] A Abbott. A post-genomic challenge: learning to read patterns of protein synthesis. *Nature*, 402:715–720, 1999.

[2] T Akutsu, K Kanaya, A Ohyama, and A Fujiyama. Matching of spots in 2D electrophoresis images. Point matching under non-uniform distortions. In *Combinational Pattern Matching. 10th Annual Symposium, CPM'99*, pages 212–222, 1999.

[3] R D Appel, J R Vargas, P M Palagi, D Walther, and D F Hochstrasser. Melanie II - a third-generation software package for analysis of two-dimensional electrophoresis images: II. Algorithms. *Electrophoresis*, 18: 2735–2748, 1997.

[4] F Aurenhammer. Voronoi diagram - a survey of a fundamental geometric data structure. *ACM Computing Surveys*, 1991.

[5] S Belongie and J Malik. Matching with shape contexts. In *IEEE Workshop on Content-Based Access of Image and Video Libraries*, 2000.

[6] S Belongie, J Malik, and J Puzicha. Matching shapes. In *Eighth IEEE International Conference on Computer Vision*, 2001.

[7] S Belongie, J Malik, and J Puzicha. Shape matching and object recognition using shape contexts. Technical report, U.C. Berkely, 2001.

[8] J E Besag. On the statistical analysis of dirty pictures. *J. Royal Statistical Society*, 48(B):259–302, 1986.

[9] P J Besl and H D McKay. A method for registration of 3-d shapes. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 14(2):239–256, 1992. ISSN 01628828.

[10] E Bettens, P Scheunders, J Sijbers, D Van Dyck, and L Moens. Automatic segmentation and modelling of two-dimensional electrophoresis gels. In *International Conference on Image Processing*, volume 2, pages 665–668, 1996.

[11] S Beucher. Extrema of grey-tone functions and mathematical morphology. In *Proc. of the Colloquium on Math. Morp., Stereol. and Image Analysis, Prague, Tchecoslovaquia*, pages 59–70, 1982.

[12] S Beucher and C Lantuejoul. Use of watersheds in contour detection. In *International Workshop on image processing, real-time edge and motion detection/estimation, Rennes, France*, 1979.

[13] A Blomberg, L Blomberg, J Norbeck, S J Fey, P M Larsen, M Larsen, P Roepstorff, H Degand, M Boutry, A Posch, and A Görg. Interlaboratory reproducibility of yeast protein patterns analyzed by immobilized pH gradient two-dimensional gel electrophoresis. *Electrophoresis*, 16(10): 1935–1945, 1995.

[14] D Blostein and N Ahuja. A multiscale region detector. *Computer Vision, Graphics, and Image Processing*, 45(1):22–41, 1989. ISSN 0734189x.

[15] F L Bookstein. Principal Warps: Thin-Plate Splines and the Decomposition of Deformations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(6):567–585, 1989.

[16] L G Brown. A Survey of Image Registration Techniques. *ACM Computing Surveys*, 24(4), 1992.

[17] M Carcassoni and E R Hancock. Point pattern matching with robust spectral correspondence. *Proceedings IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2000 (Cat. No.PR00662)*, pages 649–55 vol.1, 2000.

[18] J M Carstensen. *Description and simulation of visual texture*. PhD thesis, Institute of Mathematical Statistics and Operational Research, Technical University of Denmark, 1992.

[19] J M Carstensen. An active lattice model in a bayesian framework. *Computer Vision and Image Understanding*, 63(2):380–387, 1996.

[20] S-H Chang, F-H Cheng, W-H Hsu, and G-Z Wu. Fast algorithm for point pattern matching: invariant to transslations, rotations and scale changes. *Pattern Recognition*, 30(2):311–320, 1997.

[21] F-H Cheng. Point pattern matching algorithm invariant to geometrical transformation and distortion. *Pattern Recognition Letters*, 17:1429–1435, 1996.

[22] H Chui. *Non-Rigid Point Matching: Algorithms, Extensions and Applications*. PhD thesis, Yale University, May 2001.

[23] H Chui and A Rangarajan. A new algorithm for non-rigid point matching. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 44–51, 2000.

[24] H Cohn and M Fielding. Simulated Annealing: Searching for an Optimal Temperature Schedule. *SIAM Journal on Optimization*, 9(3):779–802, 1999.

[25] K Conradsen and J Pedersen. Analysis of two-dimensional electrophoresis gels. *Biometrics*, 42:1273–1287, 1992.

[26] T F Cootes, C J Taylor, D H Cooper, and J Graham. Active shape models-their training and application. *Computer Vision and Image Understanding*, 61(1):38–59, 1995. ISSN 10773142.

[27] D J Cross and E R Hancock. Graph matching with a dual-step EM algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1236–1253, 1998.

[28] I L Dryden and K V Mardia. *Statistical Shape Analysis*. Wiley, 1998.

[29] L Duchon. Splines minimizing rotation-invariant semi-norms in Sobolev spaces. In *Constructive Theory of Functions of Several Variables*, pages 85–100. Springer Verlag, Berlin, 1977.

[30] S J Fey. Personal communication, 2002.

[31] S J Fey and P Mose-Larsen. Image analysis method for e.g. related or similar gel electrophoresis images - using master composite image, generated by averaging selected images, to analyse and interpret existing or new images. Patent numbers WO9811508-A1, AU9741325-A, EP925554-A1, JP2001500614-W, KR2000036155-A, 1998.

[32] S J Fey and P Mose Larsen. 2D or not 2D. Two-dimensional gel electrophoresis. *Current Opinion in Chemical Biology*, 5(1):26–33, 2001.

[33] S J Fey, A Nawrocki, M R Larsen, A Görg, P Roepstorff, G N Skews, R Williams, and P Mose Larsen. Proteome analysis of saccharomyces cerevisiae: a methodological outline. *Electrophoresis*, 18(8):1361–72, 1997.

[34] C A Glasbey and K V Mardia. A review of image warping methods. *Journal of Applied Statistics*, 25:155–171, 1998.

[35] S Gold and A Rangarajan. A graduated assignment algorithm for graph matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(4):377–388, 1996.

[36] S Gold, A Rangarajan, C-P Lu, S Pappu, and E Mjolsness. New algorithms for 2D and 3D point matching: pose estimation and correspondence. *Pattern Recognition*, 31(8):1019–1031, 1998.

[37] P J Green and B W Silverman. *Nonparametric Regression and Generalized Linear Models - a roughness penalty approach*. Chapman & Hall, 1994.

[38] E Guest, E Berry, R A Baldock, M Fidrich, and M A Smith. Robust point correspondence applied to two- and three-dimensional image registration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(2): 165–179, 2001.

[39] R Hartley and A Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2000.

[40] F Hoffmann, K Kriegel, and C Wenk. Matching 2D patterns of protein spots. In *Proceedings of the fourteenth annual Symposium on Computational Geometry*, pages 231–239, 1998.

[41] F Hoffmann, K Kriegel, and C Wenk. An applied point pattern matching problem: comparing 2D patterns of protein spots. *Discrete Applied Mathematics*, 93:75–88, 1999.

[42] G W Horgan, A Creasey, and B Fenton. Superimposing two-dimensional gels to study genetic variation in malaria parasites. *Electrophoresis*, 13: 871–875, 1992.

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

[44] A K Jain, R P W Duin, and Jianchang Mao. Statistical pattern recognition: a review. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(1):4–37, 2000.

[45] H J Johnson and G E Christensen. Deformable Registration - Landmark and intensity-based, consistent thin-plate spline image registration. *Lecture Notes in Computer Science*, 2082:329–343, 2001.

[46] M Kass, A Witkin, and D Terzopoulos. Snakes: active contour models. *International Journal of Computer Vision*, 1(4):321–31, 1987. ISSN 09205691.

[47] J J Kosowsky and A L Yuille. The invisible hand algorithm: Solving the assignment problem with statistical physics. *Neural Networks*, 7(3):477–490, 1994.

[48] C Kotropoulos, A Tefas, and I Pitas. Morphological elastic graph matching applied to frontal face authentication under well-controlled and real conditions. *Pattern Recognition*, 33(12):1935–1947, 2000. ISSN 00313203.

[49] S Kumar, M Sallam, and D Goldgof. Matching point features under small nonrigid motion. *Pattern Recognition*, 34(12):2353–2365, 2001. ISSN 00313203.

[50] S Kumar, M Sallam, Dmitry Goldgof, and Kevin Bowyer. Point Correspondence in Unstructured Nonrigid Motion. In *Proceeings International Symposium on Computer Vision*, pages 289–94, 1995.

[51] M Lades, J C Vorbruggen, J Buhmann, J Lange, C von der Malsburg, R P Wurtz, and W Konen. Distortion invariant object recognition in the dynamic link architecture. *Computers, IEEE Transactions on*, 42(3):300 –311, 1993. ISSN 00189340.

[52] Hyung-Ji Lee, Wan-Su Lee, and Jae-Ho Chung. Face recognition using fisherface algorithm and elastic graph matching. *Image Processing, 2001. Proceedings. 2001 International Conference on*, 1:998–1001, 2001.

[53] P F Lemkin. Comparing two-dimensional electrophoretic gel images across the internet. *Electrophoresis*, 18(3-4):461–470, 1997. ISSN 01730835.

[54] T Lindeberg. Discrete derivative approximations with scale-space properties: A basis for low-level feature extraction. *Journal of Mathematical Imaging and Vision, JMIV*, 3(349-376), 1993.

[55] T Lindeberg. *Scale-space theory in computer vision*. Kluwer Academic Publishers, Netherlands, 1994.

[56] T Lindeberg. Scale-space: A framework for handling image structures at multiple scales. In *CERN School of Computing*, 1996.

[57] T Lindeberg. Feature Detection with Automatic Scale Selection. *International Journal of Computer Vision*, 1998.

[58] J B A Maintz and M A Viergever. A survey of medical image registration. *Medical Image Analysis*, 2(1):1–36, 1998.

[59] F Menard and F Fogelman Soulie. Application of the topological maps algorithm to the recognition of bidimensional electrophoresis images. *INNC 90 Paris. International Neural Network Conference*, 1:99–102, 1990.

[60] J M Miller, A D Olson, and S S Thorgeirsson. Computer analysis of two-dimensional gels: Automatic matching. *Electrophoresis*, 1984.

[61] F Murtagh. A new approach to point-pattern matching. *Publications of the Astronomical Society of the Pacific*, 104(674):301–7, 1992. ISSN 00046280.

[62] H Ogawa. Labeled Point Pattern Matching by Delaunay Triangulation and Maximal Cliques. *Pattern Recognition*, 1986.

[63] O F Olsen and M Nielsen. Multi-scale gradient magnitude watershed segmentation. pages 6–13 vol.1. Springer-Verlag, 1997. ISBN 3540635076.

[64] J Pánek and J Vohradský. Point pattern matching in the analysis of two-dimensional gel electropherograms. *Electrophoresis*, 20:3483–3491, 1999.

[65] L Pedersen and B Ersbøll. Protein spot correspondence in two-dimensional electrophoresis gels. In *Scandinavian Conference on Image Analysis*, pages 118–125, 2001.

[66] K-P Pleissner, F Hoffmann, K Kriegel, C Wenk, S Wegner, A Sahlström, H Oswald, H Alt, and E Fleck. Proteome data analysis and management - new algorithmic approaches to protein spot detection and pattern matching in two-dimensional electrophoresis gel databases. *Electrophoresis*, 20(4-5): 755–765, 1999. ISSN 01730835.

[67] A Rangarajan, H Chui, and F L Bookstein. The softassign procrustes matching algorithm. *Lecture Notes in Computer Science*, pages 29–36, 1997.

[68] A Rangarajan, H Chui, and E Mjolsness. A new distance measure for non-rigid image matching. *Energy Minimization Methods in Computer Vision and Pattern Recognition. Second International Workshop, EMMCVPR'99. Proceedings (Lecture Notes in Computer Science Vol.1654)*, pages 237–52, 1999.

[69] A Rangarajan, H Chui, E Mjolsness, S Pappu, L Davachi, P Goldman-Rakic, and J Duncan. A robust point-matching algorithm for autoradiograph alignment. *Medical Image Analysis*, 1(4):379–398, 1997. ISSN 13618415.

[70] A Rangarajan and J Duncan. Matching point features using mutual information. In *Workshop on Biomedical Image Analysis*, pages 172–181, 1998.

[71] K Rohr, H S Stiehl, R Sprengel, and W Beil. Point-based elastic registration of medical image Data using approximating thin-plate splines. *Lecture Notes in Computer Vision*, 1131:297–306, 1996.

[72] S Sclaroff and A P Pentland. Modal matching for correspondence and recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 17(6):545–561, 1995. ISSN 01628828.

[73] J Serra and P Soille, editors. *Mathematical morphology and its applications to image processing*. Kluwer Academic Publishers, 1994.

[74] M Skolnick. Application of morphological transformations to the analysis of two-dimensional electrophoretic gels of biological materials. *Computer Vision, Graphics, and Image Processing*, 35:306–332, 1986.

[75] Z Smilansky. Automatic registration for images of two-dimensional protein gels. *Electrophoresis*, pages 1616–1626, 2001.

[76] M Sonka, V Hlavac, and R Boyle. *Image processing, analysis, and machine vision*. Brooks/Cole Publ., 1999.

[77] J Sporring, M Nielsen, L Florack, and P Johansen, editors. *Gaussian scale-space theory*. Kluwer Academic Publishers, 1997.

[78] R Sprengel, K Rohr, and H S Stiehl. Thin-plate spline approximation for image registration. In *18th Internat. Conf. of the IEEE Engineering in Medicine and Biology Society*, pages 1190–1191, 1996.

[79] A Tefas, C Kotropoulos, and I Pitas. Using support vector machines to enhance the performance of elastic graph matching for frontal face authentication. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 23(7):735–746, 2001. ISSN 01628828.

[80] B Triggs, A Zisserman, and R Szeliski, editors. *Vision Algorithms: Theory and Practice, International Workshop on Vision Algorithms, held during ICCV '99, Corfu, Greece, September 21-22, 1999, Proceedings*, volume 1883 of *Lecture Notes in Computer Science*, 2000. Springer. ISBN 3-540-67973-1.

[81] L Vincent. Morphological grayscale reconstruction in image analysis: applications and efficient algorithms. *Image Processing, IEEE Transactions on*, 2(2):176–201, 1993. ISSN 10577149.

[82] L Vincent and P Soille. Watersheds in digital spaces: an efficient algorithm based on immersion simulations. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 13(6):583 –598, 1991. ISSN 01628828.

[83] T Voss and P Haberl. Observations on the reproducibility and matching efficiency of two-dimensional electrophoresis gels: Consequences for comprehensive data analysis. *Electrophoresis*, 2000.

[84] G Wahba. *Splines Models for Observational Data*. Series in Applied Mathematics, Vol. 59, SIAM, Philadelphia, 1990.

[85] V C Wasinger, S J Cordwell, A Cerpa-Poljak, J X Yan, A A Gooley, M R Wilkins, M W Duncan, R Harris, K L Williams, and I Humphery-Smith. Progress with gene-product mapping of the mollicutes: Mycoplasma genitalium. *Electrophoresis*, 16(7):1090–4, Jul 1995.

[86] Y Watanabe and K Takahashi. A fast structural matching and its application to pattern analysis of 2-D electrophoresis images. In *Proceedings 1998 International Conference on Image Processing*, pages 804–8. IEEE Comput. Soc, 1998.

[87] Y Watanabe and K Takahashi. An analysis system for two-dimensional gel electrophoresis images of genomic DNA. In *Proceedings. Fourteenth International Conference on Pattern Recognition (Cat. No.98EX170)*, volume 1, pages 368–371. IEEE Comput. Soc, 1998.

[88] Y Watanabe, K Takahashi, and M Nakazawa. Automated detection and macthing of spots in autoradiogram images of two-dimensional electrophoresis for high-speed genome scanning. In *Proceedings. International Conference on Image Processing*, volume 3, pages 496–499. IEEE Comput. Soc, 1997.

[89] M R Wilkins. From Proteins to Proteomes Large-Scale Protein Identification by Two-Dimensional Electrophoresis and Amino Acid Analysis. *Bio Technology*, 14:62–65, 1996.

[90] L Wiskott, J-M Fellous, N Kuiger, and C von der Malsburg. Face recognition by elastic bunch graph matching. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 19(7):775 –779, 1997. ISSN 01628828.

[91] Z Zhang. Iterative point matching for registration of free-form curves. Technical Report 1658, Institut National de Recherce en Informatic et en Automatique, INRIA, 1992.

Ph.D. theses from IMM

1. **Larsen, Rasmus.** (1994). *Estimation of visual motion in image sequences. xiv* + 143 pp.

2. **Rygard, Jens Moberg.** (1994). *Design and optimization of flexible manufacturing systems. xiii* + 232 pp.

3. **Lassen, Niels Christian Krieger.** (1994). *Automated determination of crystal orientations from electron backscattering patterns. xv* + 136 pp.

4. **Melgaard, Henrik.** (1994). *Identification of physical models. xvii* + 246 pp.

5. **Wang, Chunyan.** (1994). *Stochastic differential equations and a biological system. xxii* + 153 pp.

6. **Nielsen, Allan Aasbjerg.** (1994). *Analysis of regularly and irregularly sampled spatial, multivariate, and multi-temporal data. xxiv* + 213 pp.

7. **Ersbøll, Annette Kjær.** (1994). *On the spatial and temporal correlations in experimentation with agricultural applications. xviii* + 345 pp.

8. **Møller, Dorte.** (1994). *Methods for analysis and design of heterogeneous telecommunication networks.* Volume 1-2, *xxxviii* + 282 pp., 283-569 pp.

9. **Jensen, Jens Christian.** (1995). *Teoretiske og eksperimentelle dynamiske undersøgelser af jernbanekøretøjer. viii* + 174 pp.

10. **Kuhlmann, Lionel.** (1995). *On automatic visual inspection of reflective surfaces.* Volume 1, *xviii* + 220 pp., (Volume 2, *vi* + 54 pp., fortrolig).

11. **Lazarides, Nikolaos.** (1995). *Nonlinearity in superconductivity and Josephson Junctions. iv* + 154 pp.

12. **Rostgaard, Morten.** (1995). *Modelling, estimation and control of fast sampled dynamical systems. xiv* + 348 pp.

13. **Schultz, Nette.** (1995). *Segmentation and classification of biological objects. xiv* + 194 pp.

14. **Jørgensen, Michael Finn.** (1995). *Nonlinear Hamiltonian systems. xiv* + 120 pp.

15. **Balle, Susanne M.** (1995). *Distributed-memory matrix computations. iii* + 101 pp.

16. **Kohl, Niklas.** (1995). *Exact methods for time constrained routing and related scheduling problems. xviii* + 234 pp.

17. **Rogon, Thomas.** (1995). *Porous media: Analysis, reconstruction and percolation. xiv* + 165 pp.

18. **Andersen, Allan Theodor.** (1995). *Modelling of packet traffic with matrix analytic methods. xvi* + 242 pp.

19. **Hesthaven, Jan.** (1995). *Numerical studies of unsteady coherent structures and transport in two-dimensional flows.* Risø-R-835(EN) 203 pp.

20. **Slivsgaard, Eva Charlotte.** (l995). *On the interaction between wheels and rails in railway dynamics. viii* + 196 pp.

21. **Hartelius, Karsten.** (1996). *Analysis of irregularly distributed points. xvi* + 260 pp.

22. **Hansen, Anca Daniela.** (1996). *Predictive control and identification - Applications to steering dynamics. xviii* + 307 pp.

23. **Sadegh, Payman.** (1996). *Experiment design and optimization in complex systems. xiv* + 162 pp.

24. **Skands, Ulrik.** (1996). *Quantitative methods for the analysis of electron microscope images. xvi* + 198 pp.

25. **Bro-Nielsen, Morten.** (1996). *Medical image registration and surgery simulation. xxvii* + 274 pp.

26. **Bendtsen, Claus.** (1996). *Parallel numerical algorithms for the solution of systems of ordinary differential equations. viii* + 79 pp.

27. **Lauritsen, Morten Bach.** (1997). *Delta-domain predictive control and identification for control. xxii* + 292 pp.

28. **Bischoff, Svend.** (1997). *Modelling colliding-pulse mode-locked semiconductor lasers. xxii* + 217 pp.

29. **Arnbjerg-Nielsen, Karsten.** (1997). *Statistical analysis of urban hydrology with special emphasis on rainfall modelling.* Institut for Miljøteknik, DTU. *xiv* + 161 pp.

30. **Jacobsen, Judith L.** (1997). *Dynamic modelling of processes in rivers affected by precipitation runoff. xix* + 213 pp.

31. **Sommer, Helle Mølgaard.** (1997). *Variability in microbiological degradation experiments - Analysis and case study. xiv* + 211 pp.

32. **Ma, Xin.** (1997). *Adaptive extremum control and wind turbine control. xix* + 293 pp.

33. **Rasmussen, Kim Ørskov.** (1997). *Nonlinear and stochastic dynamics of coherent structures. x* + 215 pp.

34. **Hansen, Lars Henrik.** (1997). *Stochastic modelling of central heating systems. xxii* + 301 pp.

35. **Jørgensen, Claus.** (1997). *Driftsoptimering på kraftvarmesystemer.* 290 pp.

36. **Stauning, Ole.** (1997). *Automatic validation of numerical solutions. viii* + 116 pp.

37. **Pedersen, Morten With.** (1997). *Optimization of recurrent neural networks for time series modeling. x* + 322 pp.

38. **Thorsen, Rune.** (1997). *Restoration of hand function in tetraplegics using myoelectrically controlled functional electrical stimulation of the controlling muscle. x* + 154 pp. + Appendix.

39. **Rosholm, Anders.** (1997). *Statistical methods for segmentation and classification of images. xvi* + 183 pp.

40. **Petersen, Kim Tilgaard.** (1997). *Estimation of speech quality in telecommunication systems. x* + 259 pp.

41. **Jensen, Carsten Nordstrøm.** (1997). *Nonlinear systems with discrete and continuous elements.* 195 pp.

42. **Hansen, Peter S.K.** (1997). *Signal subspace methods for speech enhancement. x* + 226 pp.

43. **Nielsen, Ole Møller.** (1998). *Wavelets in scientific computing. xiv* + 232 pp.

44. **Kjems, Ulrik.** (1998). *Bayesian signal processing and interpretation of brain scans. iv* + 129 pp.

45. **Hansen, Michael Pilegaard.** (1998). *Metaheuristics for multiple objective combinatorial optimization. x* + 163 pp.

46. **Riis, Søren Kamaric.** (1998). *Hidden markov models and neural networks for speech recognition. x* + 223 pp.

47. **Mørch, Niels Jacob Sand.** (1998). *A multivariate approach to functional neuro modeling. xvi* + 147 pp.

48. **Frydendal, Ib.** (1998.) *Quality inspection of sugar beets using vision. iv* + 97 pp. + app.

49. **Lundin, Lars Kristian.** (1998). *Parallel computation of rotating flows. viii* + 106 pp.

50. **Borges, Pedro.** (1998). *Multicriteria planning and optimization. - Heuristic approaches. xiv* + 219 pp.

51. **Nielsen, Jakob Birkedal.** (1998). *New developments in the theory of wheel/rail contact mechanics. xviii* + 223 pp.

52. **Fog, Torben.** (1998). *Condition monitoring and fault diagnosis in marine diesel engines. xii* + 178 pp.

53. **Knudsen, Ole.** (1998). *Industrial vision. xii* + 129 pp.

54. **Andersen, Jens Strodl.** (1998). *Statistical analysis of biotests. - Applied to complex polluted samples. xx* + 207 pp.

55. **Philipsen, Peter Alshede.** (1998). *Reconstruction and restoration of PET images. vi* + 132 pp.

56. **Thygesen, Uffe Høgsbro.** (1998). *Robust performance and dissipation of stochastic control systems.* 185 pp.

57. **Hintz-Madsen, Mads.** (1998). *A probabilistic framework for classification of dermatoscopic images. xi* + 153 pp.

58. **Schramm-Nielsen, Karina.** (1998). *Environmental reference materials methods and case studies. xxvi* + 261 pp.

59. **Skyggebjerg, Ole.** (1999). *Acquisition and analysis of complex dynamic intra- and intercellular signaling events.* 83 pp.

60. **Jensen, Kåre Jean.** (1999). *Signal processing for distribution network monitoring. xv* + 199 pp.

61. **Folm-Hansen, Jørgen.** (1999). *On chromatic and geometrical calibration. xiv* + 238 pp.

62. **Larsen, Jesper.** (1999). *Parallelization of the vehicle routing problem with time windows.xx* + 266 pp.

63. **Clausen, Carl Balslev.** (1999). *Spatial solitons in quasi-phase matched struktures. vi* + (flere pag.)

64. **Kvist, Trine.** (1999). *Statistical modelling of fish stocks. xiv* + 173 pp.

65. **Andresen, Per Rønsholt.** (1999). *Surface-bounded growth modeling applied to human mandibles. xxii* + 125 pp.

66. **Sørensen, Per Settergren.** (1999). *Spatial distribution maps for benthic communities.*

67. **Andersen, Helle.** (1999). *Statistical models for standardized toxicity studies. viii* + (flere pag.)

68. **Andersen, Lars Nonboe.** (1999). *Signal processing in the dolphin sonar system. xii* + 214 pp.

69. **Bechmann, Henrik.** (1999). *Modelling of wastewater systems. xviii* + 161 pp.

70. **Nielsen, Henrik Aalborg.** (1999). *Parametric and non-parametric system modelling. xviii* + 209 pp.

71. **Gramkow, Claus.** (1999). *2D and 3D object measurement for control and quality assurance in the industry. xxvi* + 236 pp.

72. **Nielsen, Jan Nygaard.** (1999). *Stochastic modelling of dynamic systems. xvi* + 225 pp.

73. **Larsen, Allan.** (2000). *The dynamic vehicle routing problem. xvi* + 185 pp.

74. **Halkjær, Søren.** (2000). *Elastic wave propagation in anisotropic inhomogeneous materials. xiv* + 133 pp.

75. **Larsen, Theis Leth.** (2000). *Phosphorus diffusion in float zone silicon crystal growth. viii* + 119 pp.

76. **Dirscherl, Kai.** (2000). *Online correction of scanning probe microscopes with pixel accuracy.* 146 pp.

77. **Fisker, Rune.** (2000). *Making deformable template models operational. xx* + 217 pp.

78. **Hultberg, Tim Helge.** (2000). *Topics in computational linear optimization. xiv* + 194 pp.

79. **Andersen, Klaus Kaae.** (2001). *Stochastic modelling of energy systems. xiv* + 191 pp.

80. **Thyregod, Peter.** (2001). *Modelling and monitoring in injection molding. xvi* + 132 pp.

81. **Schjødt-Eriksen, Jens.** (2001). *Arresting of collapse in inhomogeneous and ultrafast Kerr media.*

82. **Bennetsen, Jens Christian.** (2000). *Numerical simulation of turbulent airflow in livestock buildings. xi* + 205 pp + Appendix.

83. **Højen-Sørensen, Pedro A.d.F.R.** (2001). *Approximating methods for intractable probabilistic models: - Applications in neuroscience. xi* + 104 pp + Appendix.

84. **Nielsen, Torben Skov.** (2001). *On-line prediction and control in nonlinear stochastic systems. xviii* + 242 pp.

85. **Öjelund, Henrik.** (2001). *Multivariate calibration of chemical sensors. xviii* + 182 pp.

86. **Adeler, Pernille Thorup.** (2001). *Hemodynamic simulation of the heart using a 2D model and MR data. xv* + 180 pp.

87. **Nielsen, Finn Årup.** (2001). *Neuroinformatics in functional neuroimaging.* 330 pp.

88. **Kidmose, Preben.** (2001). *Blind separation of heavy tail signals. viii* + 136 pp.

89. **Hilger, Klaus Baggesen.** (2001). *Exploratory analysis of multivariate data. xxiv* + 186 pp.

90. **Antonov, Anton.** (2001). *Object-oriented framework for large scale air pollution models.* 156 pp. + (flere pag).

91. **Poulsen, Mikael Zebbelin.** (2001). *Practical analysis of DAEs.* 130 pp.

92. **Keijzer, Maarten.** (2001). *Scientific discovery using genetic programming.*

93. **Sidaros, Karam.** (2002). *Slice profile effects in MR perfusion imaging using pulsed arterial spin labelling. xi* + 191 pp.

94. **Kolenda, Thomas.** (2002). *Adaptive tools in virtual environments. - Independent component analysis for multimedia. xiii* + 112 pp.

95. **Berglund, Ann-Charlotte.** (2002). *Nonlinear regularization with applications in geophysics. xiii* + 116 pp.

96. **Pedersen, Lars.** (2002). *Analysis of two-dimensional electrophoresis gel images. xxii* + 162 pp.