Feature Visualization in Large Scale Imaging Mass Spectrometry Data

This manuscript was typeset by the author with the LaTeX 2_ε Documentation System. Text editing was done using the LyX package in a *Cygwin* environment.

The image on the cover shows a voxel representation made with VTK of a spectral datacube obtained using an imaging Fourier transform infrared spectrometer. The framed images on the left of the backside show microscopic and multiple spectrometric samples from crystallized droplets described in Chapter 4 of this thesis. The framed image on the right displays one principal component of a cross-section of a chicken embryo described in Chapter 6.

Cover design by Fred Zurel.

Produced by F&N Eigen Beheer.

Feature visualization in large scale imaging mass spectrometry data /
Alexander Broersen − 2009.

A catalogue record is available from the Eindhoven University of Technology Library.
Ph.D.-thesis − ISBN 978-90-786-7552-5
NUR 980
Subject headings: principal component analysis / feature visualization / mass spectrometry / registration
ACM Computing Classification System (1998) : I.3.5, I.4.3, I.4.10, I.5.3

Feature Visualization in Large Scale Imaging Mass Spectrometry Data

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de
Technische Universiteit Eindhoven, op gezag van de
Rector Magnificus, prof.dr.ir. C.J. van Duijn, voor een
commissie aangewezen door het College voor
Promoties in het openbaar te verdedigen
op dinsdag 3 maart 2009 om 16.00 uur

door

Alexander Broersen

geboren te Hoorn

Dit proefschrift is goedgekeurd door de promotoren:

prof.dr.ir. R. van Liere
en
prof.dr. R.M.A. Heeren

The research reported in this thesis was carried out at CWI, the Dutch national research institute for Mathematics and Computer Science, within the theme Visualization and 3D User Interfaces, a subdivision of the research cluster Information Systems.
This work was carried out in the context of the Virtual Laboratory for e-Science project (`http://www.vl-e.nl`). This project is supported by a BSIK grant from the Dutch Ministry of Education, Culture and Science (OC&W) and is part of the ICT innovation program of the Ministry of Economic Affairs (EZ).

To Jolien

Contents

Preface v

1 Introduction 1
- 1.1 Mass spectrometry . . . 1
- 1.2 Analysis . . . 3
- 1.3 Features . . . 4
- 1.4 Contributions . . . 6
- 1.5 Research objective . . . 7
- 1.6 Approach . . . 7
- 1.7 Thesis outline . . . 8
- 1.8 Publications from this thesis . . . 9

2 Spectral analysis: a survey 11
- 2.1 Introduction . . . 11
- 2.2 Data acquisition . . . 13
  - 2.2.1 Imaging spectrometers . . . 14
  - 2.2.2 Noise . . . 16
  - 2.2.3 Multiple measurements . . . 17
- 2.3 Feature extraction . . . 17
  - 2.3.1 Filtering . . . 19
  - 2.3.2 Selection . . . 22
  - 2.3.3 Classification . . . 25
  - 2.3.4 Comparison . . . 28
- 2.4 Visualization . . . 30
- 2.5 Summary and conclusion . . . 35

3 PCA-based feature extraction 37
- 3.1 Goal . . . 37
- 3.2 Data preprocessing . . . 38
  - 3.2.1 Format . . . 38

3.2.2 Filtering . . . 39
3.2.3 Unfolding . . . 39
3.2.4 Weighting . . . 40
3.3 PCA-based methods . . . 41
3.3.1 PCA . . . 41
3.3.2 VARIMAX after PCA . . . 43
3.3.3 PARAFAC . . . 43
3.4 Results . . . 44
3.4.1 Quantitative comparison . . . 44
3.4.2 Qualitative comparison . . . 46
3.4.3 Performance . . . 48
3.5 Summary and conclusion . . . 51

4 Feature-based registration **53**
4.1 Introduction . . . 53
4.2 Approach . . . 55
4.2.1 Principal Component Analysis . . . 55
4.2.2 Mean Squared Error . . . 56
4.2.3 Entropy . . . 57
4.2.4 Algorithm . . . 59
4.3 Results . . . 60
4.3.1 Two collections . . . 60
4.3.2 Application . . . 61
4.3.3 Comparison . . . 63
4.3.4 Enlarged dataset . . . 64
4.4 Discussion and Future work . . . 64
4.5 Summary and conclusion . . . 69

5 Feature visualization **71**
5.1 Introduction . . . 71
5.2 Related work . . . 73
5.3 Method . . . 74
5.4 Applications . . . 75
5.5 Discussion . . . 80
5.6 Summary and conclusion . . . 82

6 Feature zooming **83**
6.1 Introduction . . . 83
6.2 Related work . . . 85
6.3 Method . . . 87
6.3.1 Binning and PCA . . . 88
6.3.2 Selection and zooming . . . 89
6.4 Results . . . 90
6.4.1 Spectral zooming . . . 92
6.4.2 Spatial zooming . . . 92

6.5 Discussion . . . 93
6.6 Summary and conclusion . . . 94

7 High-resolution feature visualization **95**
7.1 Goal . . . 95
7.2 Parametric feature visualization . . . 96
7.2.1 Extraction and visualization . . . 97
7.2.2 Principal Component Analysis . . . 99
7.2.3 Convolution . . . 100
7.2.4 Correlated geometric shapes . . . 101
7.3 Results . . . 102
7.4 Discussion and future work . . . 105
7.5 Summary and conclusion . . . 106

8 Conclusions and future research **107**
8.1 Conclusions . . . 107
8.2 Directions for future research . . . 108

Bibliography **111**

List of Figures **123**

Glossary of Terms **125**

Index **129**

Summary **133**

Samenvatting (Dutch summary) **135**

Curriculum Vitae **139**

Preface

It's coming up... It's coming up... It's coming up... It's there!
Dare — **Gorillaz**[1]

When I first decided to take up the challenge of a research position at the Center for Mathematics and Computer Science (CWI), I was unsure what it would bring me. It seemed even more uncertain if I would succeed in delivering the expected 'little book' at the end of the journey. But here I am, the work is done. I am so pleased I was able to wrap up all the research in four years. The writing proved a challenge in itself, but it goes to show that perseverance is key. It is time for a change of scenery, but not before I take this opportunity to thank everyone who supported me along the way.

First of all, I would like to thank both my supervisors: Robert van Liere and Ron Heeren. They did not only provide an interesting starting point for a research topic, but also navigated me in the scientific world with their clear vision and inspiration. They gave me every opportunity to find my own way, while keeping an eye on my compass until all the work was done.

At the FOM Institute for Atomic and Molecular Physics (AMOLF), I could always turn to Lennaert and Liam with any questions about the—sometimes overwhelming—area of mass spectrometry. Maarten Altelaar, Els Bon and others proved indispensable: they provided all of the spectral datasets used in the presented examples. For discussions on software engineering at the AMOLF, I would like to thank Ivo Klinkert and Marco Konijnenburg.

I enjoyed the time spent with my roommates Breght, Arjen, and Chris, as well as my other colleagues at the CWI and the members of the visualization group of the Virtual Laboratory for e-Science (VL-e) project. Their stimulating views on visualization issues as well as ideas that came up in discussions provided me with much insight in the technical and academic world.

Special thanks go to the members of the reading committee: Prof. van Wijk, Prof. Roerdink, and Prof. Jansen for their thorough proofreading, which was a tremendous

[1] A sample from Shaun Ryder used in the chorus of 'Dare' from the album 'Demon Days'.

help in improving the presentation of my work in this thesis. In addition, I'm honored to have Prof. Florac, Dr. Luiten and Prof. van Hee as members of my committee as well.

Finally, some acknowledgements on a more personal note. To Jolien, who sacrificed too many evenings providing indispensable support by reading all texts and providing readable alternatives when necessary: I am infinitely grateful. Naturally, any remaining mistakes are my own. I would also like to thank family and friends (you know who you are) for their interest and—especially my *paranimfs* Jolien and Paul—their limitless support: thank you for being there.

Amsterdam, *Alexander Broersen*
January 2009

Chapter 1

Introduction

Imaging mass spectrometry is a powerful technique to measure the spatial distribution of molecular content on complex surfaces of samples. It combines high-resolution microscopic imaging tools with the analytical capabilities of spectrometry. The resulting measurements can be used for microscale analysis. The size of these measurements is ever increasing, since developments in spectrometry instrumentation allow for data acquisition in continually higher mass and spatial resolution.

Since current visualization techniques can not yet fully utilize the increasing resolution and size of the measured datasets, new techniques have to be developed. With enhanced visualization techniques, analysis of complex datasets can be supported and improved. This thesis aims to show that analysis of imaging mass spectrometry data can be improved by introducing new approaches for automatic selection and visualization of features in large scale datasets.

1.1 Mass spectrometry

Mass spectrometry (MS) is the process of measuring atoms and molecules present in a material by determining the mass and charge of their ions. The presence of a combination of different ions can uniquely identify a material. Basically, this technique can be compared to measuring the different wavelengths present in (visible) light. Each wavelength—or color in the case of visible light—has a certain intensity. The combined intensities of all wavelengths determine how the color of a beam of light is perceived. A prism is able to separate a beam of light into its spectral colors creating a spectrum in which colors are arranged according to their wavelength. Similarly, a mass spectrometer is able to break up ions from the surface of a physical sample material. The distribution of these separated ions is represented in a mass spectrum in which ions are arranged according to their relative mass.

Varying intensity values in a mass spectrum represent the presence of (molecular fragments of) chemical compounds. Measured ions with nearly identical masses and charges are grouped together in order to create peaks in the mass spectrum. Each peak indicates the presence of a chemical compound with a particular mass. The height of a peak indicates the amount of ions measured in relative proportion to the

height of another peak. A mass spectrum can be expressed as the function $f(m)$ where m represents the mass and $f(m)$ denotes the intensity value in the spectrum on that particular mass. Combinations of peaks in a mass spectrum create different specific spectral profiles. Each profile uniquely describes the composition of a chemical compound in a material sample. This spectral profile is similar to a spectrum within visible light, but with a composite of chemical properties rather than colors.

In *Imaging* MS [Ben87; McD07], the location (x,y) of each mass spectrum is added, which results in a three-dimensional (3D) dataset $F(x, y, m)$. The x and y coordinate of the position on the surface of a particular sample is measured for each of the mass spectra. The measured values can be combined to create one 'spectral datacube'. A spectral datacube could be compared with a digital color picture composed of three color bands, for instance red, green and blue (RGB). Since their RGB-value on each position is known, their combined intensity creates a specific color on that location in the picture. Each separate color band shows how a color is distributed in the picture as a single intensity image. Instead of creating a digital picture with the colors from a sample, imaging mass spectrometry measures the distribution of the chemical compounds on the surface of a sample. Similar to the terms 'color band' or 'color channel', each position m in a mass spectrum is called 'spectral band' or 'spectral channel'.

An expert spectrometrist uses mass spectra to deduce the chemical, physical, or even biological properties of the compounds present in a sample of an unknown material. Analysis of a spectral measurement is based on the presence of peaks in the intensity in the mass spectrum. These peaks have to be located, interpreted, and compared to other peaks in the mass spectrum to be able to characterize the chemical structure in the sample material. Besides comparing spectral peaks, spectral datacubes also enable a mass spectrometrist to compare peak intensities on different locations. The heights of the peaks are used to create an image with the spatial distribution $F'_m(x, y)$ of a single spectral band m in the spectral datacube.

When different peaks in a spectrum show a similar spatial distribution, these peaks could originate from the same compound molecule. This principle can be illustrated with a simplified example of the chemical compound 'sodium chloride' (NaCl), also known as table salt. After a mass spectral measurement of a sample that contains this compound, there is a peak in the acquired mass spectrum around 23 u (unified atomic mass unit) for the sodium ion (Na^+) and two peaks around 35 u and 37 u for chloride ions ($^{35}Cl^-$ and $^{37}Cl^-$). When both sodium and chlorine peaks occur with a similar spatial distribution in the measured dataset, it is likely that they originate from the salt crystals within the sample. The heights of the peaks are a measure for the amount of ions that are present. If the sodium peak in this example is 100%, the $^{35}Cl^-$ peak would be at 75% and the $^{37}Cl^-$ peak would be at 25%, according to the natural ratio in which these isotopes exist.

Imaging MS has many useful applications for microscale analysis of cells and tissue sections from biological samples. Mass spectral measurements enable detection of differences in the molecular composition of the surface of a sample material. These differences may be used to determine whether a tissue sample is healthy or contains tumors. Besides detecting chemical differences, imaging MS allows for spatial localization of differences within a tissue sample on a high resolution. The spatial distribution of, for instance, different peptides and proteins can be obtained. The ability

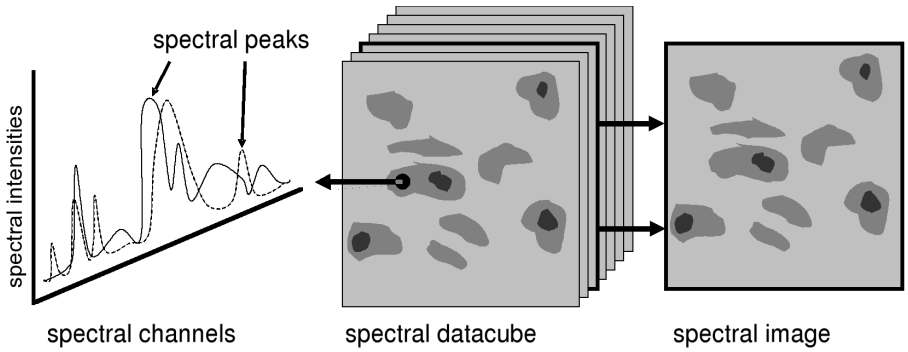


Figure 1.1: *A schematic of the spectral view on peaks and spatial view on an image, both from the same 3D datacube.*

to classify cells on a molecular level can provide more insight into the effect of, for instance, disease, drug treatment, or environment on the metabolism in biomedical or pharmaceutical research.

1.2 Analysis

Traditionally, spectral datacubes are analyzed by switching between the mass spectrum and images of a single spectral band. This way, interesting peaks, spatial distributions, and similarities between peaks and their locations can be detected. First, an overall spectral view of the datacube is used to locate peaks of interest. Figure 1.1 shows a schematic representation of several spectral peaks. After peak detection, images of selected peaks are created in order to examine their—combined—spatial distributions (as can be seen on the right of Figure 1.1). For instance, two different cells can be recognized in the schematic representation of the spatial distribution of the peaks. One or more interesting locations in the spatial view can be selected to examine their spectral profile and characterize the material on a certain location. An illustrative representation of the complete spectral datacube used to create both views is shown in the middle of Figure 1.1.

Current imaging MS techniques produce spectral datacubes with many variables in the spectral as well as both spatial dimensions. Exploration of mass spectral datacubes is not an easy task, because every peak and spectral image has to be examined in order to find potentially interesting spectral peaks or distributions. Existing approaches in spectral analysis apply data-reduction techniques to simplify the exploration of spectral datacubes by reducing and optimizing the variables. Dimension reduction is the process of reducing the number of variables in a dataset. Common practice is to apply dimension reduction to create a smaller dataset in which the most interesting properties in the data are grouped—also known as a feature space—in both spectral and spatial dimension. This way, the complexity is reduced and the accuracy for analysis is improved [Gra06].

Dimension reduction can be divided into three phases: filtering and variable selection, and model-fitting. Filtering and variable selection improves the quality of the data by removing noise. The quality can be improved by for instance techniques like

down-binning, finite impulse response filtering, (de-)convolution, and wavelet filtering. Other common variable selection techniques to reduce the influence of noise are for example thresholding, spectral peak-picking, or data-decomposition strategies. In the second phase, a model is fitted with the selected data in order to extract a subset of variables that describe the original data. Different methods for multivariate analysis can be used to describe the data with less variables, for instance correspondence analysis, principal component analysis, factor analysis, independent component analysis or variations of these. The general problem with the choice of a dimension reduction technique is the need for accurate models for noise estimation and fitting of the data onto the new variables. Before an appropriate reduction technique can be evaluated, the characteristics of the desired feature-space for imaging mass spectrometry have to be defined first.

1.3 Features

This thesis expresses its objectives and approach in terms of 'features'. We define a feature in imaging MS as:

> one or more distinct spectral peak(s), recurring at several locations in a recognizable spatial pattern

Multiple peaks are linked together as a feature when they recur with the same ratio of intensity on several locations. These peaks should be linked when induced by the presence of the same chemical compound. This way, a chemical compound is represented by a single feature. The intensity values of each of these peaks can be organized in an image, which is the spatial distribution of the feature. A pattern is recognized if this spatial distribution resembles a known image of, in this case, a biological sample. An expert biologist can determine if the spatial distribution has a recognizable pattern which is specific for the sample that is being measured. An example of a feature in a sample of nervous tissue is, for instance, the distribution of cholesterol at the edges of a specific group of neuronal cells. With this definition, we can focus our aim to the detection, visualization, and use of features in the analysis of spectral datacubes.

Detection

The detection of features depends on the detection of spectral peaks. A spectral peak is identified examining neighboring intensity values in a mass spectrum. After identification, different peaks have to be compared to find similarities, for example a recurring ratio between intensities of different peaks. Similarities can be found by comparing the spatial intensity distributions of different identified peaks or by applying statistics on the spectra. Similar peaks are selected and grouped together in a single feature.

Feature detection must be robust when a low signal-to-noise ratio is present in a spectral datacube. Robust feature detection is sensitive in identifying peaks and specific in selecting peaks to be grouped in a feature. Peak identification is complicated by noise, as it can lower the intensities within a peak (also known as a ' signal') and raise the intensity of the neighboring values. The ratio between peak height and

intensity of noise is called the signal-to-noise ratio. Moreover, a feature should not contain every peak detected: only those peaks of which the intensity ratio has a significant contribution to a particular feature should be selected. When many peaks with an insignificant contribution are selected, a feature becomes cluttered and less distinctive.

Visualization

A detected feature is traditionally visualized with two separate views: a spectral and a spatial view. All spectral peaks of a feature are visualized in a spectrum (the spectral view). The spatial distribution of these peaks is visualized in an intensity image (the spatial view). Although both views are visualized separately, they are related as they represent the same feature from two perspectives. A single spectral peak can be distributed among several spectral bands. To be able to view the spatial distribution of an entire spectral band, all images on those spectral bands are added together to create one intensity image.

Visualization of features has to be accurate with proper contrast in both the spectral and spatial view. When spectral datacubes have a sparse distribution of intensity values, it is common practice to add all spectral intensity values from a single peak to improve the contrast between this peak and the neighboring spectral bands. This way, the spectral peaks can be visualized with more accuracy. Similarly, the accuracy and contrast in the spatial view can be improved by adding neighboring intensity images. With this technique, however, structural information is lost in both views. The individual spectra can not be distinguished from each other by removing the spatial dimension in the spectral view. Similarly, individual peaks can not be distinguished in the combined intensity image.

Interpretation

After feature detection and visualization, identification and interpretation is left to an expert, who should be enabled to select, zoom in on, and compare different visualized features in an analysis. This is important in order to be able to place features in the right perspective. Each spectral measurement is performed with a different hypothesis. Identifying and interpretation which features are important in a measurement may differ as well. Therefore, a user should make a final selection of appropriate features. Some basic tools have to be available to assist identification and interpretation.

A user should be enabled to select appropriate features and exclude uninteresting features in a particular measurement. A less detailed, top-level view of the complete dataset makes it possible to place different features in the same perspective. After this, potentially interesting features can be selected for further inspection. Different, more detailed features could exist within a selection and can be selected by the implementation of a zooming function. A final selection of features should be visualized in one combined overview to be able to compare their spectral properties and spatial distributions. This comparison is the most accurate when features can be compared on the highest level of detail.

The following requirements on the detection, visualization, and analysis of features from imaging mass spectrometry data are identified:

- robust peak identification and selection requires a high signal-to-noise ratio;
- a more accurate representation of features requires a combined view of the spectral and spatial properties;
- users should be enabled to zoom in on and compare multiple features.

1.4 Contributions

Recent technological developments not only allow mass spectrometric imaging at higher spatial resolution, but also with shorter acquisition times, larger surfaces, and higher spectral resolution. Because of the large amount of detail produced with these spectrometric techniques, manual analysis of the data is an intensive and error-prone task. In many cases, the measurement itself is not the most time-consuming, but analysis of the results becomes more elaborate due to the large amount of data obtained. New tools and techniques for reducing and processing these large datasets have to be developed to support analysis.

Many dimension reduction methods are already available to transform data from imaging spectrometry into a feature space (a smaller dataset in which characteristics remain present). Each filter and decomposition method makes implicit and explicit assumptions about the underlying mathematical model of the data. Different parameters control a model's explicit assumptions. In order to create a generic approach and keep the parameter-space as small as possible, the model should have as few explicit assumptions as possible. This way, feature detection does not depend on the type of sample measured. The multiple requirements mentioned in the previous section can be met with this feature-based approach.

The key strategy in this work is the application of Principal Component Analysis (PCA) for automatic feature detection in spectral datacubes. PCA is a simple approach for dimension reduction and feature selection against a low computational cost (see Section 2.3.3). It is non-parametric and uses statistics to create a stable and unique solution that is able to extract different 'components' from a dataset. Each extracted component consists of related spectral peaks accompanied by their combined spatial distributions. Given the definition of a feature, it can be stated that each extracted component potentially contains a feature. Still, the resulting components have to be inspected to determine whether or not they contain interesting features. Therefore, improved visualization techniques are needed to locate potential features in extracted components.

In this thesis, a wide range of visualization techniques based on automatic feature extraction is presented. With extracted features, datacubes can be aligned automatically. In this application, datacubes can originate from measurements in different areas of the same physical sample. Since interesting artifacts can be distributed among several datacubes, they may be not be recognized when viewed partially. Therefore, datacubes have to be spatially aligned and combined to create a complete overview of the data. With extracted features, transfer functions can be generated. With these functions, spectral and spatial properties of a feature are highlighted simultaneously within a single 3D view of the datacube. Features can be used to zoom in on and extract specific parts within a spectral datacube on higher resolutions. The

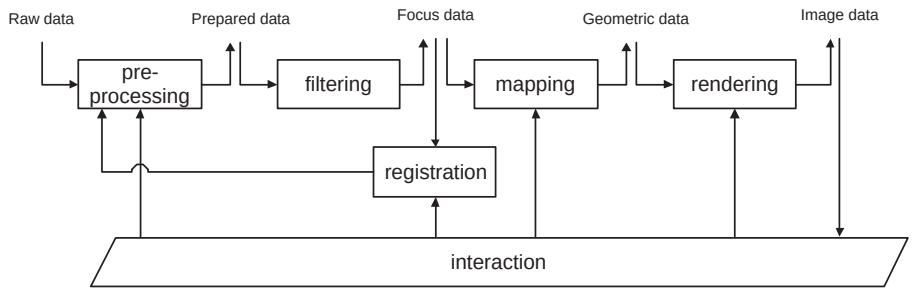


Figure 1.2: *The dataflow in the visualization pipeline of this approach.*

resulting high-resolution features are parametrically visualized as 3D abstract geometric shapes. This improves analysis, because it enables an analyst to compare and examine several features on different levels of detail.

1.5 Research objective

The central research objective addressed in this thesis is: to combine PCA with visualization techniques to solve detection, visualization, and analytical issues in the analysis of imaging spectrometry data. This objective is specified by the following research questions:

- How can PCA be used for robust feature detection in large imaging spectrometry datasets?
- How can features be used to improve registration, zooming, and visualizations of spectral datasets?

1.6 Approach

An overview of our approach can be visualized in a dataflow model (in Figure 1.2), based on the extended 'visualization pipeline' of Dos Santos and Brodlie [San04]. The preprocessing step in their model is computer-centered and provides computational methods to fit a model and enrich data. In our approach, this preprocessing step is user-centered to allow for data-enrichment on a specific level of detail. Then, the prepared datacube is filtered to be able to select different sections of the data to be mapped for visualization. According to Dos Santos and Brodlie, this is a user-centered step, because an appropriate model and a number of parameters has to be defined for filtering. We use PCA for filtering, since it is able to automatically extract different features without any parameters. In the next step in this visualization process, the filtered data (also 'focus data') is mapped onto abstract representations. In our approach, focus data (or components) are mapped either by a transfer function or a parametric definition of geometrical shapes. Finally, after rendering, an expert can select, compare, and interpret components with interesting features on different levels of detail.

In our approach, we added registration to the dataflow model. In this step, several spectral datacubes are aligned and combined into a single dataset in order to spatially extend datacubes. The feature-based alignment takes place after feature detection on the focus data (in this case the extracted principal components). When the same feature is present in different datacubes, these cubes can be spatially aligned after which an extended datacube is created from the two original (raw) datasets. Again, the different steps in our visualization pipeline are used to analyze the extended datacube and improve results in feature detection.

1.7 Thesis outline

The chapters in this thesis are structured according to the steps mentioned in the approach in the previous section. Each chapter focuses on a different feature-based technique in order to address different issues concerning visualization of mass spectrometry data.

First, a background survey on spectral analysis is provided in Chapter 2. It includes a general introduction of data acquisition in imaging spectrometry. Different methods for the extraction of features are compared as well as different approaches for the visualization of spectral data.

Chapter 3 presents an overview of PCA and PCA-based methods for detecting and extracting features from spectral datacubes. We discuss preprocessing of mass spectral data, PCA, additional rotational optimization, and a method for factor regression. The results are compared quantitatively and qualitatively, together with some performance characteristics.

In Chapter 4, a robust method for automatic feature-based registration is developed. First, features are detected using PCA. Then, an additional signal quality metric ensures that only those regions with enough signal are considered by a similarity metric. Several spectral datacubes are combined to provide better detection and extraction of features.

Chapter 5 describes a visualization technique that applies PCA to create transfer functions for volume rendering of a spectral datacube. These volumetric visualizations enable us to observe and explore features with connected spectral and spatial properties in a single 3D view. Applications demonstrate the additional value of these visualizations.

Chapter 6 presents a technique for spectral and/or spatial zooming of extracted features. This technique is especially useful for spatially extended datasets, when using the method presented in Chapter 4. Features of interest can be selected for further analysis on different levels of detail. Moreover, features with unwanted artifacts can be removed to reduce noise.

Chapter 7 provides an approach to visualize features in 3D with distinct boundaries and at the highest resolution possible. Three parameters regulate the selection of similar spectral peaks, the level of detail, and the size of the extracted feature shapes. An application shows how resulting features are visualized and interpreted.

Finally, conclusions regarding the objectives in this thesis are presented in Chapter 8. In addition, directions for future research are proposed.

1.8 Publications from this thesis

Most chapters in this thesis are based on the following publications, which appeared in the peer-reviewed proceedings of international conferences and journals.

- A. Broersen, R. van Liere and R. M. A. Heeren, Comparing three PCA-based Methods for the 3D Visualization of Imaging Spectroscopy Data, *Proceedings of the IASTED International Conference on Visualization, Imaging, & Image Processing*, 2005, pp. 540–545. [Bro05b] (**Chapter 3**)

- L. A. Klerk, A. Broersen, I. W. Fletcher, R. van Liere and R. M. A. Heeren, Extended Data Analysis Strategies for High Resolution Imaging MS: New methods to deal with extremely large image hyperspectral datasets, *International Journal of Mass Spectrometry*, 2007, 260(2–3), pp. 222–236. [Kle07] (**Chapter 3**)

- A. Broersen and R. van Liere, Feature Based Registration of Multispectral Datacubes, *Proceedings of the IASTED International Conference on Visualization, Imaging, & Image Processing*, 2006, pp. 543–548. [Bro06] (**Chapter 4**)

- A. Broersen, R. van Liere, A. F. M. Altelaar, R. M. A. Heeren and L. A. McDonnell, Automated, Feature-based Image Alignment for High-resolution Imaging Mass Spectrometry of Large Biological Samples, *Journal of the American Society for Mass Spectrometry*, 2008, 19(6), pp. 823–833. [Bro08a] (**Chapter 4**)

- A. Broersen and R. van Liere, Transfer Functions for Imaging Spectroscopy Data using Principal Component Analysis, *Proceedings of the Eurographics / IEEE VGTC Symposium on Visualization*, 2005, pp. 117–123. [Bro05a] (**Chapter 5**)

- A. Broersen, R. van Liere and R. M. A. Heeren, Zooming in Multi-spectral Datacubes using PCA, *Proceedings of the SPIE / IS&T Symposium on Electronic Imaging*, 2008, pp. 68090C. [Bro08b] (**Chapter 6**)

- A. Broersen, R. van Liere and R. M. A. Heeren, Parametric Visualization of High Resolution Correlated Multi-spectral Features Using PCA, *Proceedings of the Eurographics / IEEE VGTC Symposium on Visualization*, 2007, pp. 203–210. [Bro07] (**Chapter 7**)

Chapter 2

Spectral analysis: a survey

Chapter 1 introduced the analysis and visualization of imaging mass spectrometry datasets. An approach was proposed to do spectral analysis by feature detection, visualization of the resulting features and analysis of a selection by expert interpretation. The objective of this thesis was stated together with an outline of the approach taken to reach that objective.

This chapter provides a more detailed overview of current approaches and corresponding methods and techniques for the analysis of spectral data. A comparison is made of these existing methods and techniques with their purpose, strengths, and weaknesses. A minimal subset of tools is chosen to be able to implement the suggested exploratory visualization approach for analysis of spectral imaging data.

2.1 Introduction

The purpose of spectral analysis is to extract information of the molecular composition of a material of interest. In general, spectral data describe the interaction between matter and radiation as a function of either wavelength or frequency. One can determine properties of the matter by measuring the absorption, emission or scattering of radiation. Therefore, the materials of interest could be located far away, for instance on the surface of stars or planets in the field of astronomy or remote sensing. On the other hand, the materials of interest can also be microscopically small. In all cases, these chemical substances can be analyzed by examination of their spectra. The analysis in this thesis is limited to the discovery of features in biological samples with the help of strategies and tools for visualization. By comparing existing methods, we have to choose which would be most appropriate for this approach.

It depends on the purpose of the analysis and the state and location of the material of interest which method—or combination of methods—is the most appropriate. The goals, parameters and limitations of an individual spectral measurement are too versatile and the results too complex to be able to perform analysis without expert knowledge [Har84]. Therefore, tools for analysis and visualization have to be developed that facilitate the discovery and interpretation of extracted features. This versatility of goals complicates making an exhaustive comparison between currently

implemented methods for analysis. The final step in the analysis and interpretation will be left to an expert spectrometrist. By a visual presentation of the results, the complexity can be reduced and a user can gain more control on and insight into the extracted features.

A spectral dataset can be modeled in terms of pure spectral profiles (also called 'spectral endmembers' in the field of light spectroscopy) and/or spectral images (also called 'abundance images' in the field of imaging light spectroscopy) [Kes03]. A pure spectral profile is the spectrum resulting from one material or chemical compound in the case of mass spectrometry. The ratio between the intensities of the peaks in such a pure spectrum is always the same and can be considered a basic building block of a chemical compound. Each material or chemical compound has a certain concentration. These concentrations can vary on different spatial locations in a spectral datacube. A general linear function model that describes a spectrum $f[m]$ is

$$f[m] = \sum_{n=1}^{N} \beta_n X_{mn} \tag{2.1}$$

where m is the independent variable of a spectral band, β a vector with concentrations of the chemical compounds and N the number of distinct chemical compounds present in the dataset. An overview of the most important variables is shown in Table 2.2. The coefficients in the columns of matrix X_{mn} are the pure spectral profiles. Depending on the chemical compound, each spectral profile consists of one or more peaks.

According to Lohnes [Loh98], spectral analysis can be performed from two different points of view: a quantitative and a qualitative view. This work focuses on a qualitative exploration prior to a quantitative analysis by an expert. This focus is chosen, because identifying unknown constituents present in complex surfaces on samples is often too complex to be modeled completely automatically. The complexity is caused by the large number of different pure spectral profiles and the large number of peaks that can be present in one spectral profile. When only the presence of a chemical compound has to be detected, there is no need to fit an exact, complex model on the spectral data. The search for and detection of chemical compounds in a spectral dataset can be called qualitative exploration.

Commonly used sequential stages in an approach for a qualitative analysis are: data acquisition, feature extraction, and the visualization of features as shown in Figure 2.1. The first stage is the acquisition of the spectral data from a material sample. In the second stage, specific features have to be selected and extracted from the complete dataset. Those features are visualized in the third stage to be interpreted

variable	range
spectral variable	$m = 1 \dots M$
compound/component	$n = 1 \dots N$
horizontal location	$x = 1 \dots X$
vertical location	$y = 1 \dots Y$
spatial coordinate	$xy = 1 \dots XY$

Table 2.1: *Different important variables with their ranges.*

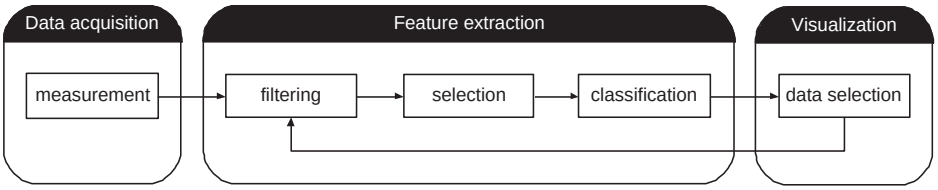


Figure 2.1: *Three stages in a spectral analysis.*

by an expert. The expert should be enabled to make a selection to explore in more detail, after which the complete cycle of feature extraction and visualization starts again.

The data in the process from data acquisition to visualization can be described by different mathematical functions. An overview of these descriptions is shown in Table 2.2. As mentioned before, there are many techniques available to support the extraction of features. To be able to compare them more easily, they can be divided in three subsequent categories: for the process of filtering, selecting features, and classifying features. The different techniques and methods used in these stages are described in more detail in the next three sections.

2.2 Data acquisition

Spectral data is acquired by a spectrometer. Whereas spectro*scopy* is a more general term to describe the study of spectral data, spectro*metry* usually refers to the actual process of measuring spectral data. These measurements can be classified according to the spectrum emitted from or absorbed by a material in some form of energy. This energy can be measured in the form of electromagnetic radiation (e. g., light), acoustic, electrons, or ions. There are many different types of spectrometers, each with its own characteristic properties and specific output of spectra, for instance the number of spectral bands. *Imaging* spectrometers have the added functionality of obtaining spectra for a large number of positions separately, where these positions

data type in the process	function	parameters
continuous spectrum	$f(m)$	m: spectral variable
spectral datacube	$F(x,y,m)$	x, y: spatial coordinates, m: spectral variable
spectral image	$F'_m(x,y)$	x, y: spatial coordinates, m: spectral variable
spectral noise	$\tilde{f}(m)$	m: spectral variable
multiple datacubes	$F'(x,y,m)$	x, y: spatial coordinates, m: spectral variable
discrete spectrum	$f[m]$	m: spectral variable
filtered spectrum	$\hat{f}[m]$	m: spectral variable
selected spectra	f'_k	k: number of selections
decomposed spectra	$P_{n\times m}$	n: number of components, m: spectral variables
decomposed distributions	$Y_{n\times xy}$	n: number of components, x, y: spatial coordinates

Table 2.2: *Different mathematical descriptions used in the process.*

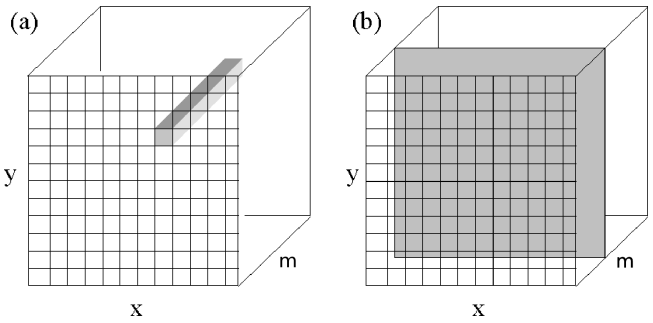


Figure 2.2: *(a) A single spectrum and (b) a single image in the spectral datacube.*

are typically arranged in a regular grid. This way, the resulting dataset has the format of a spectral datacube: intensity values having two spatial (x and y) and one spectral coordinate (denoted as m). A representation of a spectral datacube is shown in Figure 2.2. The grey areas in the datacube represent (a) a single spectrum on one location and (b) a single image on one spectral band.

Typically, spectral image data lends itself best for a qualitative analysis approach, as the spatial appearance can provide substantial information for a correct classification. A spatial map of each spectral band can be obtained and spectral data with added spatial information also provides more possibilities to provide a better view on the noise that is present in a measurement. This knowledge opens up more opportunities to reduce the influence of noise for feature extraction in a spectral analysis.

Each method for spectral imaging results in a different kind of spectral dataset with different resulting quantities (e. g., wavelength, energy), ranges, resolution, dimensions and of course different influences of noise. It is common practice in spectral analysis to make multiple measurements of the same material of interest. A larger region on the material can be covered by changing the spatial offset of each measurement, as well as the reduction of noise by comparing duplicate results. The following subsections will elaborate on different imaging spectrometers, noise, and the use of multiple measurements.

2.2.1 Imaging spectrometers

There are many types of imaging spectrometers. Each uses a different spectral imaging technique in remote sensing. Landsat's Thematic Mapper (TM), for instance, records 8 spectral bands. NASA's Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) can record 244 spectral bands [Geb98]. There are also different techniques for acquiring spectra from small biological samples on a higher resolution [Alt07]. One technique is 'Fourier Transform InfraRed' (FT-IR) imaging spectroscopy [Lev05; Wee02]. A second technique is the Time-of-Flight Secondary Ions Mass Spectrometry (TOF SIMS) [Vic02], which is primarily used in this thesis. It can be used in combination with the Matrix-Assisted Laser Desorption/Ionization (MALDI) technique [Kar88]. When TOF SIMS is used in an additional Large Area Mosaic mode [McD07], specific locations with a higher resolution compared to a normal measurement can be recorded. Whereas imaging spectroscopy measures a light spectrum according to its

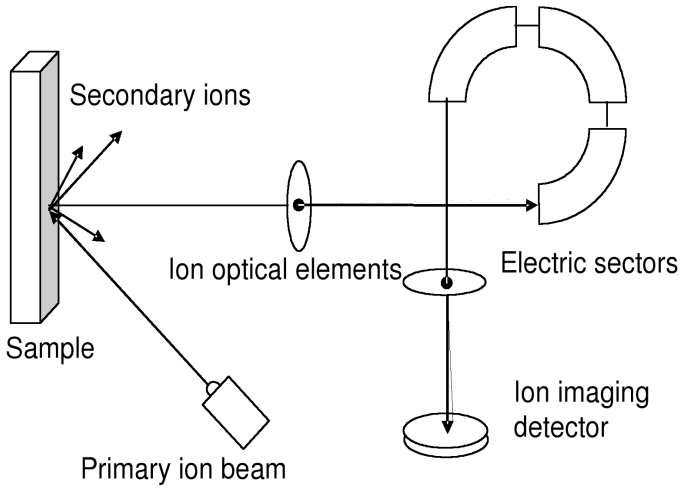


Figure 2.3: *Schematic representation of an imaging TOF SIMS instrument.*

photon energy, a mass spectrum can be measured by, for instance, the time it takes ions to reach a detector. This amount of time depends on the mass-to-charge ratio of the ion (see Figure 2.3) from which the particle can be identified.

Mass spectrometers can produce millions of spectral measurements from the chemical compounds present on the surface of a recorded sample. The primary ion beam in Figure 2.3 scans the spatial surface of a sample by removing ions from it. If a surface is measured for a longer period of time, there are more ions removed from the surface. The resulting raw spectral data is basically a cloud of single 3D points where each point (x, y, m) represents one measured ion. These 3D points can be used to create a 3D datacube $F[x, y, m]$ in which the 3D points are put into 3D cells in a spatial grid with Cartesian coordinates (x, y) with a certain spectral channel m. Measurements that have the same location and the same spectral channel are summed in a single cell. Most ($\sim 90\%$) of the 3D cells within this mass spectral datacube do not contain a value. This is in contrast with the spectral datacubes resulting from measurements that involve capturing light on different frequencies. Each 3D cell in these mass spectral datacubes normally contains a scalar value for the intensity of a particular spectral channel on a location.

The first developments of the Secondary Ions Mass Spectrometry (SIMS) technique appeared in the early 1940s. These experiments were used to analyze oxides and metals [Ric07]. Fifty years later, improvements led to the preservation of the spatial relationship between the ions. This resulted in the first applications for imaging MS in which the mass spectra of an entire spatial region could be obtained. Surface analysis of biological material was first demonstrated by Benninghoven [Ben94]. Innovations on enhanced ion generation made it possible to improve spatial resolution and quality of the resulting spectral images [Pac99]. Single cell and direct tissue analysis by investigation of large intact biomolecular species became possible with imaging MS as a proteomic tool [Aeb03].

This thesis focuses mainly on the visualization of data from imaging mass spec-

trometers, because of their capability to measure at a high spectral and spatial resolution. A mass spectrometer can record the presence of particles on a molecular and atomic scale, which results in complex and large datasets. Moreover, its ability for imaging is essential, because the location of each identified substance in a biological sample is important for the recognition of the spatial pattern of a subject of interest by an expert. Also, imaging enables us to reduce noise in a measurement using the additional available spatial information.

2.2.2 Noise

Inherently, every measurement contains a desired signal and some degree of noise, usually expressed as the Signal-to-Noise ratio (S/N) in the field of spectral analysis. The signals are in this case the spectra in the datacube. Intrinsically, a measured spectrum $\tilde{f}(m)$ consists of part signal $f(m)$ (from Equation 2.1) and part distortion ϵ_{mn} as defined in

$$\tilde{f}(m) = f(m + \tau) + \epsilon_{mn} \tag{2.2}$$

Besides the signal-independent distortion ϵ_{mn} on spectral band m and spectral profile n, there is an additional noise factor τ that results in a spectral shift of peaks in a spectrum. This spectral shift causes a broadening of spectral peaks if all spectra in a datacube are added together in a spectral view on the datacube. The independent distortion can be fitted on a chemical model of the sample, but can also be modeled by applying the right statistics on the chemical model, if the nature and variability of the noise is known.

Noise is any unwanted signal interfering with a desired signal and can be observed by differences between expected and measured intensity values. In spectral datacubes, noise is not only present in a spectrum, but also in between spectra. It usually has a Gaussian distribution. In TOF SIMS however, it is Poisson distributed, because this technique is event-based. This distribution is caused by the uncertainty associated with the rate of arrival of ions at the detector. Other sources of noise in TOF SIMS include: chemical noise, electrical noise, shot noise, and calibration noise. Chemical noise is caused by unwanted artifacts (substances or contaminants) on the surface of a sample, due to preparation or impurities. It can also refer to the material in which a specimen of interest is embedded, called the 'matrix material'. One can say that shot and electrical noise are caused by the physics of an instrument, namely by the distortions at the ion source or internal noise in the circuits of the detector. The random noise in a measurement is mainly caused by the electrical variability within the detector. Variation in the height of the surface of a sample can be observed by a small shift of τ in the mass spectrum. This phenomenon can be perceived as calibration noise in the spectral dimension.

In the ideal case, an analysis is not influenced by noise in the data. However, all techniques for measurement of spectral data invariably are. It complicates the extraction of features from the data as it is uncertain whether or not a feature is either noise or an interesting occurrence in the measured sample. Therefore, noise reduction techniques have to be applied to reduce the influence of noise on the analysis as much as possible. Many filters are available to reduce different kinds of noise in signal, image, and volumetric datasets. All filters need prerequisite information about the nature of the noise, which varies in almost every experiment. Instead of filtering,

it is also possible to reduce noise by removing specific parts from a dataset that mostly contain unwanted artifacts. In this qualitative approach, the interpretation of a feature is left to the experience of an expert rather than the design of tailor-made noise reduction techniques. The influence of noise can be reduced by methods of feature extraction, but also by taking advantage of having multiple measurements of the same sample.

2.2.3 Multiple measurements

Multiple measurements with a different spatial offset can be taken from the same sample. This is a common strategy in imaging to be able to image a larger spatial area. A spatially enlarged datacube

$$F'[x, y, m] = \sum_{n=1}^{N} F_n[x_n, y_n, m] \tag{2.3}$$

is created, where N is the number of spectral datacubes and (x_n, y_n) are the spatial coordinates of the added spectra. For instance, satellites are taking several separated images of earth. It is not possible to create an image of the complete surface with the same quality and resolution of each separate image. When put together, these images form a high-resolution map of the complete surface of the earth. However, this approach requires that images are fitted together correctly, which process is referred to as 'registration'. The strategy of taking multiple measurements is also applied to spectral imaging of biological samples. The different measurements have to be registered first in order to take advantage of this strategy. Unfortunately, most imaging mass spectrometers are not able to provide a precise offset between two different measurements, if any at all.

It is important to determine the offset between multiple measurements, because it allows for the creation of one combined dataset with larger spatial dimensions. With these larger dimensions, feature extraction could improve, because the number of measurements increases. Moreover, since there are several measurements of the same region, noise can be reduced in overlapping regions.

The field of image processing offers many approaches for registration of images. However, current literature does not provide any examples of implementation of registration of imaging mass spectrometry data. This could be caused by the unique nature of mass spectra (the difference in quality of spectral images due to their specific noise). In our approach (analysis by feature exploration), we attempt to register spectral datacubes using feature extraction. Several measurements are combined into a single dataset which would contain less noise than the individual ones. This will improve the extraction of features from a combined spectral datacube.

2.3 Feature extraction

After data acquisition, features can be extracted. Many extraction methods are developed in 'chemometrics', each with different prerequisites, goals, and performance considerations [Lis05]. An overview of the objective and limitations of each exploratory method has to be created, before one or more appropriate methods are selected for

this approach. The description of a feature in Section 1.3 can be reformulated with Equation 2.1 as:

> one or multiple correlated column(s) in X_{mn} that create a recognizable spatial pattern with their intensity values.

In this description, X_{mn} is the matrix with pure spectral profiles. In other words, a feature represents one or several correlated chemical compound(s), which spatial distribution can be recognized. Two chemical compounds are correlated if there is a linear relation between the peaks in their spectral profiles. The correlation between these spectral variables can be expressed in terms of 'multicollinearity'. This refers to a situation in which the correlation coefficient between two or more independent variables is equal to 1 or -1 (positively or negatively correlated). Positively correlated spectral variables indicate that the two chemical compounds represented by those variables could originate from the same molecule. Negatively correlated spectral variables indicate the presence of mutually exclusive chemical compounds. When the correlation coefficient is equal to 1 of -1, these variables are linearly dependent and called 'collinear'. In this case the relationship $\beta_1 X_{m1} + \beta_2 X_{m2} + \cdots + \beta_m X_{mn} = 0$ exists, where $\beta_m \in \mathbb{Z}$ are constants and X_m are the explanatory (in this case the spectral) variables. In our definition of a feature, if two or more spectral profiles are collinear, they are put in one single feature. This way, chemical compounds with the same spatial distribution are put together because it is likely that there is a relation between these compounds. In this approach, the objective of feature extraction is to automatically highlight these relations and identify them by studying their spatial patterns.

Methods for extraction can be categorized by a large number of different properties, resulting in a diversity of taxonomies. Unfortunately, many methods do not seem to belong exclusively to a single category. Therefore, in spectral analysis, methods are mainly categorized according to the consecutive steps necessary for the process of feature extraction. These steps are for instance: dimension reduction, endmember determination, and inversion to estimate the fractional abundance of the endmember spectra [Kes03]. Other methods [Hil06] have a preprocessing phase (such as smoothing and peak detection) prior to a classification phase. From a system-processing point of view, methods can be classified according to their input, output, model description, and constraints. A model describes the statistical structure in a function by mathematical rules. The taxonomy presented in this section classifies feature extraction methods with as little overlap as possible, according to a specific partial goal within the process of extraction.

A hierarchical taxonomy of methods for feature extraction is presented in Figure 2.4. The process of feature extraction is divided into three categories: filtering, variable selection, and classification. Each category has two distinct subcategories to further differentiate the methods. Filtering is a transformation of the data, with (binning) or without (convolution) reducing the number of variables. Selection is grouping parts of the data, with (peak-picking) or without (clustering) a transformation. Classification is finding the underlying components in the data, with (regression) or without (decomposition) a residual term.

One common goal of all methods is the reduction of noise to improve the quality of the extracted features. In most taxonomies, data or dimension reduction is often

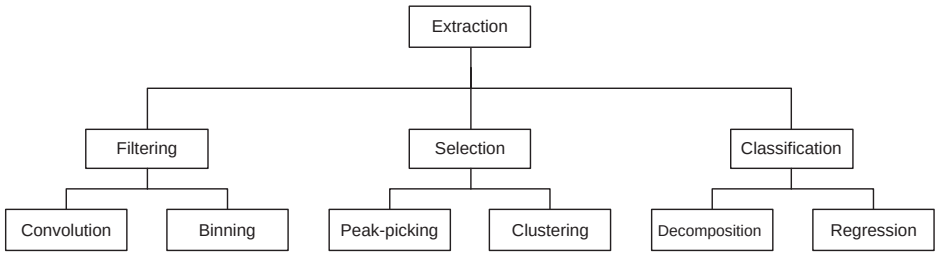


Figure 2.4: *Hierarchical taxonomy to classify methods for feature extraction.*

a separate category, but in this overview it is considered a property of the method. This choice was made, because different methods for data reduction can be placed within each separate category. The general properties of each method are explained in an overview in the following subsections along with a comparison of the major methods. A selection of appropriate methods is implemented for the approach taken in this study.

2.3.1 Filtering

Filtering is a common approach to reduce noise, to reduce the amount of data and/or to improve the signal-to-noise ratio. In the spectrometry literature, this step prior to selection and classification is also referred to as 'preprocessing'. It can be implemented by removing data-points or by transformation of the data, which both usually lead to a reduction of data. Many methods for filtering exist in the field of one-dimensional (1D) signal processing. Besides 1D-approaches, there are many two-dimensional (2D) methods for filtering in the field of image processing as well. Both 1D and 2D methods can be used in processing spectral datacubes, provided they have both a spectral signal and an image component. A combination of a 1D and 2D filter can be implemented as a 3D filter, which acts differently in the spectral dimension compared to both spatial dimensions.

It is considered necessary to improve the quality of data with respect to noise to be able to extract features more accurately. Also, the performance of feature extraction can be improved when noise is removed from a dataset and the remaining data-points have a higher signal-to-noise ratio. Improvement is only possible when the appropriate method is chosen. This choice should be based mainly on the type of spectral measurement, but the goal of the measurement and the goal of the analysis are important as well. Some aspects in the raw data which could be important for interpretation are sometimes removed or incorrectly transformed when a generic noise reduction filter is applied. For example, small peaks in a single mass spectrum are based on individual counts of ions. These peaks could easily be considered noise when they are unjustly present according to a selected filter model. With mass spectrometry data, each detected ion could be part of a significant peak when neighboring spectral channels and positions are considered. The same problem exists in the case of a spectral image. If the expected pattern in a spectral image can not be predicted, it is not possible to smooth pixels according to neighboring pixels. Mostly, there are no clear-cut borders or other recognizable spatial artifacts between different chemical

substances that can be used in a simple image filter.

Detailed models of the data are needed to be able to filter spectral data by removing or transforming data-points [Cly06; Pla06]. In most cases of mass spectrometry data, a spectral model is too complex and too large to be successfully fitted onto a resulting measurement. Only when a desired spectral profile is known after classification, multivariate regression techniques can be applied to filter the measured spectra. This is also known as 'calibration' of the data in the field of mass spectrometry. All filtering techniques that use regression need—estimations of—model information to be able to apply filtering. A remaining group of non-regression filtering methods can be divided into two categories that both implement a transformation of data: (de-)convolution and binning.

(De-)Convolution

Convolution filters transform data by replacing values with a weighted average on several data-points that are located near each other. The discrete convolution

$$\hat{f}[m] = (f \star g)[m] = \sum_{l=0}^{n-1} f[l] \cdot g[m-l] \tag{2.4}$$

defines the convolution operator $\star$ which takes two functions: the spectral function f of length $n \in \mathbb{N}$ and g of length $m \in \mathbb{N}$ where $n \leq m$ that produces the composite function $\hat{f}$. This composite function is a modified version of f and can be described as a weighted average of f. A *de*convolution filter does exactly the opposite: it is the inverse of a convolution filter. Its objective is to find f where g represents an estimated transfer function of an instrument. In deconvolution, an estimation of g would be made to obtain f from a measured signal $\hat{f}$ by reducing the instrumental noise. For instance, Ritter et al. [Rit04] implemented a deconvolution based on the known instrument response profile determined by a mass peak of silicon. However, in most mass spectral measurements, it is not always possible to get an appropriate estimation for g. One efficient group of convolution filters is that which uses wavelet transforms [Dro03]. These transforms enable to perform operations on images at multiple resolutions. To be able to apply a Discrete Wavelet Transform (DWT) to individual spectra one needs a model of the signal or wavelet function, a scaling function or sampling window, and a threshold on the resulting coefficients.

All values in a measured spectrum f are smoothed by g and thus can reduce the independent noise present in a measurement. After transformation, there should be less noise caused by the variability of a measurement in the data. Deconvolution can adjust differences in data-points locally and reduces influence of noise, provided a correct model for g is chosen. For instance, small spectral shifts in the location of peaks in a mass spectrum can be corrected if their model τ in Equation 2.2 is known. The estimation of a density function is less complex than creating a complete model of a dataset. Since the intrinsic shape of a peak in a raw spectrum changes with the mass, the adaptive properties of the wavelet transform are a good choice for filtering. Still, a wavelet function and a scale have to be chosen or estimated for the filtering process.

Besides using spatial windows, there are several ways to implement convolution and deconvolution filters. In signal and image processing, the classic Fourier trans-

form is commonly used for the implementation of smoothing by convolution filters. Vogt [Vog04] implemented 3D wavelet compression after which multivariate analysis is applied to the compressed dataset. Wolkenstein [Wol97; Wol99] also applied 3D wavelet filtering, but on a small number of spectral images to implement segmentation. Although stationary wavelet transform is a technique without a sampling window, a choice has to be made at which level to filter. According to Brown [Bro99], there are no gains in multivariate calibration performance by the application of convolution filters. He mentions three principal effects: reduction of magnitude of higher-frequency components, introduction of correlated noise, and reduction of any high-frequency components of the noise-free signal by the filter.

Binning

Another filtering approach is binning, also known as 'down-binning' or 'bucketing'. Binning can be compared with the creation of a histogram that maps a number of observations into a smaller number of discrete categories. Spectra are usually binned by taking the sum of a number of consecutive spectral variables. In binning, a reduced spectrum $\hat{f}$ is created by

$$\hat{f}[n] = \sum_{m=(n-1)\cdot w+1}^{n\cdot w+w} f[m] \tag{2.5}$$

where k is the width of a single bin (i. e., the number of variables), $m \in \mathbb{N}$ are the dependent spectral variables or observations in the spectrum f and $n \in \mathbb{N}$ are the new, binned variables ($n < m$) in $\hat{f}$. The value of w can be fixed or variable for each bin. Binning reduces both the size of a dataset by a factor w and the influence of noise by a simple transformation of the dependent variables. Binning is a predominant method in mass spectrometry to increase signal-to-noise ratio and reduce dimensionality in the spectral dimension. Neighboring spectral variables are grouped together by summing their intensity values into a single, new spectral variable. This way, the heights of the spectral peaks in a binned datacube are increased, while the resolution is decreased. Although mostly implemented in the spectral dimension, binning can also be applied to spectral image planes. It will combine a group of neighboring pixels into a single new pixel. Again, a resulting image has a lower spatial resolution, but the independent noise has less influence on the image.

The signal-to-noise ratio will increase by applying binning at the expense of the resolution, but without much computational effort. The bins can be of a fixed width or they can be of variable size, using manual inspection or automated algorithms. An example of binning with equal-width of $w = 2$ is shown in Figure 2.5, with in (a) the original variables m and in (b) the new variables n. Calculations are straightforward, as specific models do not have to be estimated when using bins with a fixed width. Mass spectrometry data has a relatively large spectral resolution compared to other spectrometry methods. Therefore, binning is ideal for data compression in the spectral dimension and simultaneously increases the signal-to-noise ratio. Similarly, spatial binning could be an interesting approach provided a spectral datacube is spatially extended by combining multiple spectral images into a single dataset.

The high resolution of imaging spectrometry data perfectly allows for binning. Different types of binning and peak selection can be used [Car03; Ran05a] before

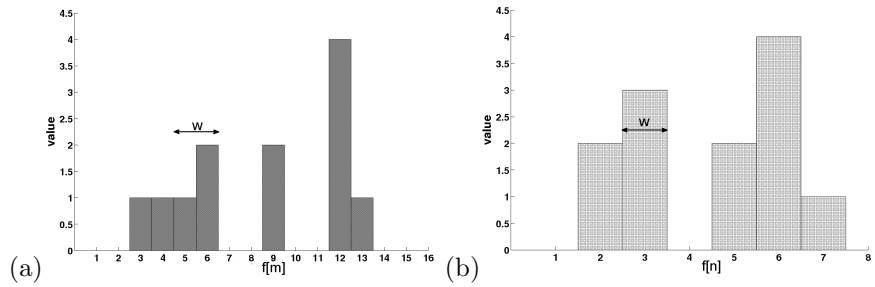


Figure 2.5: *Binning with equal-with bins with size $w = 2$ of (a) the original variables m into (b) the new variables n.*

applying multivariate analysis methods to analyze and quantify mass spectra (see Subsection 3.2.2 for more details). Wickes [Wic03] stated that the best image denoising algorithm is down-binning, instead of wavelet or boxcar filtering. When binning is applied ideally, each spectral peak is put into a single bin [Dav07]. A practically impossible task, because the distances between these peaks vary. Another obvious disadvantage of binning is that separated, neighboring peaks are combined into a single peak. Therefore, information is lost in the filtering process.

2.3.2 Selection

The previous subsection elaborated on filtering methods for noise reduction and data compression. The data selection methods in this subsection try to reach the same goals by the selection and grouping of similar data-points. The main difference is that filtering can be applied prior to selection, to improve the results of a method for feature selection. For example, instead of filtering with a fixed bin-size, a peak-selection approach locates peaks after which each peak can be used as a new variable. Inherently, this is the most optimal way to bin a spectral dataset, provided there is a robust and accurate selection of peaks. However, rather than solving the problem of finding an appropriate bin-size, it become a problem of peak-selection. Besides specialized methods designed to distinguish and select spectral peaks, there is a variety of generic approaches to reduce data. These methods select different parts of a dataset based on—statistical—properties they have in common.

Data reduction methods can be considered methods for selection. The influence of noise or unimportant artifacts can be reduced when they are not included in a selection. If noise has a different statistical model of distribution, variables can be singled out or projected into a new selection of variables, called 'factor'. Therefore, approaches for variable selection are mostly referred to as common Factor Analysis (FA). The description of the methods for FA is ambiguous. Besides data reduction by selection, these methods can also be categorized as a filtering approach for reduction of noise. This group of methods is even able to classify a dataset when the number of classes is known. This classification approach is described in more detail in Subsection 2.3.3 about classification in feature extraction.

Numerous different implementations exist for FA, with different statistical assumptions, parameters, target distributions, and performance issues. Examples of

more complex variations on FA are commonly used for classification problems and can be found in Subsection 2.3.3. Two of the most commonly applied variations of FA are Principal Component Analysis [Hil06] (PCA) and correspondence analysis. A main difference with other FA derivatives is that PCA does not need a fixed and predefined number of factors or components to decompose the data. Also, most methods for FA assume that multiple models are present in the data [Här03] of which the common variance in each factor is considered. It depends on the method of computation and the model which variables are combined in one factor. PCA merely distinguishes the covariance of the total variance present in the data and the results are independent of the method of computation. Correspondence analysis distinguishes itself by the assumption that data must be positive and distributed according to Chi-square distribution. Most of the remaining methods for selection of spectral data can be divided into two categories: peak-selection and clustering methods.

Peak-picking

A general discrete peak-fitting function $f'[s,t]$ can be defined as

$$f'[s,t] = \sum_{m=1}^{M} f[m]\psi_{st}[m] \tag{2.6}$$

with $f[m]$ the spectral signal and $\psi_{s,t}[m]$ as a family of discrete wavelet basis functions, s as the scale, and t as the translation factors. High values of $f'[s,t]$ indicate peaks of scale s at position t. These peaks are selected by a threshold on the values. The occurrence of spectral peaks is a common characteristic in the field of spectral analysis. One or more peak(s) indicate the presence of a specific chemical compound. Besides the presence of noise, there are two characteristics that complicate peak-selection or peak-picking. A peak is mostly distributed among several, neighboring spectral bands, depending on the resolution and the type of measurement. A single apparent peak can consist of several peaks that are located closely together in a spectrum. Both characteristics complicate an automatic selection of peaks. One way to solve this is to perform interactive peak-selection, in which an expert spectrometrist recognizes and manually selects a peak in a spectrum. Additionally, a spectral library with known peak locations can be used to make a correct selection. A spectrometrist still has to make the final interpretation, because libraries can not contain all chemical compounds with their corresponding spectral peaks.

Proper peak-selection ensures a spectral dataset can be reduced in size without losing important details. If a specific peak is correctly detected and selected, the intensities are combined or transformed into one or more new spectral variable(s). This way, the influence of noise is reduced together with the number of excess spectral variables. In a mass spectrum, small fluctuations of noise can not be filtered by applying a threshold on signal-intensities. The signal-to-noise ratio is too low in this type of spectral data, to be able to distinguish between a spectral peak and noise.

There are many implementations [Coo07; Ran05a] of peak-detection algorithms that use database entries, model information, or parameterized differences between sequential spectral bands. For instance, Morris [Mor05] implements peak detection with the discrete wavelet transform applied to the mean spectrum instead of a single spectrum. Most implementations of peak-selection approaches do not perform well

with individual spectra, because their signal-to-noise ratio is too low to distinguish a peak from surrounding noise. Therefore, peak-selection methods should be applied to a combined rather than a single spectrum. Alternatively, a spectrum could be filtered by, for instance, a convolution approach. According to Chen [Che03], self-modeling curve resolution can be seen as a sub-category of FA. It is a way to determine spectral profiles without prior information about the location of peaks or reference data. This kind of curve resolution does need an estimation of pure profiles and noise to make the estimation.

Clustering

In clustering, a dataset $F = \{f_1, f_2, \ldots, f_n\}$ of $n = x \times y$ spatial entities is mapped on a smaller set of K clusters $f'_K = \{c_1, c_2, \ldots, c_K\}$ where $K < n$. The membership of a spectrum f_n to a cluster c is defined by minimizing an objective function

$$O(F, f') = \sum_{k=1}^{K} \sum_{i=1}^{n} d(f_i, c_k) \tag{2.7}$$

where $d(f, c)$ is a distance function between the entity f_n and a prototype entity c (i. e., the average of all points in the cluster) of a cluster. In general, clustering of data covers a wide range of methods. The idea behind Cluster Analysis (CA) is to partition a dataset into several subsets, in such a way that each subset has similar properties. Similarities are expressed by a distance measure. A distance measure can be used, for instance, to find similar spectra located in spatial neighborhood. Techniques for clustering can be divided into three categories: partitional (distance-based, model-based), hierarchical (linkage, model-based) and density-based (mode-seeking, graph-based). Most of them require a distance function and spatial information if available.

Cluster analysis is applied to group areas with similar spectra to be able to recognize spatial patterns with the same chemical composition. It is similar to a peak-selection approach, but adds location information to the distance measure. Thereby, cluster analysis can be used for image segmentation including the spectral information. Tran [Tra05] investigated clustering methods applied to multi-spectral data and defined clustering as: "to help to understand relationships of objects by similarity". When visualized, subsets can instantly provide insight into these relationships. Clusters can be left out of further analysis or focused on in more detail by applying another cluster analysis on a resulting subset [Fle06].

There are many implementations of clustering methods from the field of digital imaging. These implementations are based mainly on the spatial characteristics and intensity values in a single image. Clustering results depend heavily on the defined measure for similarity. Again, a model of the data or similarity measure has to be defined and the expected number of clusters have to be given in non-hierarchical clustering [Kes03]. Linear Discriminant Analysis (LDA) assumes Gaussian conditional density models and only needs the desired number of clusters as input. Derivatives on cluster analysis exist (e. g., Support Vector Machines), that do not need an additional model. Instead, the parameters in the distance measure are estimated by training with a correctly partitioned dataset. The availability of partitioned datasets complicates the applicability of this type of data selection in spectral datasets. Meth-

ods in feature extraction that incorporate more sophisticated models for clustering can be categorized as classification methods.

2.3.3 Classification

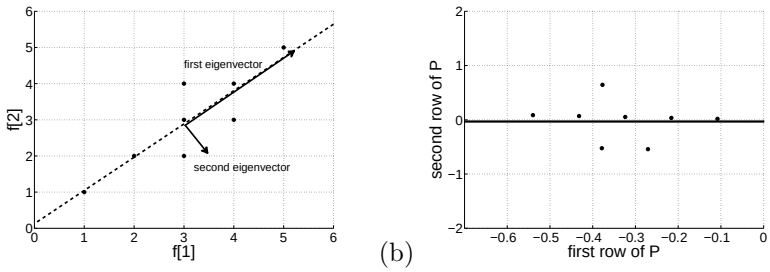
The process of classification is similar to certain methods for peak-picking and clustering. In general, classification is the labeling of individual objects based on quantitative information on one or more properties of these objects. In imaging spectrometry, classification is the actual labeling of data-points that belong to the same spectral group and/or spatial region. Labeling is done either by a defined similarity measure, by expert knowledge, or by applying statistics. Similar to applying filtering methods before selection, a primary selection can be made prior to classification. Making an appropriate pre-selection is beneficial to the performance of a classification method. Firstly, a pre-selection reduces the search-space for a classification. Secondly, influences of noise and artifacts can be removed which should make the classification more robust.

Generally, classification is applied to reduce the influence of noise by analyzing spectral data and subdividing a datacube in recognizable pieces with the same characteristics. Anderson [And00] describes mathematical models for exploratory multivariate data analysis. Classification makes it possible to first explore correlations supporting an analyst to generate hypotheses. There are many non-linear approaches and variations of decomposition and regression techniques. This overview does not focus on the use of non-linear classification techniques for dimension reduction [Maa07]. According to Nascimento [Nas06], spectral data can not be correctly unmixed and classified by for instance Independent Component Analysis (ICA) and Independent Factor Analysis (IFA). Both expect statistical independence of non-Gaussian distributed data instead of correlations in Gaussian distributed data as is the case with mass spectral data. Furthermore, both can be implemented as higher order and/or non-linear methods.

The process of classification can be implemented by matrix decomposition or by fitting estimated factors in regression analysis. In other words, classification is an attempt to find X_{mn} and β_m where the concentration vector β_m can have different values on different spatial coordinates in $F[x, y, m]$. Both multivariate methods can be preceded by several filtering methods [Lee08]. Regression analysis requires training, calibration datasets, or proper statistical estimators. In methods for classification by decomposition, there is also an implicit model of the data. Regression analysis and methods for classification by decomposition are closely related to each other. The use of Neural Networks (NN) for classification is another way to examine linear as well as non-linear relations in the data [Chi01; San02]. The NN tries to learn how to classify a dataset, by giving itself feedback with a so-called cost function [Hut96]. For instance, a posterior probability function can be used to estimate the statistical model in a dataset. Performance and design issues prevent fruitful implementations of NN for the exploratory analysis of spectral data.

Decomposition

The goal of decomposition is to break up a dataset in a number of smaller components to gain more insight how these smaller parts contribute to the whole dataset. A



(a) (b)

Figure 2.6: *Two spectral variables $f(1)$ and $f(2)$ that are positively correlated with (a) the two eigenvectors and (b) the data projected on the new dimension of the first component.*

decomposition is usually expressed in one part with the correlations between the variables and the smaller components, and one part with the projections of the variables onto the smaller components. A general model for a matrix decomposition is

$$D = P^T \cdot Y \tag{2.8}$$

where D is a 2D m by n matrix with observable variables and instances (also 'samples') of those variables, Y is a n by n ('scores') matrix with the projected variables, and P is a n by m transformation (also 'loadings') matrix that contains the correlation between the spectral variables m and the components n. One way to find suitable matrices for Y and P is by PCA using eigenvalue decomposition (see Subsection 3.3.1 for more details). The values in these components can be both negative as well as positive, which shows whether the variables are positively or negatively correlated. The number of components is limited only by the smallest dimension of D. Usually, decomposition methods are integrated in regression algorithms to estimate a component with certain constraints. For example, two positively correlated spectral variables $f[1]$ and $f[2]$ are as shown in Figure 2.6. When these variables are collinear, they can be described by one new variable by rotation of the ordinant axes using the eigenvectors. This variable is projected onto a new axis by the eigenvector in P with the largest eigenvalue as shown in Figure 2.6b. The second eigenvector with the second largest eigenvalue describes the residual variance if the variables $f[1]$ and $f[2]$ are not perfectly collinear. In this way, all variables are decomposed according to their variance. The number of variables is reduced if there are any linear relationships between them. The noise in the data (described by the second eigenvector in this example), can be removed by using only the first principal component (first eigenvector) of the projected data.

Features can be extracted from mass spectral data by decomposing the datacube into several components. The matrix P contains spectral profiles, which resemble different rows in X_{mn}. The rows in Y contain the different values for β_m on each spatial location in the datacube. A decomposition method does not use estimates, specific model information, or expert knowledge. This group of methods has to rely on the relations within the given data-points to create a decomposition and a potential classification. Unsupervised classification is applied, when datasets are too large to be able to manually create a training set. Also, relations could be too complex to be put into a complete model in advance. A third reason to use plain decomposition is when expert knowledge about the relations in the data or its distribution is not

available. A disadvantage is that the solutions are never exactly modeled due to noise in the data. Another disadvantage is that it is hard to verify the results, because it is unclear what the classification is based on.

In this category of classification methods, implicit constraints are used only. For instance, an orthogonality constraint on P decorrelates the different components and merges multicollinear peaks. Closely related is the Singular Value Decomposition (SVD). Both use second order statistics to re-express data in terms of basis functions. These methods do not incorporate a model of the noise on top of the chemical model of the data. Implicitly, this additional noise factor can be found in the components that explain the least of the—Gaussian—variance in the data. Contrastingly, methods belonging to the group of FA estimate the chemical model and leave the noise out of the final decomposition in factors. Another widely used method is Multiple Correspondence Analysis [Kie91] (MCA), sometimes referred to as 'weighted PCA' because both components are weighted using a statistical model, for instance χ^2.

Factor regression

Another method used to determine original spectral profiles X_m in a measured dataset is factor regression. This method uses a limited number of components (or factors) and has a separate term for the residual. A loading matrix L is estimated by the spectral profiles that are expected to be present in a dataset or by putting constraints on these variables. Also previous classification results from, for instance PCA, can be used as an initial estimation of L. These values are used to model a dataset in factors using regression. The factor model used in regression is generally of the form

$$D = L \cdot V + U \tag{2.9}$$

where D is a m by n matrix of observable variables and instances. The matrix L is a m by k matrix of factor loadings or unobservable constants, V is a k by n matrix of unobservable random variables and U is a m by n matrix with a unique, uncorrelated or error variance for each factor. This model is called the 'factor model' and is related to the decomposition model with a couple of differences. The number of factors explained by this model is k and is predefined as a constraint on the model. This model tries to factorize a dataset in a smaller number of components compared to the decomposition model. The coefficients in X_{mn} are estimated by the k columns in L. This model is also applied to the Multivariate Curve Resolution (MCR) [Tyl06]. A supervised classification method by regression is trained by examples outside the dataset, by a data-model, or by expert knowledge. These training or calibration datasets have to be supplied to be able to create a classifier for the data, for instance maximum likelihood. In some cases, when only a few distinct properties in a dataset are known, the use of a data-model could be advantageous. Constraints and limitations on a statistical distribution can provide a practical model of the data for classification. Using pure spectra, known mixtures, clustering and region selection, an expert has to determine endmembers. These endmembers are used for correct classification by applying a regression model on the measured spectra.

The use of previous classification results for L will enable an accurate selection of those specific properties in a dataset for a correct classification. Classification results include those spectral profiles that contain several spectral peaks which are already

identified. A linear function can be fitted through the data-points if there is only one independent (spectral) variable. There are many spectral variables in a spectral dataset. Therefore, a linear regression model has to be found that fits the linear combination of these spectral variables and does not have to be a linear function of these variables. Classification will fail, when a non-representative training set is used in a regression or when a wrong model of the data or noise is implemented. Linear dependencies between independent variables (multicollinearity) make linear regression impossible unless constraints are put on the estimators in the model. An expert should perform a final interpretation and verification of the results before invalid assumptions lead to the wrong conclusions. Previous results of correct classifications can be stored in a database and used in future supervised classification.

Many examples of regression schemes exist in literature. Most of them are designed for a specific application for classification. When Multivariate Linear Regression (MLR) is performed on the resulting scores of a PCA, it is called Principal Component Regression (PCR). It is used to make an estimation of the different spectral profiles present in a sample. Preacher [Pre02] explains exploratory factor analysis as a more general approach to regression analysis. All factor analysis methods have in common that they need a model of the data and noise. With this, an estimation can be made which can iteratively be improved by for instance (Partial) Least Squares (PLS) regression or one of its many variants [Wag04]. Projection Pursuit (PP) is similar to FA to discover groups of data and outliers in projection subspaces. Constraints are put on spectral data components with non-negative matrix factorization from Pauca [Pau06]. Parallel Factor Analysis (PARAFAC) by Bro [Bro97] provides a unique solution, in contrast with other methods for FA. This means that a solution has no rotational freedom and can provide a more robust chemical model of the data.

2.3.4 Comparison

In order to implement the desired feature extraction approach, a selection has to be made in the aforementioned methods for noise reduction. Three categories (filtering, selection and classification) for feature extraction were distinguished, but some methods can be placed in alternative categories. One reason is the large overlap in functionality of the methods. Many of the mentioned methods are variations on the same theme. There is a large variety of alternative methods or specific implementations, but this overview tries to distinguish the main approaches and some of the differences in this large collection of methods. As mentioned in the previous subsections, most methods can be used side-by-side in the same analysis: filtering to improve a primary selection and selection to improve the classification. However, their common goal is the extraction of features and reduction of the influences of noise. This subsection provides an overview of the methods and their properties.

All of the methods mentioned are designed to reduce the influence of noise in an analysis. Since each of them has a different approach to reduce noise, different advantages and disadvantages have to be considered in a specific application. This comparison is not intended to formally quantify all available methods. Instead, it aims to provide a framework on which the decision of finding the appropriate methods can be based. Certain criteria have to be defined to be able to find an appropriate subset of methods that can be used for the exploratory qualitative analysis of (imaged) mass

method\property	selection	reduction	parameters
down-binning	−	+	window
convolution filter	−	−	transfer function
deconvolution filter	−	−	transfer function+model
wavelet filtering	−	+	wavelet function+window
cluster analysis	+	+	#components
principal component analysis	+	+	none
factor analysis	+	+	#components+model
multiple correspondence analysis	+	−	model

Table 2.3: *Different methods with noise reduction properties for peak detection.*

spectral data. A main selection of methods with their accompanying scores on certain criteria can be found in Table 2.3.

In this approach, automatic selection of spectral variables is important. Consequently, an expert can inspect and interpret a selection. Not only is a manual selection within large datasets a time-consuming exercise, but chances are that interesting data is missing in a selection. This brings us to the next property that is important in the analysis of mass spectral data: the ability to reduce data. In large datasets, it is hard to get a complete view of all of the different relationships. Many methods for feature extraction tend to reduce the search-space by dimension reduction or compression. Instead of dimension reduction by subset selection (or filtering), the number of dimensions can also be reduced by transformation into or projection onto a new and smaller set of variables. There is a difference between dimension reduction and—lossy or lossless—compression [Kaa01], although both have the goal to retain as much of the signal as needed prior to classification. Dimension reduction intends to improve classification by eliminating irrelevant variables in contrast with compression which eliminates redundancy in the data as much as possible.

A next criterion can be put on the parameters and—as a result—the complexity of each method. Since each spectral measurement and its intended goal is different, it is desirable to have as few parameters as possible. Of course, without any model parameters, feature extraction becomes less accurate compared to having a full chemical model and estimation of noise. Finding the optimal parameter setting for an experiment is a difficult task. Not many mass spectrometrists are knowledgeable in the subtle differences of each setting. Experts still have to interpret and quantify the results in exploratory analysis. A quick and simple feature extraction could be more efficient than time spent on creating an exact parameterized model. In this comparison, methods become less complex when they have less parameters as described in a survey of nonlinear dimension reduction techniques by Van der Maaten [Maa07]. This study compared nonlinear to linear methods and found that nonlinear methods do not perform better, because of the curse of dimensionality, overfitting of local models, oversampling, and the sensitivity to outliers. Iterative methods are likely to be more complex compared to direct methods as the final point of convergence is not fixed, but depends on the used parameters. Quantification of computational complexity in extraction methods is an intricate area. This complexity depends on the dataset, on specific parameters, and is difficult to express in comparable figures [Pla04]. An

exhaustive and more formal comparison is beyond the scope of this overview.

In our approach, the methods with the least prerequisites are chosen to be implemented in the set of tools. The many available spectral variables make it possible to choose those methods that are able to make a trade-off between resolution and improvement of the signal-to-noise ratio. The primary goal of feature extraction in this approach is to be able to make a selection of possibly interesting features. A final classification and interpretation is left to the expert to whom the results are presented. Since the data is multicollinear, the most simple approach for selection is PCA. It enables the selection of data and is well-known for its data reduction capabilities. Another important property is that it is not necessary to set any parameters. Unfortunately, in order to be able to apply PCA to a collection of mass spectra, the signal-to-noise ratio has to be improved prior to selection, for instance by filtering. Otherwise, the variance in the noise dominates extracted components, because PCA uses covariances as a criterion to make a distinction between different components. The most simple method to reduce and filter spectral data is binning. The purpose of binning is not only to increase the signal-to-noise ratio of the spectral peaks, but primarily to reduce the number of variables to be able to apply PCA. One mass spectral peak is measured among several spectral variables. A higher signal-to-noise ratio is established by combining neighboring spectral variables instead of smoothing them. Together, these methods create a minimal subset of tools to allow for the extraction of features in imaging mass spectrometry data. Others have successfully used PCA as well as binning for feature detection. The difference with this work is the explicit use of binning in feature detection and separation of positive and negative (correlated) parts in a component. Interesting peaks are often selected manually [Lho01] before PCA instead of being able to use the full dataset for exploration. Additionally, feature detection is separated from the visualization of features, which allows for enhancements in the representation of features.

2.4 Visualization

After extraction, features are visually presented to an expert. This expert has to be enabled to make an appropriate selection in the spectral and/or spatial dimension for a more detailed inspection and interpretation of the automatically highlighted features from the data. Besides the interpretation of the spectral peaks, an expert focuses on recognizing spatial patterns in spectral data. Visual cues on extracted features will provide better contrast against other parts of the dataset. The ability to leave out identified artifacts and noise by selection creates a better focus on the more interesting parts of the dataset.

According to the dimensionality, visualization techniques can be divided into four categories: spectrum, image plane, datacube, and feature. Müller exploited PCA to optimize each separate step in the generic visualization pipeline [Mül06]. Landgrepe [Lan00] distinguished visualizations into three categories: the spectral space, the image space, and the feature space. One or more spectra can be visualized in a plot of the signal in the spectral space. In the image space, the intensity values are visualized by an image plane on a specific spectral band. In the feature space, the results are parameterized and could be represented by an abstract shape or icon. Another category for visualization is added in the approach in this work. This is

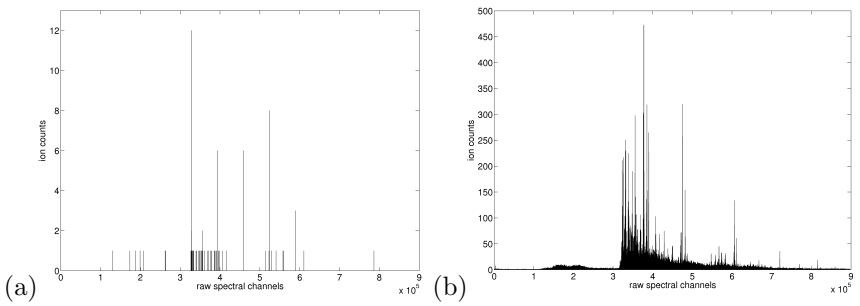


Figure 2.7: *(a) An example of a single raw spectrum with raw channels and (b) the sum of all spectra with raw channels of neuroblastoma cells.*

the visual representation of the datacube in three dimensions. This representation is closely related to the visualization of features, but can be distinguished by the fact that the data-points are not parameterized before display. The following subsections provide a more detailed overview of each of these categories.

Spectrum

Line-plots are 1D representations of a spectral signal. They are a representation of one or more spectra in a single view, as shown in Figure 2.7a and b. Multiple spectra measured at different locations can be visualized by multiple lines, although another common approach is to display the sum (or average) of all spectra. A mass spectrum shows the isotopic distribution of a sample material. Each data-point in a spectral datacube represents a detected ion. These data-points are joined together by a line. Several detections, located closely together in the spectral dimension, create a spectral peak. In other words, peaks in mass spectra are those regions with a relatively higher density than their surrounding spectral regions.

The display of a spectrum in a line-plot is the traditional view in the analysis of spectral data. This way, peaks in the spectral signal are instantly visible. In the case of a mass spectrum, multiple spectra are added together to improve the signal-to-noise ratio for better distinction of peaks. This is an operation on multiple spectra, and therefore removes the spatial distinction between spectra and decreases the resolution of a spectral image plane. An alternative method to improve distinction between peaks is to filter a single mass spectrum with for instance a binning or a convolution method. This improves the signal-to-noise ratio, but with a decrease in spectral resolution. Again, there is a trade-off between noise that could occlude a peak and the level of detail of a peak's visualization.

A spectrum is plotted with the spectral units on the x-axis and the intensities on the y-axis. The spectral units are usually expressed in the mass-to-charge ratio instead of 'channels' that are measured by the mass spectrometer. Historically, a spectrum is analyzed by locating the peaks in a spectrum. Therefore, alternative approaches for feature extraction should keep the possibility to view the results spectrally.

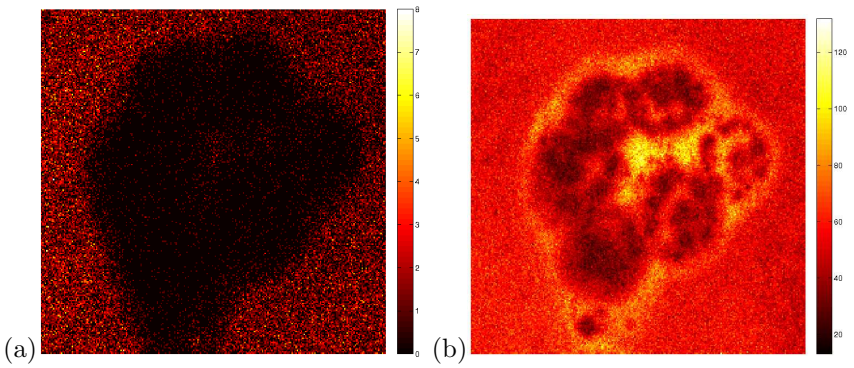


Figure 2.8: *(a) An example of the spatial distribution of intensity values of a group of* 250 *summed channels combined in one spectral band and (b) the spatial distribution of the sum of all spectral bands of neuroblastoma cells in a* $150 \times 150\,\mu m$ *field of view.*

Image plane

A spectral image plane is a 2D representation of the intensities on one or more spectral band(s) as shown in Figure 2.8a and b. Usually, multiple spectral bands are combined and visualized in the same image. When all spectral bands are summed and displayed in the same image, it is usually referred to as the Total Ion Count (TIC) image. Different colors are assigned to different signal intensities. This creates a spatial distribution of measured signals. Just as all positional information is lost in the spectral view, all spectral information is lost in a spatial view. Many small biological structures can be recognized in an image. A spectral image is therefore a useful representation to identify and interpret different chemical structures in an extracted feature.

An image representation instantly shows the spatial patterns in a measurement, but knowledge about the chemical composition is still needed to find an appropriate spectral region. Not all spectral regions contain distinguishable spatial patterns or have enough contrast to be used for interpretation. Most images resulting from mass spectrometry do not contain edges, corners, or blobs with enough contrast in order to be detected automatically. Results have to inspected manually for interpretation.

Each intensity value on a specific spectral band is converted into a pixel with a specific color in a color-scale. The observation made by Prutton [Pru99] states that some mass spectra can contain information about the surface depth as well as the masses of the detected ions. This information can be visualized by creating an image plane of a single peak. Each—relatively small—difference in mass can be visualized by assigning colors to different mass-to-charge ratio located nearby [McD03]. This creates a map where different regions in mass-to-charge ratio are assigned to different colors. As a result, this image contains a height-map of the surface of the sample after a measurement with an appropriate technique.

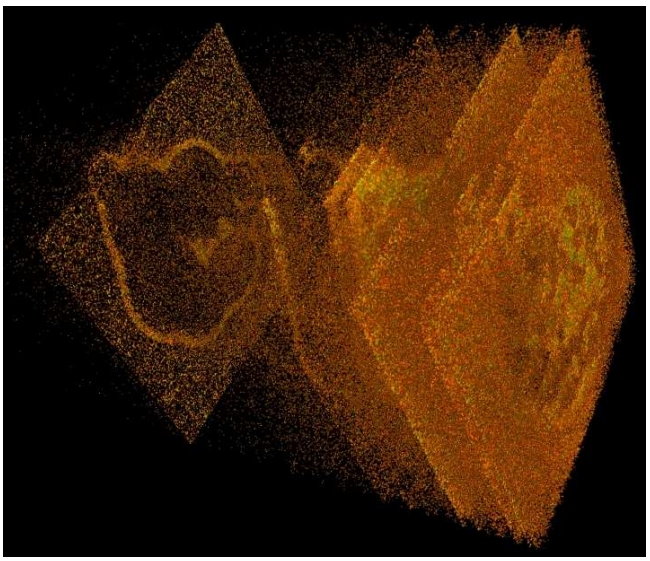


Figure 2.9: *An example of a 3D representation of a spectral datacube of neuroblastoma cells in which the lowest intensity values are transparent.*

Datacube

A datacube is a volumetric representation of the acquired spectra with their spatial positions. This three-dimensional (3D) representation is basically a sequence of image planes, stacked on top of each other as shown in Figure 2.9. Whereas spatial information is lost in the spectral view and spectral information is lost in a spatial view, a combined view could be useful to highlight these relations.

An eminent problem in this representation is the occlusion of the data-points within the boundaries of the datacube. Problems with occlusion are usually solved with tools that interactively change the view in a 3D visualization. Changing the view on a solid datacube will not solve the occlusion problem. The introduction of selection tools can help to interactively remove uninteresting parts of the datacube. Additionally, transparency can be used to provide more insight in occluded parts. Since a datacube has no 'real' physical meaning, interpretation of the resulting shapes can be troublesome. Therefore, the applicability of a 3D representation of a datacube for analysis remains questionable.

Some authors experimented with 3D visualization of imaging mass spectrometry data [Sme07]. In 1997, Kenny [Ken97] experimented with compression, volume rendering and segmentation of datacubes from multi-spectral analytical electron microscopy. Although his implementation was limited by hardware, no further progress was made in this area even when hardware improved. Other experiments with multi-resolution visualization strategies on sparse data from neutron spectroscopy were made by Bustinduy [Bus05]. Although the same problems were recognized, this implementation was able to visualize some spectral variables with a limited spatial resolution. Small differences in the spectral dimension can be visualized as 3D differences in height to represent surface characteristics.

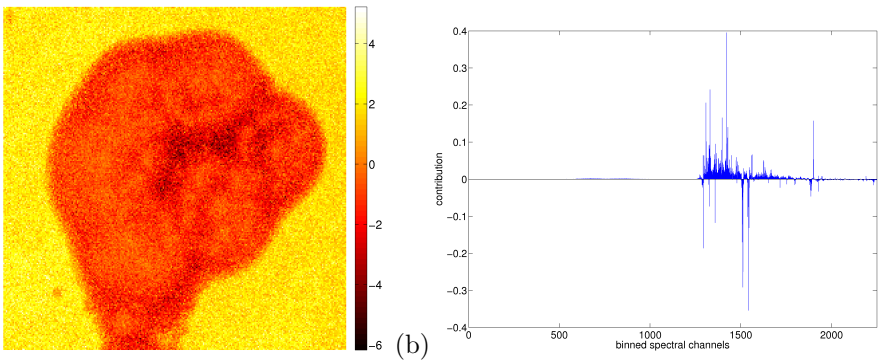


Figure 2.10: *(a) An example of the spatial component of a projected feature and (b) the spectral component of the same feature, from a datacube of neuroblastoma cells in which groups of* 250 *channels are summed together.*

Feature

Instead of visualizing the original data values, an abstract representation can be made to highlight properties of—parts of—the data. Basically, all abstract, filtered or projected representations of spectral data can be considered visualizations in a newly defined feature space. Many results of the feature extraction techniques can be put into this category. Most of these methods create a subset of the original dataset or make projections onto new variables as shown in Figure 2.10a and b.

Feature visualization can be helpful, because the original intensity values do not always provide the necessary insight into the data. As the data is transformed and the dimensions change, however, it is difficult to extract any quantified information from a dataset. The metrics and models in the feature extraction have to be very accurate to be able to interpret a visualization of features instead of using the original data values. In the field of visualization, one common technique for dimension reduction is Multi-Dimensional Scaling (MDS). This visualization technique allows the exploration of similarities and dissimilarities within a dataset.

In an approach for exploratory analysis, it is not harmful to rely on feature visualization. A final quantification and interpretation has to be done with original data-values, preferably on the highest possible resolution. PCA is able to provide both spectral and spatial components as a result of an analysis. Both components are projections from the original dataset and therefore can be called features. Meanwhile, these features can still be represented in a spectral or spatial view. The main difference with real values is that both spectral and spatial components contain negative data values, whereas the real intensity values do not. These negative values are inherent to the way PCA decomposes the data. These visualizations can therefore be useful for feature discovery, but interpretation has to be based on the real data values.

2.5 Summary and conclusion

This chapter presented an overview of the methods used for the analysis of imaging spectrometry data. The analysis in the proposed approach is divided into three stages: data acquisition, feature extraction and feature visualization. A detailed description of currently applied methods is given for each stage. This was done to obtain a specific subset of methods that is the most appropriate in this qualitative approach for analysis. The choice was made according to the defined intentions and goals for analysis, applicable in the extraction and visualization of features from acquired spectral image data.

The high-resolution capabilities and available spatial information make imaging mass spectrometry ideal for the analysis of biological samples. PCA in combination with a binning function is most suited for extracting features from imaging mass spectral data. Both methods increase the signal-to-noise ratio and reduce the amount of data. The binning function can create different levels of detail in visualizations. Additionally, PCA is able to distinguish between positive and negatively correlated spectral profiles. The visualization of (spectrally and spatially linked) features supports an analyst in the exploration of a spectral datacube.

In the remaining chapters of this thesis, different issues are addressed to facilitate feature extraction and visualization in order to support exploratory analysis. The main subset of methods used to implement these applications are: binning, PCA, and (primarily 3D) visualization to support the analysis and exploration of spectral data. The combination of these methods should make it possible to implement basic feature extraction and visual exploration of the results by an expert.

Chapter 3

PCA-based feature extraction

Chapter 2 provided an overview of different methods for the acquisition of spectral imaging data, extraction of features and visualization of the results. In our approach, two simple methods were chosen for qualitative analysis. These methods provide filtering, selection and classification of imaging spectrometry data. With these two methods, different tools were developed to assist an expert in the exploratory analysis of large spectral datacubes.

Chapter 3 will provide more details on implementing feature extraction on spectral datacubes by using PCA or PCA-based methods. Firstly, a spectral datacube has to be transformed and binned. Then, features are extracted from the datacube in three ways: by applying PCA and two PCA-based methods. For all three methods, the resulting features are compared quantitatively and qualitatively. The extracted features are used to solve issues with the visualization of mass spectral imaging data, as described in the remaining chapters of this thesis.

3.1 Goal

In our approach, the goal of extracting features from a spectral datacube is to automatically create a potentially interesting selection for analysis. Interesting features are data-points which have similar properties, thereby distinguishing themselves from other parts in a dataset. In the case of spectral data, these features are the spectral peaks that recur in several spectra or on different locations. Basically, the goal is to discover the original spectral peaks X_{mn} and their concentrations β_n in

$$F'_{x,y}(m) = \beta_1 X_{m1} + \beta_2 X_{m2} + \cdots + \beta_N X_{mN} \tag{3.1}$$

that produced a spectrum $F'_{x,y}(m)$ on a certain location (x, y) where n is the number of original spectral profiles and M the number of spectral variables. PCA-based methods are able to automatically create selections of features according to the variance present in the spectral variables without data-dependent prerequisite parameters. These unknown values for X_m can be approximated with a system of linear equations if there are enough measured instances $F'_{x,y}(m)$, i. e. an overdetermined system. These

selections are highlighted in a visual representation of the data, to be interpreted by an expert.

If each spectral peak has to be selected, interpreted and compared manually, chances are certain features are not noticed. For instance, if the goal of an analysis focuses on a different part or aspect in the dataset. Automatic feature extraction on the other hand will—besides being more sensitive—also reduce the manual effort in selecting individual or regions of data-points. Therefore, a method for automatic extraction is less time-consuming, but still relies on an accurate manual interpretation of the results.

The raw data of an imaging mass spectrometry experiment has to be prepared before a decomposition method can be applied. This phase is usually referred to as preprocessing of data. Several transformations have to take place before a decomposition method can be applied to the data. Different PCA-based methods can be applied after preprocessing. The mathematical models and prerequisites of three decomposition methods with different constraints are compared. These three methods are quantitatively and qualitatively compared to decide which is the most appropriate for this exploratory approach. A quantitative comparison is made by comparing the results of the methods with an a-priori known spectral data cube; i. e. a ground truth. A qualitative comparison of resulting components of a case-study is made by an expert.

3.2 Data preprocessing

Data preprocessing is the transformation of a dataset, which is necessary to be able to apply specific methods for analysis. The transformations for preprocessing can be divided into four different groups. First, the imaging MS measurements have to be transformed and combined into one dataset with a certain format, a spectral datacube in this case. Second, the spectral datacube has to be filtered to reduce it to an appropriate size and improve the signal-to-noise ratio. Third, this datacube is transformed into a 2D matrix for the application of the decomposition algorithm. An optional fourth transformation is to weight the different variables or instances against each other. In order to be able to estimate the weights, additional knowledge of the measurement is needed. It should be mentioned that the use of weighting as a means of transformation is sometimes called into question when the variance estimates of the data elements have a high degree of uncertainty [Kee05].

3.2.1 Format

PCA is traditionally performed on a 2D matrix with the samples of the dataset in one dimension and the dependent variables in the other dimension. A spectral datacube has to be converted to such a matrix with preservation of the spectra and images. In order to do so, the raw mass spectrometry data has to be placed in this spectral datacube format. A mass spectral measurement consists of a collection of ion detections. Each detection is a 2-tuple with a number representing the coordinates of a location. The two coordinates of one location can be derived from this number. The other part of the 2-tuple is a channel number, from which a mass-to-charge ratio can be determined. All these 2-tuples are combined into one datacube. One measurement

can contain more than $2.5 \cdot 10^7$ 2-tuples. This process transforms all measurements into a discrete datacube $F[x, y, m]$.

One way to reduce data is to use the raw channel data instead of converting its corresponding mass-scale. Accuracy is lost when converting the 32-bit channel integers into floating point numbers for the m/z values. Another advantage of using the channel integers is that the sparse matrix format can be used in MatLab^{TM}. This compression uses a Harwell-Boeing format which leaves out the storage of zero-counts in the mass spectrometry data without loss of information. The spectrum on a certain spatial location represents one column in matrix X. A slice of the datacube at a particular wavelength represents one row.

3.2.2 Filtering

The data has to be filtered to reduce the size and improve the signal-to-noise ratio. In this approach the minimal filtering is done by binning, also known as down-binning, bucketing, or bagging. This filtering approach is closely related to the discretization of continuous measurements or the mapping of measurements to categories by means of a histogram.

Binning can be applied in the spectral dimension as well as in the spatial dimension. Usually, there are many more spectral variables compared to the spatial variables in mass spectrometry. Because of their large amount of data-points, spectral datacubes are commonly reduced by spectral binning in the field of mass spectrometry. With this technique the spectral and spatial dimension are reduced to any desirable size by putting two or more consecutive channels or four or more neighboring counts into one combined bin. No ion-counts are lost and the signal-to-noise ratio is reduced at the same time. PCA needs this spectral reduction to be able to treat each peak as one variable instead of having more samples across multiple channels. Without spectral reduction, PCA finds correlations in peak distributions instead of correlations between spectral peaks.

A simple linear binning function $\hat{f}[i]$ on the spectral variable $m \in \{4, 5, 6, \ldots\}$ with equal-width bins with size $w \in [2, \lfloor m/2 \rfloor]$ and the binned spectral variable $i \in [0, \lfloor m/w \rfloor]$ is defined as

$$\hat{f}[i] = \sum_{j=0}^{w-1} f\left[i \cdot w + j\right] \tag{3.2}$$

on the sequence of spectral measurements. The discrete function describing the complete measured spectrum is divided into equally sized bins by summing the sequential values. Spatial filtering has to be applied to the datacube format instead of on a 2D matrix to be able to use the neighboring intensity values. In the 2D matrix format, these spatial relations are only implicitly present. In spatial filtering, spatially neighboring cells are combined into a single new cell.

3.2.3 Unfolding

The spectral datacube has to be 'unfolded' after acquisition to be able to perform a multivariate analysis. Bro [Bro97] describes it as: "simply a way of rearranging a multi-way array to a matrix". The measured 3D datacube $F[x, y, m]$ is converted into

2D M by XY matrix D with m, the spectral dimension and xy the spatial dimension. Several matching spectral and spatial components are extracted using multivariate analysis on unfolded matrices. Commonly, the resolution of the components is limited to the amount of memory that is available to store a partial solution, for instance the covariance matrix in a PCA. Therefore it is desirable to use the smallest possible matrix D with the highest possible resolution to obtain good results.

A 2D matrix is constructed by unfolding each X by Y image in the spatial dimension into an XY-dimensional vector. This vector represents the spatial dimension at a particular mass-to-charge ratio. The 2D matrix, D, consists of each unfolded spatial vector, see Equation 3.3. Typically, these are $2 \cdot 10^6$ by 256×256 matrices for, respectively, the spectral and spatial dimension. Each row in

$$D = \begin{bmatrix} d_{1,1} & d_{1,2} & \cdots & d_{1,XY} \\ d_{2,1} & d_{2,2} & \cdots & d_{2,XY} \\ \vdots & \vdots & \ddots & \vdots \\ d_{M,1} & d_{M,2} & \cdots & d_{M,XY} \end{bmatrix} \tag{3.3}$$

represents an image at a particular spectral variable m.

3.2.4 Weighting

Most applications of multivariate methods agree on the necessity of data weighting prior to feature extraction. There are several approaches to remove undesired artifacts, unwanted variances and corrections for relative intensity variations. The most common approaches are baseline correction, mean-centering, unit variance scaling, and different methods for normalization. Baseline correction is necessary with MALDI data (see Subsection 2.2.1) and is performed by subtracting a baseline from the complete spectrum. Mean-centering across a matrix D is performed to center the data on the origin of coordinates to remove the mean of the variables. To remove the mean from $d_1 \ldots d_{XY}$, the centering matrix $C_M = I_M - \frac{1}{M} 1_M 1_M^T$ is used for W_M in

$$\tilde{D} = W_M D W_{XY} \tag{3.4}$$

with I_M the identity matrix and 1_m the m-element vector of ones. This is only useful if variables can not be compared directly or have different offsets. An offset is a part of the structural model that is constant across one or several variables [Bro03]. Unfortunately, meaningless variables can contaminate a dataset if emphasis is put on them by mean-centering, which increases the noise in the variance. Variance scaling can also be expressed as a weight matrix with the scale as a diagonal matrix, for instance with inverse of standard deviations of D. Scaling is generally not suitable for spectral data since noise is scaled to be as important as peaks, especially when factor analysis is applied due to the risk of overfitting. Depending on the type of data, different applications of normalization exist. A correct model of normalization must be selected, since normalizing using a wrong model could introduce erroneous trends and artificial variation.

It is very well possible to combine various preprocessing techniques with the methods presented here. Mean-centering would be not appropriate in the TOF SIMS case, as data-processing is performed on full datasets and not only on peaks. The use

of spectral mean-centering would then result in negative values for mass-numbers that give zero counts and therefore the interpretation of the spectral profiles that result from PCA would be much more complicated. The advantage of the sparse format would be lost as every spectral parameter would give a certain number of mean-centered spectral counts, resulting in almost no zeros in the matrix.

Keenan and Kotula [Kee04a] applied a different weighting scheme for PCA to account for Poisson distribution in mass spectrometry data as in

$$W_m = (aG)^{-1/2} \tag{3.5}$$

$$W_{xy} = (bH)^{-1/2} \tag{3.6}$$

where a and b are constants to scale the eigenvalues. The estimates in the Poisson distribution of these variables are $a = \sqrt{d_{..}}/n$, $b = \sqrt{d_{..}}/m$, $g = d_m/\sqrt{d_{..}}$ and $h = d_n/\sqrt{d_{..}}$, where $d_{..}$ represents the total number of counts in D, m is the number of rows in D and n the number of columns. G and H are diagonal matrices with elements of vectors g and h along the diagonals. With this choice of variables, the weighting matrix aG is simply a diagonal matrix with the properly unfolded mean image along its diagonal. The diagonal of the matrix bH consists of the mean spectrum. This way, the data is transformed according to the estimated variance by the mean to distribute the Poisson uncertainty in the data more uniformly.

3.3 PCA-based methods

PCA separates peaks in different uncorrelated spectral components and can simultaneously extract spatial patterns with the distribution of those components. The direct linkage between the resulting spectral and spatial components characterizes the approach. In this work, we combine both spatial and spectral dimensions to form a 2D data matrix and apply PCA to this matrix. This results in finding correlated spatial and spectral features. This way, features are used to discriminate between boundaries of chemical elements on the material surface. We use the resulting principal component vectors to construct different tools for visualization. Features can be made opaque to highlight features of interest, while features with low variances are made transparent. We briefly describe the three different PCA-based methods for parameterless feature extraction. The methods we use are: PCA, PCA with VARIMAX rotation and PARAFAC.

3.3.1 PCA

We extract spatial and spectral components using the well-known method of applying PCA [Wal03; Las98]. In some application areas, this is also called the discrete Karhunen-Loève transform, or the Hotelling transform. In our approach, we unfold a M by X by Y datacube in such a way that a 2D M by XY matrix D is constructed. Common PCA is used to compute a sorted list of N principal components in an orthogonal N by M matrix P using eigenvector decomposition. From these eigenvectors (or 'eigenspectra'), a N by XY matrix Y is calculated by

$$Y = P \cdot D \tag{3.7}$$

that, in this case, can be denoted as the matrix with 'eigenimages'. The matrices D, P^T and Y are defined as

$$\begin{bmatrix} d_{1,1} & \cdots & d_{1,XY} \\ d_{2,1} & \cdots & d_{2,XY} \\ \vdots & \ddots & \vdots \\ d_{M,1} & \cdots & d_{M,XY} \end{bmatrix} = \begin{bmatrix} p_{1,1} & \cdots & p_{1,N} \\ p_{2,1} & \cdots & p_{2,N} \\ \vdots & \ddots & \vdots \\ p_{M,1} & \cdots & p_{M,N} \end{bmatrix} \cdot \begin{bmatrix} y_{1,1} & \cdots & y_{1,XY} \\ y_{2,1} & \cdots & y_{2,XY} \\ \vdots & \ddots & \vdots \\ y_{N,1} & \cdots & y_{N,XY} \end{bmatrix} \tag{3.8}$$

by Equation 2.8, which explains how D can be expressed using N components in P and Y.

In PCA, the following constraints apply: $P \cdot P^T$ has to be a diagonal matrix and $Y \cdot Y^T$ is a square identity matrix I. These constraints define a unique solution for P that is found by the eigenvector decomposition. An eigenvector decomposition $A \cdot P = \lambda \cdot I \cdot P$ is obtained by breaking up a m by m square covariance matrix

$$A = \frac{1}{m} \cdot D \cdot D^T \tag{3.9}$$

from the outer product of matrix D into eigenvectors P with eigenvalues in λ and the identity matrix I. This is accomplished by solving the homogeneous equation $|A - \lambda I| = 0$. The columns of both P and D are forced to be mutually orthogonal and the components are ordered according to the explained variance in the data.

The first principal components in P describe those 'loading vectors' (the spectral profiles) in the datacube with the most spectral variance. Without preprocessing, the first component has only non-negative values [Len04]. The original datacube is projected using the principal components in P as basis functions and results in a matrix Y with the spatial 'score vectors'. The rows in P are the eigenvectors in the spectral dimension and the rows in Y give the scores of the eigenvector on each spatial location. Since the principal components in P are sorted in decreasing variance according to their eigenvalues ($\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$), the first rows in Y represent the highest contribution to the spatial dimension. The last rows with the less associated variances are more likely to represent noise. The components can have both negative as well as positive values.

The 'Guttman-Kaiser criterion' [Gut54] states that components with a eigenvalue lower than one—less than one original variable—should be discarded. There is also 'Cattell's scree test' [Cat66] which basically sets a relative threshold by visual inspection based on the inflection point of the resulting curve of eigenvalues. One problem with this approach is the minimal contrast between different spectral peaks and spatial components. This results in less distinctive regions in the resulting features. Another problem is that the extracted loading and score vectors can be negative, while it is known that spectra are intrinsically positive. This problem can be overcome by splitting the spectral profiles in P in a part with positive values

$$P_n^+[m] = \begin{cases} m, & \text{if } m \geq 0 \\ 0, & \text{otherwise} \end{cases} \tag{3.10}$$

and a part with negative values

$$P_n^-[m] = \begin{cases} -m, & \text{if } m \leq 0 \\ 0, & \text{otherwise} \end{cases} \tag{3.11}$$

where n is the number of components (or rows in P) and m are the spectral variables in the columns. The matching positive spatial distributions Y_n^+ can be obtained by

$$Y_n^+ = P_n^+ \cdot D \tag{3.12}$$

and its similar counterpart Y_n^- can be found by replacing P_n^+ with P_n^- in Equation 3.12.

3.3.2 VARIMAX after PCA

The decomposition given in Equation 3.7 is not unique. There are many variations on the two-way bilinear decomposition similar to PCA [Tim01]. Various optimization criteria and constraints can produce different results. One approach is to rotate the resulting principal components to obtain a better fit on the data without affecting the decomposition using the rotational ambiguity of PCA; i. e., $D = P^T \cdot R \cdot R^{-1} \cdot Y$. Both R and R^{-1} demonstrate the ambiguity, as they can be removed or change the value of P^T and Y. The VARIMAX rotation proposed by Kaiser [Kai58] is one of the most popular criteria for rotation. It can be applied as a post-processing step on extracted principal components. VARIMAX searches for a rotation of the original components in such a way that the variance of the squared principal components is maximized. The axes of the new components remain orthogonal to each other in VARIMAX. The axes are not required to be orthogonal to each other in an oblique rotation, like for instance SIMPLIMAX [Kie94]. For each kth principal component the objective function S_k^2 is computed via

$$S_k^2 = \frac{f \cdot \sum_{i=1}^{m} \left(x_{if}^2 / h_i^2 \right)^2 - \left(\sum_{i=1}^{m} x_{if}^2 / h_i^2 \right)^2}{f^2} \tag{3.13}$$

where is f the number of principal components, m is the number of spectral variables, x_{if} is the loading of spectral variable i on component f and

$$h_j^2 = \sum_{i=1}^{f} P_{ji}^2 \tag{3.14}$$

is the communality of the ith spectral variable in P. The overall variance V, with $V = \sum_{k=1}^{f} S_k^2$ is being maximized until the increase of V drops below a certain threshold (i. e. 10^{-6} in our examples). In theory, the VARIMAX method can improve the contrast between spectral peaks, since rotating the principal component bases will result in sharper gradients in adjacent spectral peaks.

3.3.3 PARAFAC

Another variation of a PCA-like decomposition method is the PARAFAC (PARAllel FACtors analysis) model of Harshman [Har70]. Exactly the same model was independently proposed by Carroll and Chang [Car70] as the CANDECOMP (CANonical

DECOMPosition). Kiers [Kie91] has shown that PARAFAC model can be considered a constrained version of the two-way PCA. The PARAFAC generalization of PCA does not have the rotational ambiguity in its restricted model as in

$$D_k = P^T \cdot B_k \cdot Y + E_k \tag{3.15}$$

where E_k is a residual matrix. This restricted PARAFAC model [Bro98] resembles the model of SVD in which B_k would be a diagonal matrix with the singular values, where k is the number of components. This allows putting several constraints on the decomposition instead of the orthogonality constraint in the SVD. The implementation of the algorithm described by Bro [Bro97] is used in this work to put a non-negativity constraint on the decomposition on P and Y in Equation 3.7 to improve interpretation of the scores. Hereby, the implicit orthogonality constraint of the PCA model is lost.

Other than with PCA, the PARAFAC model has residuals which are not part of a model of noise or error, but of the difference between model and measured data. The components in PARAFAC can not be sorted according to the explained variance as with PCA. Therefore, the number of components has to be known in advance, also in contrast with an eigenvector decomposition. The advantage of the PARAFAC method as we use it, is that the score vectors will always be positive. The resulting features are therefore easier to interpret as only the most positive values in the score vectors instead of the most negative ones have to be included in a selection.

3.4 Results

The quality of the extracted features from three different PCA-based methods is compared. We have identified three important criteria to assess the quality. First, spatially correlated spectral features in the visualization should be distinguishable. As a rule of thumb, the higher the contrast between features, the higher the quality of the visualization. Second, these features should be recognizable as bio-molecules in complex surfaces such as cells and tissue samples in the datacube. For example, do these features represent a cell wall or a tissue, etc.? If so, how well are the recognized spectral features correlated? Finally, are the spectral and spatial features distinct in different regions in the image? These criteria will be used in Subsection 3.4.2 to qualitatively compare the presented methods.

3.4.1 Quantitative comparison

A synthetic spectral datacube was created to be able to make a quantitative comparison between the three decomposition methods. Three different spectra including some overlap in the peaks were used to create a variety in the spectral and spatial dimensions. After this, some different levels of Gaussian noise (mean: 0.000 and with a variance: between 0.0001 and 0.0500) were added to the whole datacube to make it more realistic. A spectral, spatial and 3D view on this synthetic spectral datacube are displayed in Figure 3.1.

The resulting spectral score vectors of the three methods are compared with the original spectra, our ground truth. For a quantitative analysis we use a widely used

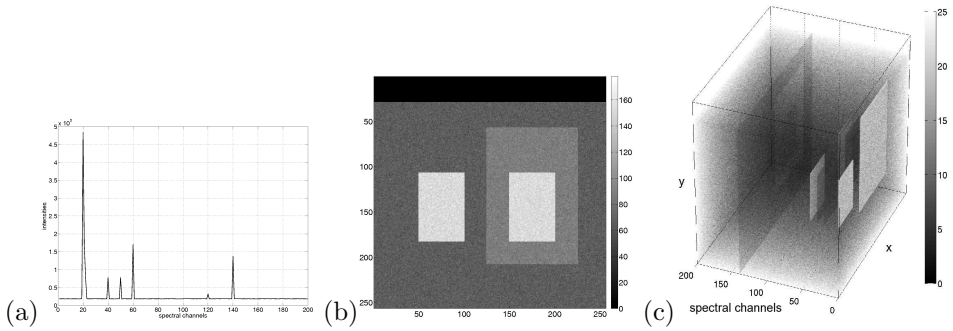


Figure 3.1: *Three views on the synthetic spectral datacube with variance of* 0.0001*. (a) A summed spectral intensity plot with on the x-axis the spectral channels and on the y-axis the intensity. (b) The summed spatial distribution of all spectral profiles. (c) A volume rendering of the complete synthetic datacube.*

Table 3.1: *The root mean squared error of the different components of each method.*

Method	PCA	PCA+VARIMAX	PARAFAC
component1	0.0744	0.0813	0.0235
component2	0.0691	0.0581	0.0112
component3	0.0753	0.0629	0.0249
Total ε	0.2190	0.2023	0.0597

measure of error, the Root Mean Squared Error (RMSE, ε) similarly used in other analyses of correlated spectral data [Sca93]. The absolute values of the resulting spectral component are compared with the synthetic component according to

$$\varepsilon_{method} = \sqrt{\frac{\sum_{i=1}^{m} \left(|component_i| - synthetic_i\right)^2}{m}} \tag{3.16}$$

The number of spectra in the datacube is represented by m and results in a ε for each method. Each method is able to distinguish between the three different spectral components, while the other components clearly contain the added noise. An overview of ε of each method is shown in Table 3.1.

This table clearly indicates that the PARAFAC decomposition results in the least amount of error. Also the VARIMAX rotation provides a better fit compared to the

Table 3.2: *The total root mean squared errors of each method with different levels of Gaussian noise.*

variance\Method	PCA	PCA+VARIMAX	PARAFAC
0.0001	0.2259	0.1983	0.0352
0.0010	0.2190	0.2023	0.0597
0.0100	0.2210	0.2038	0.1352
0.0500	0.2307	0.2219	0.1613

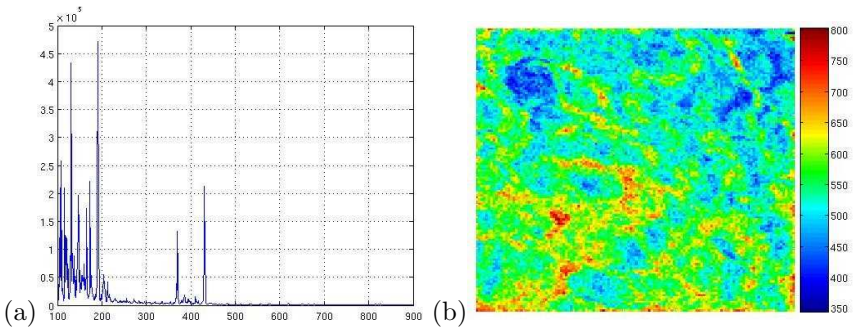


Figure 3.2: *(a) A summed spectral intensity plot with on the x-axis the mass-to-charge ratio and on the y-axis the measured intensity. (b) The summed spatial distribution of all spectral profiles in the datacube.*

use of PCA without a rotational fit. To gain better insight into the influence of the added Gaussian noise, different levels of noise are introduced as shown in Table 3.2. Table 3.2 shows that also when noise levels are rising, the ε of the PARAFAC method still remains lower than the ε of the other two methods.

3.4.2 Qualitative comparison

Imaging MS is used to analyze the spatial organization of intact biomolecules in complex surfaces such as cells and tissue samples. It is particularly useful to directly visualize peptide and protein distributions in invertebrate or mammalian tissue. In the imaging MS data used here a 15 kV Indium primary ion beam is rastered over the surface of a cryosection of the cerebral ganglia of the freshwater snail *Lymnaea Stagnalis*. A data array of 256×256 x,y-coordinates, is generated with each position containing an entire mass spectrum. Each square pixel represents an area of approximately 500×500 nm. Prior to the experiment the tissue surface has been covered with a thin layer of 2,5-dihydroxybenzoic (2,5-DHB) acid by electrospray deposition (called 'matrix material') to enhance the generation of intact biomolecular ions. The mass spectrometer used was a time-of-flight mass spectrometer. High-resolution molecular ion maps have previously shown to provide insight in the spatial organization of various biomolecules in these brain sections [McD05]. Figure 3.2 shows the summed spectral intensities and the TIC image of the spectral datacube in this example.

Manual interpretation of these types of datasets is a time-consuming procedure, where either the spectral peaks of interesting spatial features or the spectral images of interesting peaks in the spectrum are inspected. In order to identify spatially correlated spectral data (often attributed to a specific compound), statistical analysis tools are called for. Here, we qualitatively examine the results of the three different multivariate statistical analysis algorithms applied to a single MS image dataset of the brain of *Lymnaea Stagnalis*.

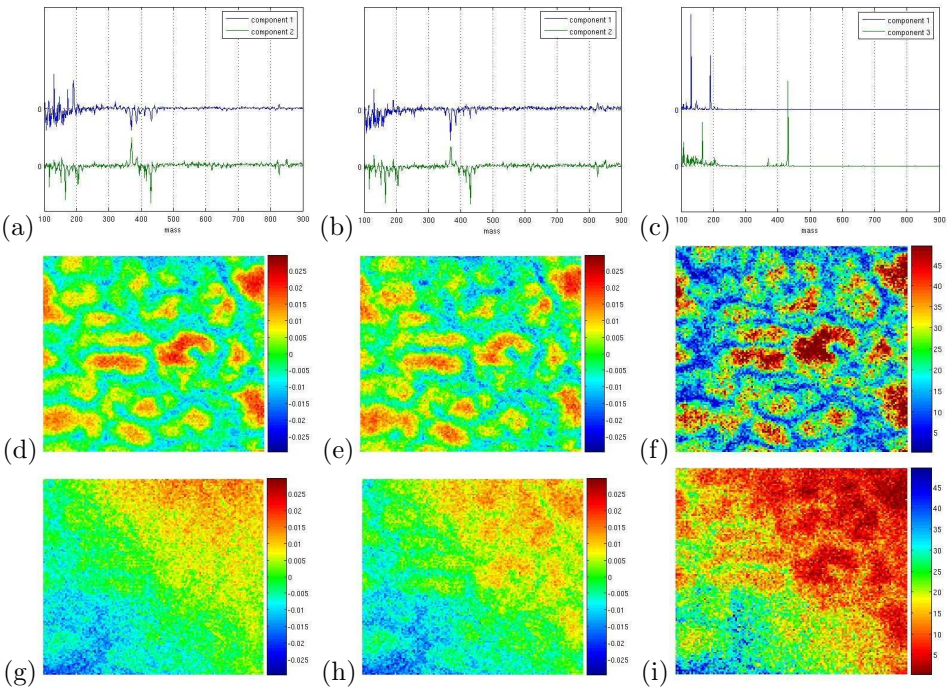


Figure 3.3: *Two spectral (a) and their two spatial (d–g) components derived using PCA. Two spectral (b) and their two spatial (e–h) components derived with PCA and VARIMAX rotation. Two spectral (c) and their two spatial (f–i) components derived with PARAFAC.*

PCA

The reconstructed score vectors in Figure 3.3a display two separated components where the first component (blue spectrum) contains predominantly spectral features that are clearly correlated to the applied 2,5-DHB matrix material as positive peaks. Intermixed with this compound negative spectral features of cholesterol (m/z 368 and 385) are also observed. As in regular spectra the 2,5-DHB peaks usually constitute the base peak in the spectrum. This method seems to underrepresent the spectral intensities. The spatial features in (d) are distinctively related to the areas in between individual cells that seem to indicate a stronger 2,5-DHB signal is found there.

The second component (green spectrum) found is again a mixture of cholesterol, but now includes positive peaks and a peak at m/z 425 that previously has been attributed to APGWamide. It also contains some higher mass lipid molecules around m/z 815. The spatial features of the individual cells in (g) are barely recognizable.

PCA with VARIMAX rotation

The second method, PCA with VARIMAX rotation, shows similar spectral features in its components but judging from Figure 3.3b an improved spectral correlation is found. Some of the individual cells can be recognized by the higher values in yellow compared with the green areas that surround them. This is also obvious from the improvement in quality of the images in (e) and (h) compared to similar spatial components from PCA in (d) and (g). A better feature contrast is found, but the individual components are not fully separated.

PARAFAC

The PARAFAC approach offers a spectral view that is more similar to the spectral view the mass spectrometrists are used to. In addition, the relative intensities and signal-to-noise ratio in the two spectral components (in Figure 3.3c) are as would be expected from these types of measurements. More importantly, a much better separation between the 2,5-DHB-related peaks and the cellular peaks is obtained. This also results in better contrast in the feature images in Figure 3.3f and i, facilitating an easier localization of the chemical compounds. Note that color-scale in (i) has only positive values and is reversed compared the similar component images in (g) and (h). This is caused by the positivity constraint in PARAFAC. This scale is reversed to make it more comparable with Figure 3.3g and h.

The smaller peaks around m/z 815 are not clearly visible on this scale in the (green) spectral component, but may be incorporated in other component spectra. It becomes clear that the compound separation and localization has significantly improved. This can be observed in the more clearly defined peaks in the spectral component and the increased contrast in the image components.

3.4.3 Performance

Performance regarding the computational effort is another aspect for the application of methods for multivariate data analysis. An estimate of computation time was made for the studied methods. All time-measurements were done on the same

computer (single processor 32 bit AMD Athlon, 2.2 GHZ, 1 GB of memory), using MatLab^{TM} 7.1 with the N-way toolbox 2.11 [Bro97] and VARIMAX implementation. Calculations were done in a 32-bit environment. This limits memory allocation and therefore the maximum size of the analyzed dataset to 4 GB. The use of a 64-bit environment would circumvent this memory problem and therefore makes the use of larger datasets possible. However, this would also increase calculation time. The size of the quantitatively analyzed datasets was chosen in such a way that the total calculation could be done without the need of virtual memory. The use of virtual memory would dramatically increase the total calculation time, because hard disk access is much slower than RAM access. This would not result in a representative measure when algorithms are compared.

Memory

The memory requirements for the application of PCA involve calculating the eigenvalue decomposition of a covariance matrix of a dataset. The environment with a PCA implementation sets limits on the size of the dataset that can be analyzed. These limits depend on the amount of (virtual-)memory that is available on a system and the maximum size of a variable containing the data. It would be beneficial regarding memory consumption, if the PCA is implemented on sparse matrix structures, because most data-points in imaging MS data contain zeros. Standard approaches for solving a non-Hermitian eigenvector problem generally have a $\mathcal{O}\left(N^2\right)$ space—or memory—complexity [Bai00]. It is not possible in these standard approaches to exploit the sparsity of the matrix to reduce memory requirements. Different iterative alternatives have been proposed [Bai00] that are able to find a subset of eigenvectors within a sparse matrix structure. These alternative approaches can be tailored by many parameters to solve the eigenvector problem for different classes of matrices. The actual performance of each approach depends on the implemented approach, the values of its parameters, and the properties of the data within the sparse matrix.

Standard memory tests within the profiling utility of MatLab^{TM} give an indication of the memory use by the implementation of MatLab^{TM} to solve the sparse eigenvector problem. MatLab^{TM} sets a maximum to the size of its workspace, the size of the largest matrix and the number of elements in the largest arrays. The MatLab^{TM} environment supports the use of sparse matrices as well as the operation for eigenvalue decomposition. There are not many ready-to-use implementations available for the eigenvalue decomposition of sparse matrices. MatLab^{TM} environment provides a simple interface to routines for processing sparse matrices and create an eigenvalue decomposition [Leh96]. This environment is widely used and provides stable and efficient implementations of the routines used in this thesis.

The used single mass spectral datacubes contain approximately 10^{11} elements of which on average $2.5 \cdot 10^7$ elements are non-zero in the used datasets. A complete datacube will fit into memory only when a sparse format is used because it exclusively stores the non-zero elements. It depends on the implementation of a sparse matrix multiplication whether or not there is enough memory to calculate the covariance matrix. Current iterative implementations for eigenvector decompositions are efficient in memory usage to be able to find the desired principal components as soon as a (sparse) covariance matrix can be constructed. The sparse covariance matrices

Table 3.3: *Indications of computation time in seconds using various methods on various samples. With #:number of components, sPCA:sparse PCA, VX:VARIMAX and PFAC:PARAFAC. The VARIMAX processing time is given as the time added to PCA.*

set	#	size	PCA	sPCA	+VX	PFAC	sPFAC
hair	7	$300 \times 256 \times 256$	3	3	+0.15	3500	2000
droplet	7	$300 \times 256 \times 256$	3	3	+0.15	1200	6500
hair	14	$300 \times 256 \times 256$	3	3	+0.25	6000	5000
droplet	14	$300 \times 256 \times 256$	3	3	+0.25	40000	50000
hair	21	$300 \times 256 \times 256$	3	5	+0.35	14000	13000
droplet	21	$300 \times 256 \times 256$	3	4	+0.35	160000	85000
hair	7	$5053 \times 64 \times 64$	$5 \cdot 10^2$	25	+0.20	900	1000
droplet	7	$5053 \times 64 \times 64$	$5 \cdot 10^2$	20	+0.20	700	600
hair	14	$5053 \times 64 \times 64$	$5 \cdot 10^2$	25	+0.30	3500	3000
droplet	14	$5053 \times 64 \times 64$	$5 \cdot 10^2$	20	+0.30	9000	8000
hair	21	$5053 \times 64 \times 64$	$5 \cdot 10^2$	30	+0.40	6000	4000
droplet	21	$5053 \times 64 \times 64$	$5 \cdot 10^2$	20	+0.40	30000	27000
LDI	7	$1850 \times 290 \times 7$	30	35	+0.15	52000	55000
LDI	14	$1850 \times 290 \times 7$	30	35	+0.30	212000	214000
LDI	21	$1850 \times 290 \times 7$	30	35	+0.40	498000	511000

in these examples were between 8 MB and 150 MB. The memory usage for the calculation of the eigenvectors within MatLab^{TM} took between 16 MB and 256 MB depending on the properties of the spectral dataset. These properties include the amount of non-zero elements in the datacube and the variance in the data.

Time

Computation time was evaluated for three datasets in Table 3.3. Three samples were studied using TOF SIMS imaging: a purely synthetic sample containing well-defined chemical components as a droplet-array, an embedded hair cross-section and a third sample was measured using Laser Desorption and Ionization (LDI)-TOF imaging: a cross-section of paint layers. Two different datacubes were used for the TOF SIMS datasets: one with a large spectral dimension and one with a large spatial dimension. The number of components was varied from 7 to 14–21. The LDI-TOF imaging datacube was analyzed at full spatial resolution (7×290) and with 1850 spectral variables.

The standard PCA method first calculates the full and exact decomposition and then restricts the resulting dataset to the requested number of components. PCA performed on sparse matrices produces an approximation by itself, not giving a full representation of the original datacube, but only resulting in the requested number of components. The resulting sparse components are approximations because the data-matrix itself has to be approximated in the sparse decomposition. The equivalent, non-sparse data-matrix would be too large to be decomposed. The difference in methodology contributes to the time-reduction that is involved in the use of sparse matrices.

The continuous nature of the LDI data, with a non-zero entry at almost each

sampling point resulted in an increased computation time when the sparse matrix format was used. This can be explained from the fact that the in-memory size is larger for the sparse-type matrix than for the full matrix, which inevitably leads to larger processing times. VARIMAX as a post-processing optimization step after PCA results in only a small increase in calculation time. This justifies the use of VARIMAX after PCA in any case to increase chemical contrast in both component images and spectra, as shown in previous subsections.

PARAFAC is clearly a much more demanding technique with an increase of a factor of 10–1000. Although it turned out to be better at resolving certain features, it is not suitable for routine use with the current standings of desk computer facilities. It could be very helpful in very complex systems or in systems where trace amounts of a certain chemical components are expected. Prior knowledge, which is favorable to make a sensible choice for the number of components to be looked for, could be obtained using PCA. Like PCA, PARAFAC turned out to be faster on sparse matrices. It should be mentioned that the random initialization as used in the PARAFAC calculations, results in a large variation in calculation time and the order of the factors. PARAFAC is a computationally much more demanding technique because it seeks an exact fit of the data using optional constraints, spread over the defined number of factors.

3.5 Summary and conclusion

In this chapter we have compared the quality of three different PCA-based methods for the 3D visualization of imaging spectroscopy data. We used PCA, PCA with VARIMAX rotation and the PARAFAC method. We compared the methods quantitatively and qualitatively. For the quantitative comparison, we used a RMSE metric to compare the methods with ground truth spectra under various noise conditions. For the qualitative comparison, we used three criteria to judge the quality of features in the resulting visualizations. These criteria were applied to interpret the visualizations of features in the brain of the snail *Lymnaea Stagnalis*.

This study shows that the PARAFAC method is clearly superior to the other methods. PARAFAC results in features that are more clearly recognizable than the other two methods (see Figure 3.3). The reason is that PARAFAC uses some model information, while PCA does not. The VARIMAX rotation uses a post-processing fitting to maximize the variance of the components which results in images and spectra with higher contrast.

We learn from the synthetic data case that although the root mean squared error becomes larger with higher noise levels, the PARAFAC method still produces the most distinctive results. We expect that these trends are similar in the real life application. The implication is that more noisy samples will still result in good visualizations.

Chapter 4

Feature-based registration

Chapter 3 described how PCA and PCA-based methods can be used to extract features from imaging spectrometry data. These methods provide an automatic extraction of spectral and image components from one measurement (a spectral datacube). When multiple measurements are made from the surface of the same sample, each measurement has to be aligned with another to create one single multi-spectral image of the sample to be used for analysis.

This chapter presents a feature-based method to facilitate the automatic registration of spectral datacubes. Features are used together with a similarity measure to find overlapping regions between multiple measurements. To exclude potential incorrect alignments, a registration has to made more robust by adding a measure to indicate the randomness within a region. This measure of randomness acts as a weight to express the similarity in the overlapping regions more accurately. The performance of this feature-based method is compared with a pixel-based registration method.

4.1 Introduction

The surface area of a typical biological sample is too large to be recorded in one measurement by an imaging mass spectrometer. In a typical high spatial resolution SIMS measurement the maximum analysis areas are approximately $50\,\mu m$. This is much smaller than many of the samples of interest, for example biological tissue samples of 2×1 cm.

To provide high-resolution imaging of the complete, large areas, the sample is divided into a mosaic of small areas (termed 'tiles') with the sample stage raster. Each tile is then analyzed with a high spatial resolution measurement. Moreover, the multiple spectral datacubes have to be combined to provide the final, complete (mosaic) dataset. Compared to the high resolution microscopic image in Figure 4.1a, the spectral TIC image mosaic discloses the chemical composition of surface material on each location. This TIC image in Figure 4.1b is created by the sum of all spectral intensities present at a single two-dimensional location. Instead of the black background in between the crystallized droplets in the microscopic image, the spectral image provides

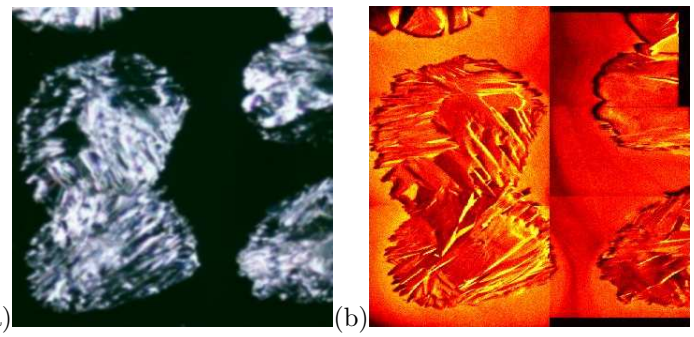


Figure 4.1: *(a) Microscopic image and (b) spectrometric TIC image mosaic of a droplet-deposition on a silicon substrate.*

spectra of the material in between the crystals as false color information.

Sample stages capable of moving through large areas (e. g., 5×5 cm) are often significantly less accurate than the resolution of the imaging mass spectrometry experiment. Therefore, it is not possible to make a measurement of a tile that is precisely located alongside another tile. Each tile is acquired with a small, but unknown amount of overlap with its neighboring tiles. These overlapping regions make it possible to reconstruct a complete mosaic with the different tiles. As a result, the high-resolution spectral images have to be aligned after data acquisition. In most examples this has been performed 'by eye'. This process is time-consuming, user intensive, and subjective. At best, a manual alignment process would include the following steps:

1. Import the data into a mathematical package capable of processing $>$ 2 GB data files.

2. Calculate the TIC images of each tile.

3. Create a layered image to manipulate relative positions of the TIC images.

4. Vary the relative positions until the 'optimum' is found.

5. Use the offsets determined in step 4 to combine the data files into a single datacube.

However, it is more common to use the proprietary software to create the images for each m/z range of interest of each tile's dataset and to apply a pixel-based registration method. The TIC images are then aligned in a graphics package, the relative offsets are applied to the specific spectral images, and the results are saved. In fact, no case studies are known in SIMS imaging in which multiple tiles were combined into one dataset and in which all raw data events could be preserved. Recently, more advanced techniques for data acquisition are able to sample a larger surface area by capturing a number of datacubes on a higher resolution. However, these kinds of state-of-the-art datacube registration still use TIC images to create one large mosaic of datacubes. These methods are very similar to pixel-based registration techniques

found in many image processing handbooks, for instance Ibanez and Schroeder [Iba03] or image registration for remotely sensed spectral data [Cha99].

This chapter describes a new feature-based, automated image alignment algorithm for imaging mass spectrometry datasets. Automatically combining several datacubes is done to increase spectral information and the range of the spatial area to improve analysis.

Usually, the uncorrelated noise prevents the optimum offsets of two images from being determined. This noise often blurs or hides edges of recognizable spatial distributions present in an image that can be used for alignment. Feature extraction by PCA reduces uncorrelated noise and can therefore be exploited for automated alignment. Also, spectral datacubes do not always contain useful features at the overlapping regions that can be used for landmarking approaches. Finally, the spectral datasets have a relatively low signal-to-noise ratio compared to other image registration problems. Although there are many tiles available, there is neither a fixed ordering nor a fixed overlap between the different measurements. In addition, each overlap is relatively small compared to the complete image.

Fortunately, instead of having to apply some additional rotation, scaling, shearing or nonrigid transformations, the tiles only have to be linearly translated in two dimensions. Another advantage is that acquisition will always provide overlap with other regions in the collection.

In the next section, we describe the algorithm as a whole, as well as methods for selecting and registering features and for adding additional weight to specific areas. Subsection 4.3 compares our feature registration method to the pixel-based method on two test collections. Finally, a discussion of the pros and cons of our method concludes this chapter.

4.2 Approach

Our registration method can be divided into three parts. In the first part, the data is reduced by selecting and extracting the most important features from the datacubes using PCA. Then, by applying the mean squared error metric to corresponding features in adjacent cubes, a minimalization landscape is constructed. This landscape represents the 'fit' of the feature when two adjacent datacubes are aligned with each other. The minima provide the regions with the most similarities. Under various conditions, however, the landscape does not provide sufficient information to robustly assume that the lowest value of the landscape is indeed the desired solution. These conditions can be described in the third step by analyzing the entropy of the landscape and to add a weight to all of the possible solutions. The resulting algorithm in Subsection 4.2.4 gives an overview of how a solution is found in the final search space.

4.2.1 Principal Component Analysis

First, PCA is applied to the datacubes to reduce them by preserving the most important features into a number of spectral components and their corresponding image components. There are many methods to decompose multidimensional data or to apply dimension reduction [Har70; Car70; Moi02], but these are better suited when many independent features—rather than only a few components—describe most of the

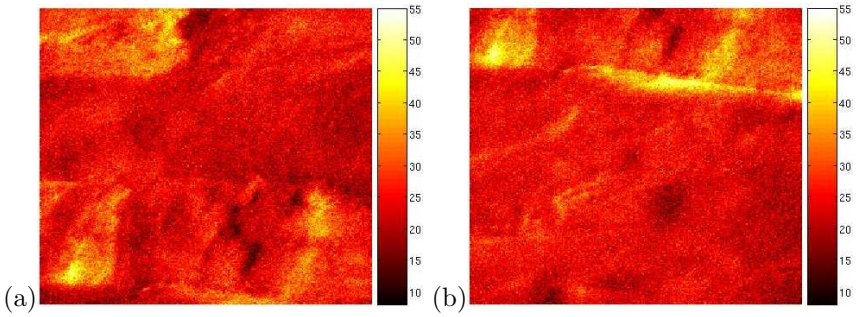


Figure 4.2: *(a) One image component and (b) an other spectrally matching image component.*

information. PCA can be used to compress the thousands of image planes into a few image components. PCA selects correlated spectral and spatial features in the datacubes, which results in different components as described in Subsection 3.3.1. Once two similar spectral components can be found in both pieces of the same collection, the corresponding image components (see Figure 4.2) can be used for registration. For example, Figure 4.2a shows an area in yellow at the bottom left corner that contains some high intensity values. A similar area can be found in Figure 4.2b in the top left corner. The similarity between both regions can be measured using an appropriate similarity metric.

4.2.2 Mean Squared Error

Although many metrics and applications already exist for the registration of images [Fon97; Ran05b; Zit03], we use the most basic approach. In most cases, there are no clear defined edges or distinguishing regions present in the resulting image components. Landmarking or region-based registration could be problematic when applied to this type of datasets. Most of these approaches use a metric to measure the difference in intensity values between two regions in two images and an optimizer to transform one of the images according to a fitness value to find a better fit. One image is the so-called 'fixed image', and the image that is being transformed is the 'moving image'. Because these spectral datacubes do not have many spatial features and could contain some noise, the complete search space has to be considered to find a suitable minimum and a correct offset between the two images. There is no longer a need for an optimizer. However, this approach is only possible when both images are small enough and—more importantly—that the moving image does not need any rotational, scaling or warping transformations.

The most simple pixel-based similarity measure is the Mean Squared Error (MSE) measure. The mean squared pixel-wise difference is calculated using

$$MSE(A,B) = \frac{1}{N}\sum_{i}^{N}(A_i - B_i)^2 \tag{4.1}$$

where A is a region in the fixed image, B a region in the moving image, N the number of pixels in these regions and i the pixel position.

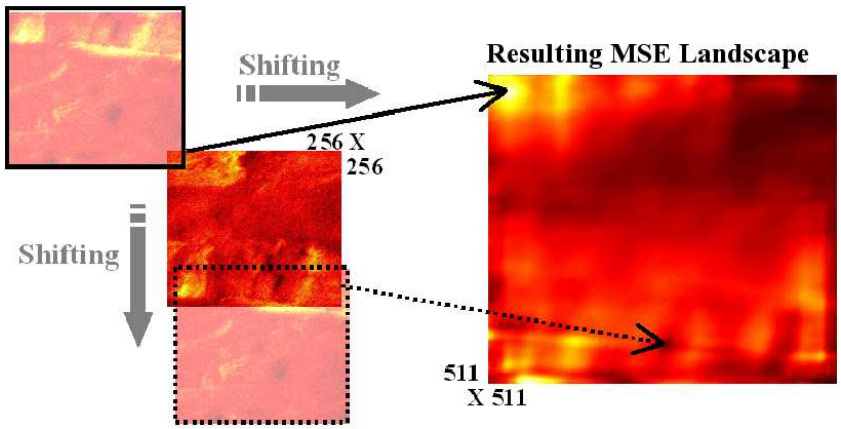


Figure 4.3: *Shifting two image components of* 256×256 *pixels to create a MSE landscape of* 511×511 *pixels of each overlapping region to find the correct minimum represented by the dark colored points.*

Linear changes in intensity between both images will result in a poor match value and poor matches result in large values of the metric. This metric is used on all subregions of two image components, as can be seen on the left in Figure 4.3. It results in a 511×511 search space in which each point is the result of the MSE metric applied to a combination of two regions. The Normalized Cross-Correlation (NCC) metric [Pra91] also uses images of the same modality. It is invariant to linear changes in intensity and it is robust to noise. Poor matches result in high values of the metric with well-defined minima and sharp peaks. The metric is sensitive to clutter, occlusion and non-linear changes in contrast. It is not used in this case, because it does not perform considerably better than the MSE metric on these spectral datasets.

4.2.3 Entropy

In the complete MSE landscape on the top right in Figure 4.3, there are still many (black) areas with low values. This is mainly caused by the fact that the intensities in certain regions do not have enough contrast and/or a high amount of randomness and simply do not contain enough information to accurately use the similarity-metric. If the subregions consist of only one pixel (at the corners of the MSE landscape), it is practically impossible to find an accurate metric for the similarity. Image characteristics can provide some statistics about the information in an image [Gon03]. These statistics can give an indication of fitness of the region in an image relative to a region in another image. This characteristic can be used to provide a weight for the MSE landscape and to create a more realistic search space in order to find the most appropriate minimum. Some commonly used texture metrics are contrast, correlation, energy, entropy, and homogeneity [Ooi06]. In this context the measure for entropy in Chalermwat [Cha99] is the most suitable to act as a weight. It provides an indication of the 'randomness' of intensities in an image using its histogram. The

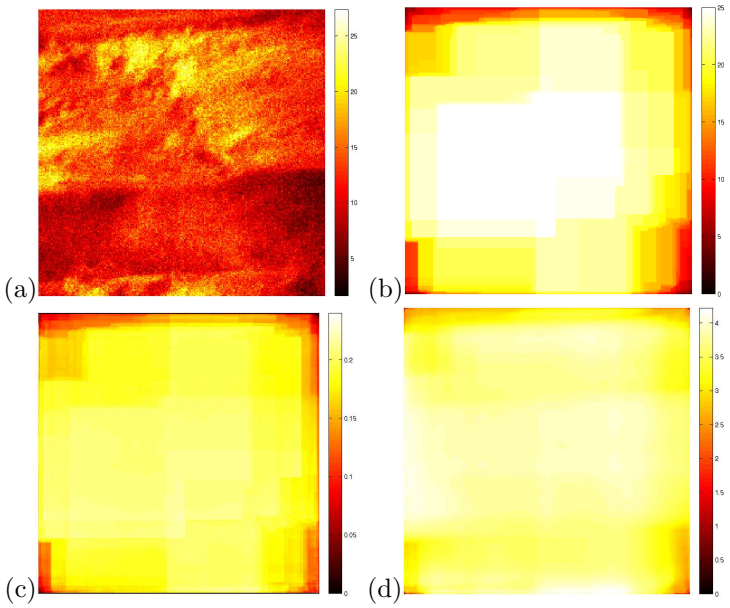


Figure 4.4: *Different texture measurements applied to the partial windows: (a) the image component, (b) differences between minimum and maximum values, (c) homogeneity, and (d) entropy.*

entropy H of each region in one image component is calculated using

$$H(A) = -\sum p_A \cdot \log p_A \tag{4.2}$$

where A is the region in an image component and p_A the histogram values of A. Many registration algorithms for multimodal images also use entropy as similarity-metric instead of weight function. This similarity-metric is called 'mutual information' and measures the mutual dependence between image regions. The mutual information of two images can be expressed in terms of entropy for registration. In this approach entropy is used as weight function to eliminate regions with high uncertainty in the spatial distribution of intensity values.

The entropy of each region of the fixed image component can be calculated and put into a landscape corresponding with the MSE landscape from Figure 4.3. The entropy landscape of the moving image component can be created similarly and combined with the fixed entropy landscape using the entry-by-entry product of both entropy values. The resulting combined entropy landscape of the image components in Figure 4.2 is shown in Figure 4.5a. This combined entropy landscape contains some outliers, mostly located at the corners where only a few pixels are being considered. An additional fit on the histogram is made to remove unwanted outliers. This is done according to the normal distribution, by using only the values in the interval $[\mu - \sigma, \mu + \sigma]$, where μ is the mean and σ is the standard deviation in the histogram. The values in the histogram to the left of this area are set to zero and the values to the right of this area are set to the remaining maximum. The remaining landscape (see Figure 4.5b) is applied as a weight for the MSE landscape.

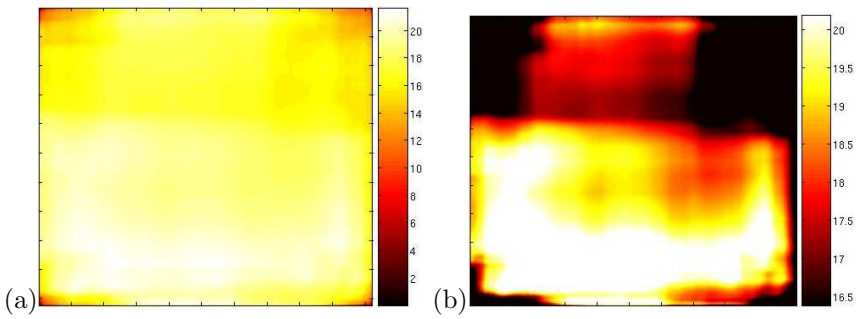


Figure 4.5: *(a) Entropy space of all regions in the image component of Figure 4.2a and (b) the standard deviation fitted entropy of the same image component.*

The black regions in entropy spaces indicate that a minimum found in the corresponding regions of the MSE landscape most likely will not contain enough information to be considered a solution. The final weighted search space can be created by dividing all values in the MSE landscape by the fitted entropy landscape. This results in a $\frac{MSE}{entropy}$ search space in which a minimum has to be found. Once the location of this minimum is known, the offset of the translation between both images can be calculated with the correct registration between both image components.

4.2.4 Algorithm

With the use of the previously described methods, a PCA-based registration algorithm was created to find the correct offset between two datacubes, if such an offset exists. If more similar spectral components are present, more minima in the $\frac{MSE}{entropy}$ search spaces can be compared in order to find the correct offset. The algorithm proceeds as follows:

1. Apply PCA on two datacubes from one collection of measurements which results in a decomposition of each datacube with a number of corresponding spectral and image components.

2. Select matching spectral components of both datacubes by comparing spectral peaks, starting from the components describing the most variance.

3. Registration of the corresponding selected image components from both datacubes by creating a MSE landscape of all possible combinations of regions in the image components.

4. Add a weight function to the MSE landscape using an entropy image characteristic applied to each region of both of the image components to exclude regions that do not contain enough spatial information to be considered a solution.

5. Select the lowest points in the first five regions with local minima in the $\frac{MSE}{entropy}$ search space in such a way that a selection of points remains which are the most appropriate solutions for a correct registration.

6. Repeat step 2 to 5 for the next two corresponding spectral components from the PCA decomposition of both datacubes in order to get two sets of points that could be solutions for a correct registration.

7. If present, select the point with the lowest value in both selections of appropriate solutions, in order to find the best solution for a registration. If there is no such point, no suitable offset between both datacubes can be found.

This algorithm can be applied to each combination of spectral datacubes in a collection of measurements. If it is known which datacubes in a larger mosaic are neighbors, this algorithm is applied to find the correct offset. When this information is not available, each combination of pairs has to be considered. Two datasets are used in the next section to compare this PCA-based algorithm with a traditional approach.

4.3 Results

The algorithm is applied to two collections of spectral datacubes resulting from imaging mass spectrometry. These datacubes have two spatial dimensions (256×256 pixels) and one spectral dimension (more than three thousand image planes). A dataset has to be read, transformed into a datacube and binned before PCA can be applied. The time it takes to perform these operations depends on the size of a dataset and the variation in the intensity distribution of the spectral variables. The time it takes to create a MSE-landscape or to create an entropy-space is independent from the properties of a dataset. With two image components of 256×256 pixels, it takes approximately 125 seconds to create a MSE-landscape of 512×512 pixels on a single processor of a 32 bit AMD Athlon (2.2 GHZ). One entropy-space of one image component is created in approximately 350 seconds.

4.3.1 Two collections

One dataset of four overlapping datacubes (see Figure 4.6a for their relative positions) is a measurement of an array of crystals as shown in Figure 4.1. They were produced from aqueous solution by droplet-deposition on a silicon substrate. The crystals consist mainly of dihydroxybenzoic acid (DHB), a compound which is widely used as matrix material in matrix-assisted mass spectrometry techniques. Measurements were done using imaging TOF SIMS in microprobe mode [Cha99] on a Physical Electronics TRIFT-2 time of flight mass spectrometer. Each imaged area was 200×200 μm^{-1}. The high abundance of DHB results in distinct peaks in the mass-spectrum. The low total signal intensity in one of the corners of each of the images is due to inaccurate alignment of the primary ion beam.

Another dataset shows a mosaic of the kneecap of a mouse which contains 85 datacubes recorded with the same mass spectrometry technique as the droplet decomposition. A small part of only five datacubes (see Figure 4.6b) was taken from the complete collection in this first approach to test this feature-based registration method. In contrast to the crystal dataset there are some datacubes that do not have an overlap with one of the other cubes in the dataset. There is no overlap between the three combinations: A-E, B-E and B-D in Figure 4.6b. Both collections of four

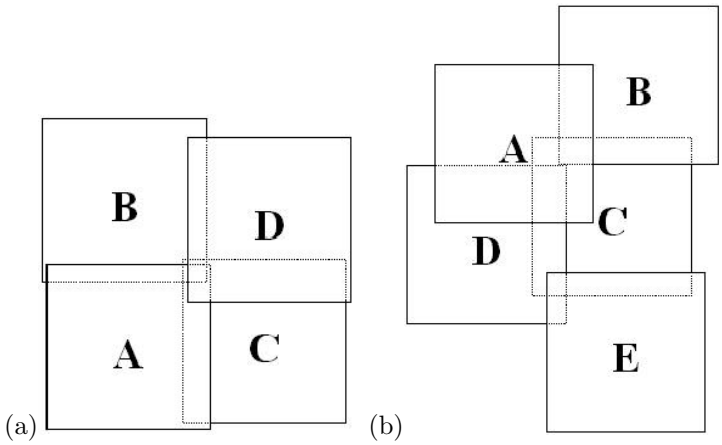


Figure 4.6: *The relative locations of (a) four pieces of the crystal dataset and (b) five pieces of the kneecap dataset.*

and five datacubes are used to compare a traditional TIC-based registration with the PCA-based registration.

4.3.2 Application

Since there has not yet been an attempt to register imaging MS datacubes automatically, the TIC images are used in a manual approach to stitch different spectral image planes. These images are constructed by taking the summation of all image planes of the spectral datacubes. Each combination of these resulting images can be registered with each other using the same $\frac{MSE}{entropy}$ metric which is used in the PCA-based approach to be able to quantitatively compare both approaches. The crystal dataset consists of four datacubes which results in six combinations, in which there is an overlap with every other datacube. The kneecap dataset consists of five datacubes which results in ten combinations, in which seven combinations do have an overlap and three of them do not. The results of this registration can be found in the next subsection.

All combinations in each dataset are being registered using this new PCA-based method. First a PCA is conducted on each datacube using the algorithm from subsection 4.2.4. Similar spectral components are being matched and each corresponding image component (see Figure 4.2) is used for the registration. This results in several MSE landscapes (see Figure 4.7) for each matching spectral component in which minima have to be found to get a most appropriate solution for a registration.

The entropy space (see Figure 4.5) of each component image is calculated and combined into the $\frac{MSE}{entropy}$ search space to add a weight to the solutions according to the amount of information that is present in each region. Figure 4.8a shows the $\frac{MSE}{entropy}$ search space of the first spectrally matched components of two datacubes. And Figure 4.8b shows the $\frac{MSE}{entropy}$ search space of the next spectrally matched components of the same two datacubes. The white regions in both images are caused by the entropy weight. The overlapping regions in these white areas do not contain enough

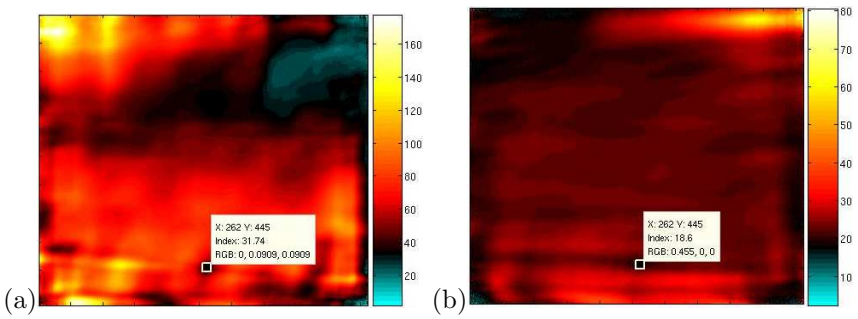


Figure 4.7: *Mean Squared Error landscape from comparing the (a) first and the (b) second principal component of C and E of the kneecap dataset in which the location of the correct solution for the registration is marked.*

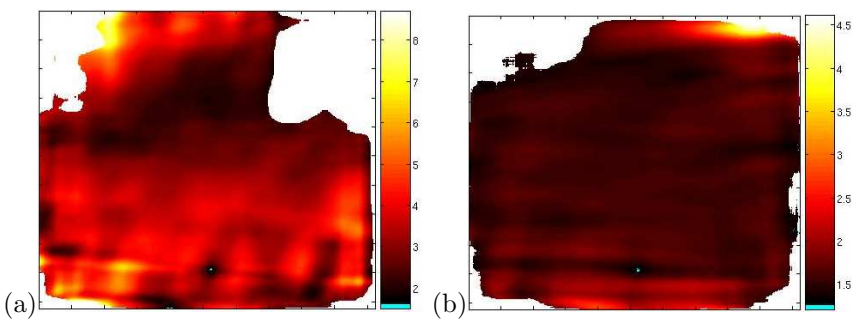


Figure 4.8: $\frac{MSE}{entropy}$ *search space of the (a) first principal component and the (b) second principal component with the possible solutions in the blue color.*

Table 4.1: *Summary of the results from the TIC-based approach.*

TIC-based	array of crystals		kneecap of a mouse	
	found	not-found	found	not found
global fit	0	6	2	5
incorrect fit	–	–	0	3

Table 4.2: *Summary of the results from the PCA-based approach.*

PCA-based	array of crystals		kneecap of a mouse	
	found	not-found	found	not found
correct fit	5	1	5-1	2
incorrect fit	–	–	3	0

intensity information to be considered in the solution. Both landscapes do have a global minimum on the same location which should be the best solution for the translational offset between both datacubes. The same procedure is applied to all other possible combinations to find those offsets or dismiss the combination if there is no joint local minimum in both search spaces and therefore not an overlap of regions.

4.3.3 Comparison

The results of the six combinations of TIC-based registration of the crystal dataset together with the ten combinations of the kneecap dataset are presented in Table 4.1. None of the correct offsets between the six combinations in the crystals dataset could be found using the TIC-based approach as shown in the row with the 'global fit'. Only one $\frac{MSE}{entropy}$ search space could be considered for each combination of datacubes, so the global minimum in the search space was used as a solution for a best fit. There were only two correct offsets found successfully within the kneecap dataset using the global minimum in the $\frac{MSE}{entropy}$ search space in this TIC-based approach. There was no indication available in the TIC-based approach whether or not the global minimum would result in a incorrect fit. This is the reason why the three incorrect fits of the kneecap dataset were put in the 'not found' column.

Table 4.2 contains the corresponding results with PCA-based registration. All but one of the correct offsets between the six combinations of the crystal dataset could be found. In some cases there was only one pair of spectral components that gave a match in a combination of datacubes. The reason is probably that there was not enough similar information present to find more than one match of spectral components that made a comparable contribution to the complete datacube. In those cases, only the match and its $\frac{MSE}{entropy}$ search space were used and its global minimum was considered as a solution. The three combinations in the kneecap dataset that did not have any overlap were correctly found as indicated in the row with the 'incorrect fit'. Unfortunately, two offsets that did exist could not be found. This was caused by the lack of entropy in regions that should provide for a correct fit. The remaining four out of six existing offsets were found and one was found, but was incorrect. This solution was found by comparing two regions without much intensity information.

The joint entropy was not low enough to dismiss it as a possible solution.

4.3.4 Enlarged dataset

After successfully finding the correct offset between two datacubes, the two datasets can be fused together by combining the raw ion counts of the different spectral mass measurements. The feature-based registration algorithm can be applied to this fused dataset and the remaining pieces in the collection. There are two advantages in using a registered and combined datacube as a new base for another registration [Bro08a].

The first advantage is the increased number of spectra that become available for a new application of PCA. When more spectral signal is used in a PCA, the components can be separated more accurately according to the variation present in the spectra. This will improve the contrast and separability between the extracted features. This is shown in Figure 4.9, which compares PCA applied to the mosaic dataset with PCA applied to a single tile. Both positive and negative component images are primarily influenced by the peaks on m/z 369 and 365. The signs of the components correlated with the different parts of the tissue section are different for the mosaic dataset and the single tile. This phenomenon is caused by PCA and depends on the variation in the dataset. Both spectral components from the mosaic dataset and the single tile are matched, after which the component images are compared in Figure 4.9. Although the signs are inverted, both components are still anti-correlated in both datasets.

The second advantage of using the combined result for a new registration is the larger spatial area that is used to create the MSE landscape. Spatial regions on the edge of a single component image are extended by the registered and added component image. As a consequence, the MSE landscape gains accuracy on the former edges that could have caused problems in finding a well-defined minimum. A MSE landscape of two combined datacubes is shown in Figure 4.10, which is used in the registration of a third tile from the same sample.

4.4 Discussion and Future work

The presented automatic alignment routine is suitable for highly multidimensional datasets, which are sparse in any single channel and possess a significant degree of uncorrelated noise. The results show that the PCA-based approach for the registration of a collection of spectral datacubes is superior to a traditional TIC-based method. Pixel-based registration of selected image components using the MSE metric with a complete coverage of the search space and an additional entropy weighting is able to correctly register two datacubes if a solution exists. We can find a more robust solution with the multiple minima from the landscapes of several extracted features instead of using only one pair of images. Some remarks can be made about the metrics used in this method with the possibility to improve and/or optimize the algorithm. This method of aligning multiple tiles has several advantages:

- By using the sequentially aligned and combined data from the first two tiles, a following alignment step includes additional overlap information.
- The tiles can be ranked according to contrast, allowing those of higher contrast to be aligned first, thus maximizing the overlapping regions for those tiles of

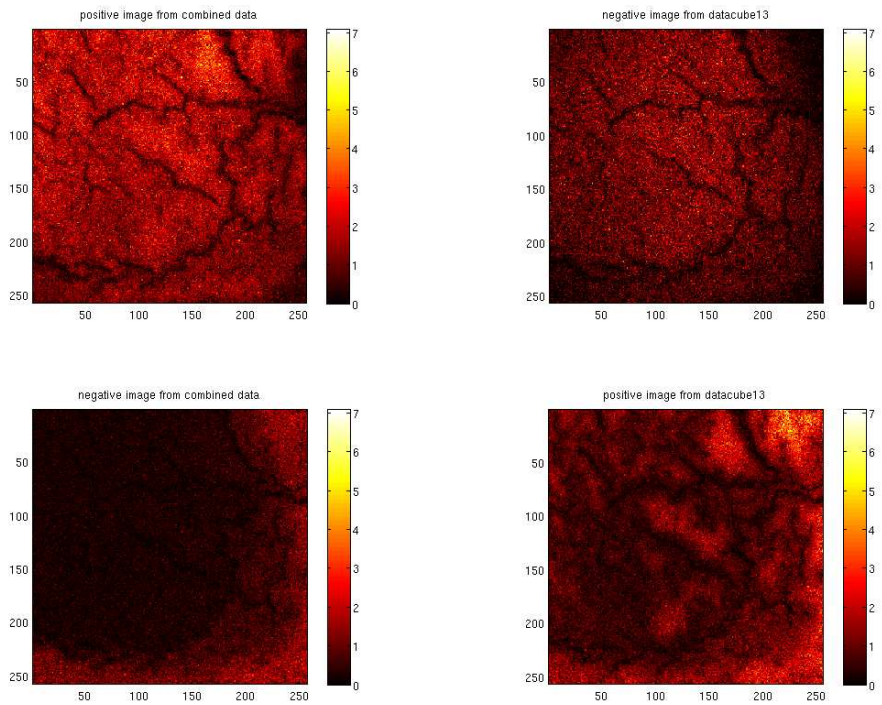


Figure 4.9: *Enhanced separability and contrast in components by PCA using a larger dataset. The positive and negative components on the left of the combined data show a better separation between peaks on m/z 369 and 365 compared with the similar components on the right from a single datacube. Note that the signs of the components are different of the mosaic dataset and the single tile.*

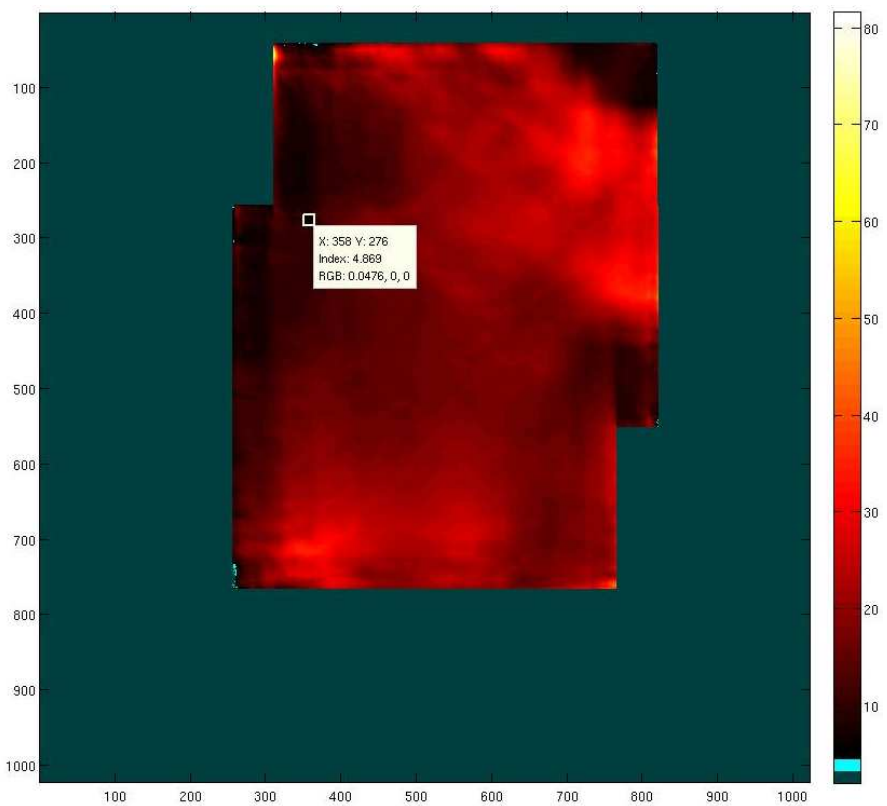


Figure 4.10: *MSE landscape for a registration using a more accurate, intermediate result of two combined datacubes.*

lower contrast.

- This method is time-saving because the process is executed in parallel using a 'divide and conquer' approach.

Discussion

This procedure was developed, because existing image alignment routines, such as those used in photography, were found to be unsuitable for the imaging mass spectrometry datasets. There are several reasons. In most mass spectral measurements, there are no sharply defined edges/corners that can be used as 'landmarks'. Landmark points are identifiable points [Dry98] used to align two images. The images created of particle counts by the mass spectrometer do not contain sharply defined shapes that can be robustly landmarked. Another problem with mass spectrometry data is that the overlapping regions between tiles are small. There are very few distinct features present in the overlapping regions, therefore it is hard to detect similarities. Even if these regions were larger, the individual spectral images contain less correlation between groups of neighboring pixels compared to photographic images due to counting statistics. This results in spectral images with a low signal-to-noise ratio that complicates feature detection. Areas with low signal, or no signal at all, are excluded in this approach by a weighting factor on the regions that are compared. In image alignment routines for photography, it is usually unnecessary to implement a weighting factor on regions in order to create a more robust alignment. Another issue that complicates the alignment of datacubes is that each datacube can contain thousands of distinct images; photography normally compares only a few single images.

Only one solution in the kneecap dataset from Figure 4.6b was marked as a 'correct fit', but incorrectly found by the algorithm. This was mainly caused because the overlapping regions with the found minimum did not contain enough intensity information. Unfortunately, the entropy did not contribute enough as a weight to eliminate it as a solution in the $\frac{MSE}{entropy}$ search space. One mistake out of ten combinations is not a serious problem when the complete mosaic has to be constructed. With the joint contributions of all correct solutions, it is not hard to find the correct location of the datacube in the mosaic. Besides the advantage of directly registering the important features with this PCA-based approach, more search spaces are available in one combination of datacubes. In those cases where no match could be found between the spectral components of two datacubes, or where the correct solution could not be found, the algorithm still does not produce an ambiguous solution, because it is able to use more search spaces. Generally, it does not have to be a problem if some of the offsets can not be found, if there is enough certainty about the ones that are found. With the TIC-based approach, there is no alternative but to use the global minimum in the $\frac{MSE}{entropy}$ search space. The MSE metric was not able to find an unambiguous offset to register the TIC images, while they can be found with the PCA-based approach. So even if some of the datacubes do not have much overlap or useful spatial features in the overlapping regions, it is still possible to create a complete mosaic when the entire $\frac{MSE}{entropy}$ search space is used.

The automatic alignment routine addresses the problem of the sample stage position being of greater uncertainty than the spatial resolution of the SIMS measurement.

There are other issues associated with mosaic imaging that the analyst should arm himself against in order to ensure the validity of mosaic measurements:

1. Sample height variation, across a large sample, would result in a systematic variation of compounds measured m/z mass across the sample and leads to ionization artifacts. Any such variation (including a localized protrusion) can be readily identified by calculating the variation of an ion's mass with position, a so-called 'height map' [McD03]. This height map can be used to correct the mass measurements but can not remove any ionization artifacts. No such variation was found with the samples in these results.

2. Chemical damage of biologic samples: the dose of the primary ion beam is a critical factor; if it is too high the chemical integrity of the sample is compromised and the SIMS spectra are no longer representative of the sample. A mosaic image requires overlap regions that will receive two-fold the normal primary ion dose. If this dose is too high it could cause the overlapping regions to display different spectral signatures and potentially skew the alignment procedure. The ion dose delivered to the overlapping areas was sufficiently low that the only effects observed were a slightly lower signal intensity in the overlapping areas. These lower intensity areas were sufficiently different to be distinguished by PCA.

Future work

Clearly, the results of the PCA step and the determination of the quality of the signal are essential elements in the success of the automated alignment algorithm. The results obtained with PCA can be highly affected by preprocessing the data, which includes denoising, selection of peaks, and even the choice of scales. For example, in SIMS the signal intensities decrease rapidly with increasing mass-to-charge ratio; the use of a logarithmic intensity scale can be used to give more weight to the higher mass, but lower intensity, molecular ions. For the automated alignment routine preprocessing was limited to binning. Previous work on SIMS data has demonstrated that binning is "the most effective technique to improve PCA performance" [Wic03]. The auto-alignment procedure benefits from PCA in noise reduction and the availability of more than one component image for a more robust alignment. Consequently, we used fast PCA methods rather than more computationally intensive variants for these large datasets.

The algorithm could be tuned by changing the registration metric using for instance reciprocal square differences, gradient difference or different implementations of mutual information. Some—combinations of—other image characteristics acting as a weight like a gradient-based metric or maybe a variogram-based approach may create some improvement in certain cases. These are more computationally intensive than the currently used metric. Several standard metrics on image texture properties were investigated, including contrast, correlation, homogeneity, energy, and entropy. It was found that the approach of local entropy, used in all the results, was the most effective at removing regions with a high randomness in intensity values.

For accuracy, it is desirable to use each combination of datacubes by the creation of the mosaic. The use of many complex metrics may improve the results, they

slow down the process significantly when using larger collections. The next step for solving this problem of creating a mosaic is the applications of the algorithm on larger collections and investigate if some optimizations are needed. Image pyramids are commonly used to reduce the search space and to reduce the computations. If the binned 128×128 image is used instead of a 256×256 image, the number of entropy calculations and the comparisons to create the MSE-landscapes are both decreased by a factor of four. Once the correct offset is found in a low resolution image, this offset could be refined using the same region in a higher resolution image or by sub-pixel interpolation.

An interactive variant of the presented registration algorithm will improve the robustness of the found solutions. An analyst could improve the results by reducing the search-space when the overlapping regions can be selected manually. This forces to locate a minimum in a more confined search-space. Similarly, if estimations can be made about the relative positions of the measurements within a raster, it reduces the number of combinations that have to be considered for the solution. Computational efforts for the interactive selection can be reduced when selection takes place on a lower resolution made possible by, for instance, the zooming technique described in Chapter 6.

4.5 Summary and conclusion

This chapter explains how the reduction of uncorrelated noise provided by PCA allows high-resolution imaging mass spectrometry datasets to be automatically aligned and combined for high-resolution analysis of large areas. The generation of mosaic images of large datasets necessitates stitching together a collection of separate imaging experiments. One advantage of feature-based registration is that the influence of noise in a datacube is greatly reduced and, therefore, will result in a more robust registration. Another advantage is that multiple attempts for registration are performed with several extracted features to improve robustness.

The three steps of PCA decomposition, spectral matching, and signal quality assurance are necessary because of the high dimensionality and sparsity of the SIMS imaging mass spectrometry data and indicate future methods of how to work with such data. The results clearly show that the entropy-weighted, mean squared error landscape of chemically matched component images can be used to automatically align high-resolution imaging mass spectrometry datasets. This algorithm can be adapted for all datasets of similar nature in imaging mass spectrometry, particularly the mass microscope being developed as part of the high-resolution imaging mass spectrometry research efforts.

Chapter 5

Feature visualization

So far, we have shown how PCA is used to extract features from imaging spectrometry data. A feature-based method was introduced for the automatic registration of spectral datacubes. Registered datacubes can be combined to create a new dataset that covers a larger spatial region. PCA can then be used to extract features from the combined dataset as well.

This chapter presents a new application of PCA: to generate multidimensional transfer functions for the visualization of spectral datacubes. These transfer functions are needed in the volumetric visualization of spectral data to isolate those regions containing interesting features. This approach is characterized by the direct linkage between the resulting spectral and spatial components of a feature. Our method enables us to create an opacity map from these components. One or more mappings can be selected to highlight features in 3D using volume visualization.

5.1 Introduction

The use of Direct Volume Rendering (DVR) is a well-known method for the visualization of 3D volumetric datasets. In most volumetric datasets, each voxel contains a scalar value that represents the density of a material on that location. For visualization, a transfer function is a mapping that assigns a color and opacity value to a scalar value. A volume renderer can draw the voxel data using the mappings specified in the transfer function. The challenge in designing an appropriate transfer function is identifying which structural properties are important for the user and which relevant features in the data should be highlighted.

Imaging spectrometry can be used to scan the structure of chemical elements on material surfaces. In contrast to a volume consisting of 3D points of scalar values, a spectral dataset consists of two spatial dimensions and a wavelength in the third dimension. Each scalar value in the volume is interpreted as the intensity on a wavelength at a 2D position on the surface of a material. Linsen [Lin05] stated objectives for a visual exploration tool for mass spectrometry data: a better understanding of the data set in its entirety, quantitative depiction of expression ratios on a global scale,

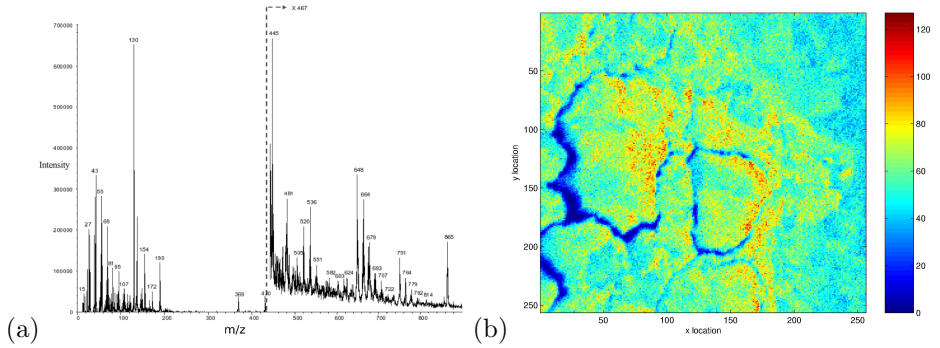


Figure 5.1: *(a) A plot of the summation of all spectra in the datacube from the cerebral ganglia of the pond snail. The right part of the plot is enlarged by a factor of* 467*. (b) The spatial distribution of spectral profiles.*

and easy visual detection of data acquisition errors. These objectives also apply to the design of visualization tools for other mass spectrometry datasets.

Since molecular compounds have a unique and known spectral profile, scientists can use spectrometry to investigate which chemical elements are present on the surface of a material, provided their spectral profile can be extracted. Unfortunately, extracting a spectral profile from a datacube is a difficult task. First, the intensity at each point in the volume consists of contributions of the spectral properties of neighboring chemical elements at that position on the surface. A robust extraction method is needed to factor the linear combination in a spectrum into the spectral profiles of each molecular compound. Second, spectra characterize themselves by different levels of scale in which peaks in the spectral profile can vary in order of magnitude. For example, consider Figure 5.1a. The sum of all spectral profiles in the datacube is plotted, with the mass-to-charge ratio on the x-axis and the measured intensity on the y-axis. For visualization purposes, the right part of the plot is magnified by a factor of 467. Various large peaks can be seen in the left part of the spectrum, while the right part of the spectrum consists mainly of very small peaks. Both types of peaks are important in the analysis of the data. Figure 5.1b shows the spatial distribution of spectral peaks. The value of a pixel represents the sum of spectral intensities at each position on the surface of the material. A color map is used to assign a color to each intensity.

Figure 5.1 is an example of how scientists use two side-by-side views to analyze data in a datacube. One view is the spectral view: it shows the sum of all spectral profiles. The second view is a spatial view: it shows the summation of the spectral profile at each position on the surface of the scanned material. It is our goal to create tools for data analysis with one integrated 3D-view to gain insight into the spatial distribution of features in the volume. This is difficult using only the above mentioned spectral and spatial views. This chapter is a first step towards achieving this goal: we demonstrate how a transfer function can automatically be generated for 3D-rendering of the datacube using PCA. Different automatically extracted components are presented in 3D visualizations to a user. The spatial and spectral characteristics of a component are displayed in one single representation. This representation provides

more information on the spectral profile and spatial distribution of a feature in one component.

5.2 Related work

Early implementations of volume rendering [Lev88] using transfer functions [Dre88] are mostly applied to data resulting from Computed Tomography (CT) scans. Many other areas in science could benefit from the techniques developed for these medical applications. For instance, Djurcilov et al. [Dju02] use techniques for visualizing 3D scalar datasets by combining uncertainty information on top of environmental data. Uncertainty information is added to the classic volume rendering equation to highlight important features by adjusting opacity and color. In these datasets each data-point has a 3D spatial location, contrary to spectral data where each data-point has a 2D spatial location and a certain spectral channel (for instance mass-to-charge ratio or wavelength). Therefore, there are few implementations of 3D visualizations for spectral data. One example can be found in Polder and van der Heijden [Pol01]: it shows a DVR of a spectral datacube and a representation using iso-surfaces. The latter produced some unexpectedly good results. Torson [Tor89]presents a system for interactive analysis of 3D data-arrays. He uses this system on spectral data on which conventional volume rendering and surface display techniques could not be used appropriately. Torson names three reasons for not applying DVR techniques on data from imaging spectroscopy. First, the data values are not varying smoothly throughout the datacube. Second, volume rendering can not easily show small local data variations superimposed on broad overall variations. Third, volume rendering provides only a qualitative view of the data. Torson's system did not provide any interactive navigation tools for volume rendering in a PC-based virtual reality, unlike Fuhrmann et al. [Fuh02], who implemented interactive navigation tools on CT data. This would be the goal for this procedure, though.

We try to solve the problem of using DVR for spectrometry data using PCA (see Wall et al. [Wal03]) to detect features. Lasch et al. [Las98] already used PCA to detect patterns in FT-IR data images. Also, Piwowar et al. [Piw01], applied PCA to recognize spatial-temporal patterns in Arctic sea ice concentrations. A closely related multi-variate image analysis algorithm is ICA. Muraki et al. [Mur00] apply ICA on multichannel volume data from Magnetic Resonance Imaging (MRI) scans to separate specific tissue characteristics, e. g., water and fat. Muraki trains a radial basis function network with sample data from the visible female dataset to generate color transfer functions.

He et al. [He96] also use stochastic search techniques to generate transfer functions for data from MRI and CT scans, with better results than the approaches relying merely on the 'trial and error' of the human factor. He's approach requires a minimum of computer aid compared to data-centric or image-centric approaches as described in Pfister et al. [Pfi01]. Due to the complexity of the task and the introduction of multi-dimensional transfer functions [Kin98; Kni01b], most research tends towards a semi-automatic approach in transfer function design for direct volume rendering of medical datasets. A minimum of user involvement is accomplished using direct manipulation widgets (see Kniss et al. [Kni01a]) to create multi-dimensional transfer functions for specific datasets [VH04].

Different approaches in creating appropriate transfer functions have to be considered in the relatively open area of using DVR to visualize the datacubes from imaging spectrometry. Existing multi-dimensional transfer functions do not handle the equally important high and low peaks in the spectral dimension very well. PCA has already proven itself in the area of statistical pattern recognition.

5.3 Method

In our method, PCA is used to extract features and create an appropriate transfer function for a volumetric visualization of the spectral datacubes. This transfer function maps the intensities in an extracted principal component to different opacity values. After PCA, a component is selected and the spectral datacube is visualized using the opacity map that is generated by the transfer function. This way, the linked spectral and spatial distribution of a feature in the datacube can be examined.

Weighted PCA

PCA has to be applied to all spectral imaging datasets (both TOF SIMS and FT-IR). As described in Subsection 3.2.4, PCA can be applied to TOF SIMS data without scaling of the spectral and/or spatial dimension. The FT-IR datacubes are not created by counting statistics, which makes mean-centering and scaling appropriate preprocessing steps. These preprocessing steps are to normalize the unfolded FT-IR spectroscopy data in matrix D, as in Subsection 3.2.3. First, we subtract each data value with the mean. This reduces the influence of extreme scalar data values. Second, data values are scaled according to the variance. This removes big variations between values. Both rows and columns of the data matrix D are preprocessed in this way

$$\begin{aligned} \tilde{D}_{rows} &= (D - \mu_{xy}) / \sigma_{xy} \\ D_{preprocessed} &= \left(\tilde{D}_{rows} - \mu_m\right) / \sigma_m \end{aligned} \tag{5.1}$$

where $\tilde{D}_{rows}$ is the data matrix D with mean μ_{xy} subtracted from each row, after which the row is scaled according to the variance σ_{xy}. The same operations (mean-centering and variance scaling) are applied to the columns, which will result in $D_{preprocessed}$. Consequently, we can treat both rows and columns in the matrix as measurements when applying PCA. Next, PCA is used to find orthogonal and normalized matrices for the spectral and spatial dimensions.

Transfer function generation

In Subsection 3.3.1 we described how the principal components and score vectors are computed in order to find spectral or spatial features in a 3D spectral volume. Opacity of the transfer function is applied for the visualization of these features. Multiple opacity functions are used for different features to isolate them from other areas. With the eigenimages and eigenspectra matrices (Y and P), the original data is projected onto new bases. To generate the opacity function of areas with the highest

variance, an addition of the first score vector with both the highest spatial and the highest spectral variance is used. The motivation is that those features are captured in one single 3D opacity map. A 3D opacity map is a datavolume in which each cell cell is assigned a certain opacity value, for instance by a transfer function.

This argument is represented more formally by

$$Y = \begin{bmatrix} i_1 \\ i_2 \\ \vdots \\ i_m \end{bmatrix} \quad P = \begin{bmatrix} s_1 \\ s_2 \\ \vdots \\ s_{xy} \end{bmatrix} \tag{5.2}$$

Here, Y is a $m \times xy$ size matrix with spatial scores and P is a $xy \times m$ matrix with the spectral loadings, which is converted to a $m \times xy$ matrix by taking the transpose. Hence, O^1 is a $m \times xy$ sized matrix. The resulting vectors are combined into one opacity transfer function. For example,

$$O^1 = \begin{bmatrix} i_1 \\ i_1 \\ \vdots \\ i_1 \end{bmatrix} + \begin{bmatrix} s_1 \\ s_1 \\ \vdots \\ s_1 \end{bmatrix}^T \tag{5.3}$$

shows how the opacity map of the first score vectors is derived. The 3D points with the highest positive and negative values in O^1 are assigned to high alpha values. All regions in the volumetric data that contribute to this first principal component are made opaque using this 3D transparency map. Similarly, the opacity maps of the consecutive score vectors can be generated. Different opacity mappings can be combined to display similarities or differences of multiple features in the original data.

5.4 Applications

Our method is applied to four examples of spectral recordings. First, the results of our method are shown when applied to a dataset created by the TOF SIMS technique. It is a dataset of a small section of the anterior lobe of the cerebral ganglia of the pond snail, *Lymnaea Stagnalis*. The second example is a visualization of the brain ventricle of a mouse resulting data from FT-IR spectroscopy. The third and fourth example is respectively an embedded hair cross-section and a synthetic sample as a droplet-array as seen in Subsection 3.4.3.

The small brains of a snail

A high spectral and spatial resolution can be obtained using TOF SIMS. In this example, atomic and molecular structures in a dataset can be identified. The boundaries of different cells can be visualized when TOF SIMS is applied to a slice of the brain of the pond snail. Figure 5.1 already showed the resulting spectra and images from this spectral scan. Some obvious features are highlighted when our method is

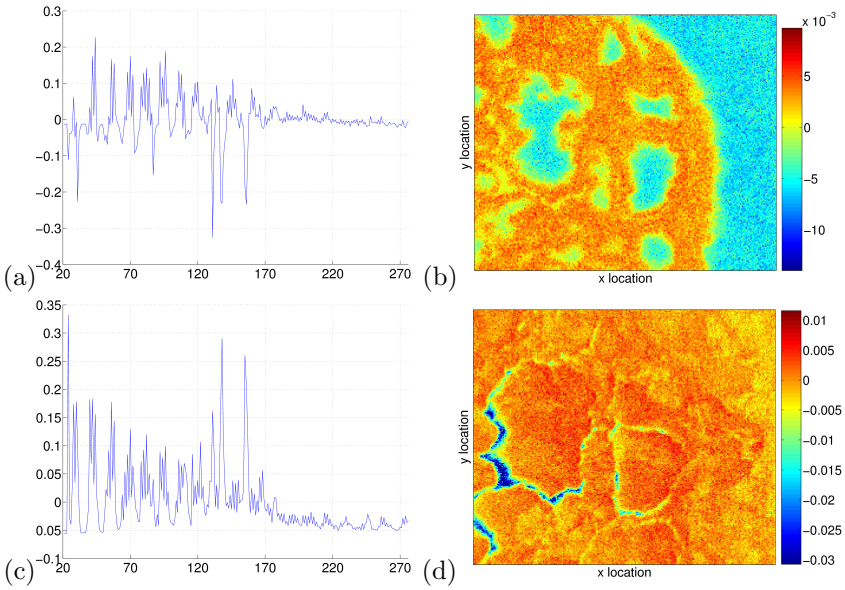


Figure 5.2: *(a–c) The first two score vectors with (b–d) the accompanying eigenimages as a result from PCA.*

applied to this dataset. The first two spectral score vectors are shown in diagram a and c of Figure 5.2.

The three lowest peaks in (a) represent the negative contributions in the spectral dimension of the component that isolates the largest amount of data. Figure 5.2b gives the corresponding spatial distribution of the material in which the cells in the dataset are embedded. The second score vector in (c) has isolated a large positive peak. This peak corresponds with the red areas in the image of Figure 5.2d that can be identified as larger cavities between the different cells. Other components that result from PCA also highlight certain areas within or between cells that contain different organic compounds.

Transfer functions can be made from the resulting components of the spectral datacubes described in Subsection 3.4.2. The principal components extracted in Figure 3.3 result in a 3D representation shown in Figure 5.3. The spectral datacube in Figure 5.3a shows that many different spectral features display a certain amount of spatial correlation. This makes it difficult to identify the individual features from these two principal components. Figure 3.3b shows similar spectral features in its components after PCA with an additional VARIMAX rotation. The features in this datacube are more clearly visible by the improved spatial correlation. A better feature contrast is found, but the individual components are not fully separated. Figure 3.3c and d both show similar features resulting from PARAFAC. These datacubes show improved distinction between the spectral planes of the feature, but PARAFAC is also able to separate the feature from (b) in two individual features.

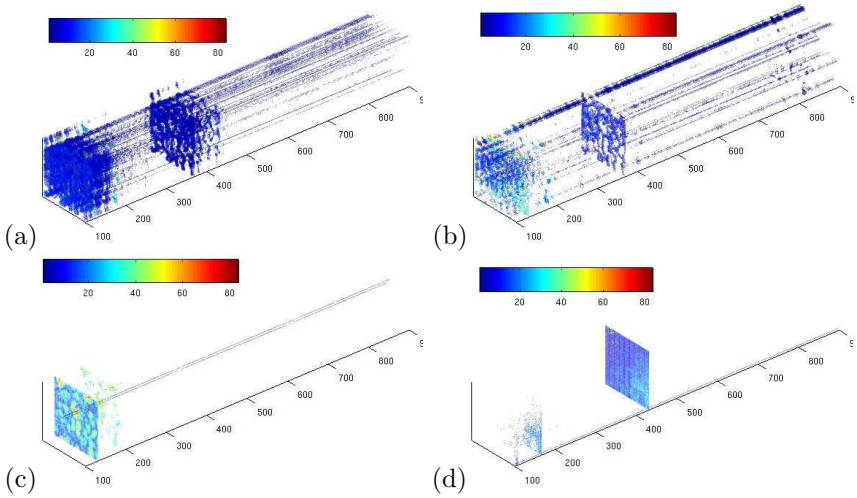


Figure 5.3: *The spectral datacube with a resulting transfer function from (a) PCA, (b) PCA with VARIMAX rotation and (c–d) two different components from PARAFAC.*

The bigger brains of a mouse

The second example is a slice of the central brain ventricle of a mouse. FT-IR spectroscopy is applied to identify different chemical functional groups. The resulting spectral and spatial components are combined in different 3D maps of which the second, third and fourth are displayed in Figure 5.4.

The long axis represents the spectral dimension that ends in the front on wavelength of 4000 cm^{-1}. The first component mainly highlights the differences between the uninteresting regions (the red and blue areas). The red and orange regions of the components represent positive correlations, whereas the blue regions represent negative correlations. Both positive and negative regions could be of interest for the identification of the functional groups. In Figure 5.4a, the isolated blue region at wavelength 1550 cm^{-1} represents the location of amide groups. The blue regions at wavelength 3300 cm^{-1} and 3400 cm^{-1} represent hydroxy groups and amino groups. Different regions are clearly distinguished in the resulting 3D maps. These maps can be used as an opacity map on top of the original data as shown in Figure 5.5.

The same volumetric data can be loaded in VolView, a visualization package from Kitware. This package offers many tools [Mar05] to interactively create an appropriate transfer function based on the 'trial and error' method with an initial estimation of color and opacity transfer function based on the histogram of the data. This initial guess for an appropriate transfer function is shown in Figure 5.5c. This package can not distinguish between the spatial and spectral dimensions in the spectral data, because it uses the scalar values the same way in all three dimensions. VolView can not differentiate between small but important differences between values in the spectral dimension when they are dominated by large peaks that are present in other regions in the datacube.

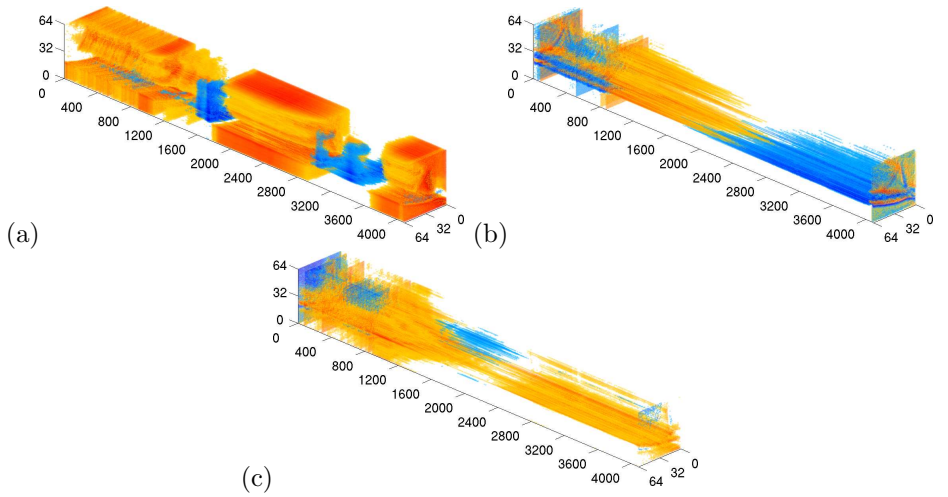


Figure 5.4: *The (a) second, (b) third and (c) fourth component mappings with the negative contributions in blue and positive contributions in red.*

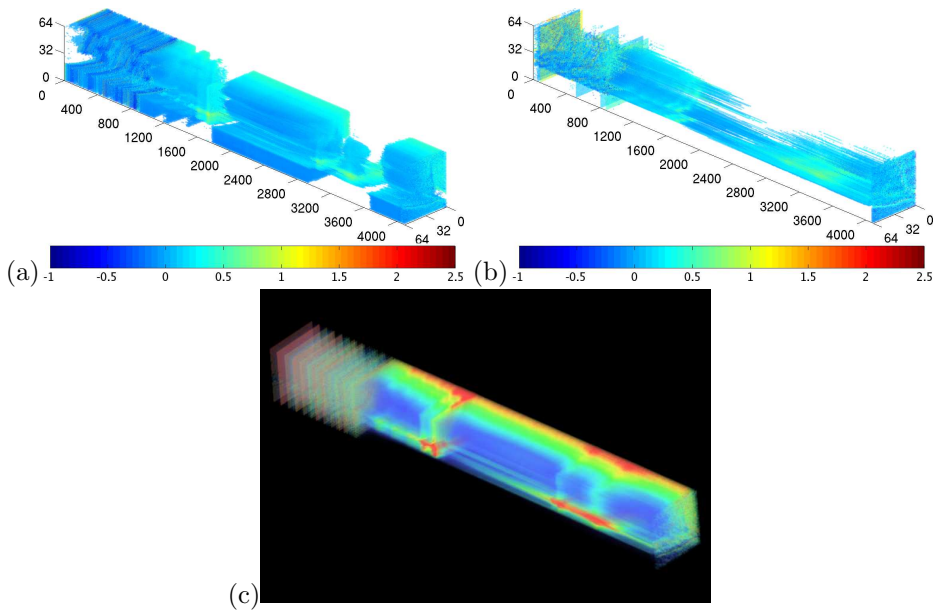


Figure 5.5: *(a–b) Two resulting component mappings applied to the original data compared with (c) the VolView representation of the same dataset.*

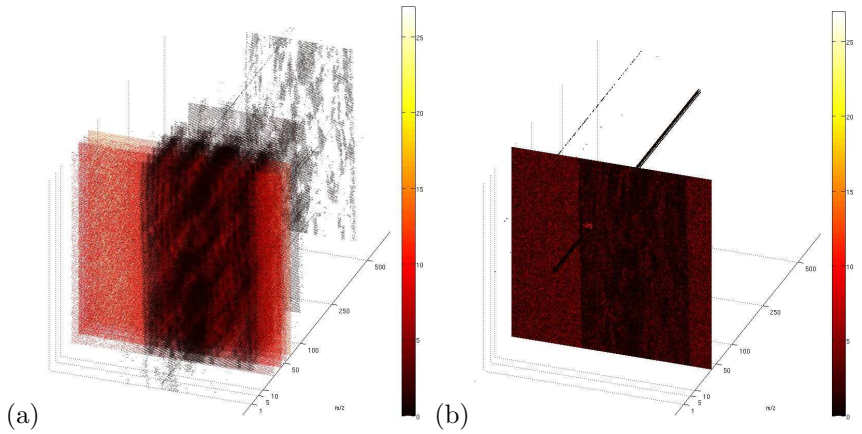


Figure 5.6: *A representation of the complete spectral datacube of the embedded hair with an overlay using the (a) first principal component, PC1, and (b) the thirteenth principal component.*

A hair and a droplet-array

The third example is a cross-section of a hair. PCA was able to extract the location of a particular feature in the hair, but the spectral view revealed that it was caused by a salt-crystal. A combined view would instantly reveal the connection between both views. Each pair of extracted scores and loadings can be combined in a single three-dimensional overview to gain more insight in the correlations between spectral profile and location. Each value in the cube is the intensity on a certain position in a spectral plane and is given a color using the 'hot'-color map from MatLab^{TM}. Because most data-points have value zero within a mass spectrometry dataset, the complete datacube would result in an image of a black box. Large parts of this box can be discarded as they do not contain any interesting properties. Again, an opacity map is generated to hide uninteresting features within the datacube which is created by the extracted components. Instead of a continuous switch between spectral and spatial view, a complete view of the cube can directly reveal this connection. A user is able to interactively rotate the cube and instantly get an overview of all the data in three dimensions.

The complete spectral datacube of the hair is shown in Figure 5.6a and b. Only the high values in the spectral profile and image component of the first principal component are made opaque by the opacity map. This highlights that component in the original datacube which contains mostly the areas and peaks from the hair itself. The component with the extracted features from the crystal is shown in Figure 5.6b. It clearly shows the relation between the highlighted image plane on $m/z\,39$ and the small group of pixels on the location of the crystal, while other areas of the datacube remain hidden. The significant peak on $m/z\,39$ in the spectral component highlights the complete image plane at this spectral position. Similarly, the high intensity of the pixels in the spatial component results in the appearance of a 'rod', spanning the whole spectral dimension of the datacube. The number of points that are shown can be adjusted by changing the threshold in the opacity map. This

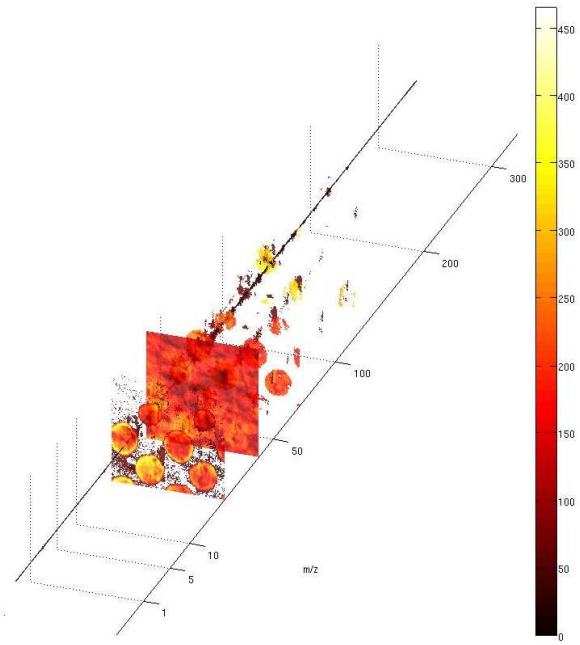


Figure 5.7: *A representation of the complete spectral datacube of the PVP droplet series with the 6th PARAFAC factor.*

representation provides better overall insight in the data by visualizing the direct correlation between spectral peaks and spatial occurrences. Figure 5.7 shows several isolated drops in the spectral datacube using the sixth PARAFAC factor. The different components or factors can be highlighted together or separately in the same datacube by combining their opacity maps. The resulting three-dimensional view becomes more accurate and discriminating once the resulting components contain more contrast. This advantage makes it easier to compare the quality of results from the different multivariate analyses.

5.5 Discussion

The current practice in the analysis of spectral data is illustrated in Figure 5.1. Two plots, one for the spectral and one for the spatial information, are used to view the spectral datacube. In this chapter, we have introduced a method to view spectral data in three dimensions. The 3D view is used to gain insight into the spatial distribution of features in the volume, which is very difficult to do using only the spectral and spatial views. We have discussed how a transfer function can be generated using PCA. It is applied in both the spectral and spatial dimension of the individual images, as well as in the dimensions of the spectra of all images.

Using PCA in both dimensions of the datacube allows us to address the two major problems that were mentioned in the introduction. First, intensity at each point in the data consists of contributions of many spectra. Applying PCA in the spectral

dimension results in the principal components of a number of spectra. Second, spectra characterize themselves by different levels of scale. PCA is a technique that computes variances. Hence, the technique will find variances between spectra with very large and very low peaks.

We now discuss the advantages and disadvantages of the proposed method:

- Although the first principal component is used to highlight the data regions which have the highest variances, this does not necessarily mean the most interesting feature is captured by the first component. For example, a principal component could contain high variances in the spectral dimension and very low variances in the spatial dimension. As a consequence, each opacity function should still be manually checked by the user.

- Contrary to the original data, component vectors can contain negative values. These negative peaks may be just as important as the positive, but it is uncertain how the composite opacity map is affected when the peaks have a positive spectral but negative spatial contribution.

- The method normalizes the data as much as possible and reduce variances in both dimensions (see Equation 5.1). This is difficult to realize in both dimensions without losing the direct relation between spectral and spatial components. In some cases, it is not possible to filter outliers completely in the preprocessing steps: their values are too deviant to successfully auto-scale the data and lose the extreme values. An alternative data-scaling technique can be used to remove these outliers automatically through a threshold function to filter extreme values.

- As mentioned above, PCA is used to analyze spectra and highlight contrasting features in images. Even small differences between spectra and images can be detected when they are correctly preprocessed. Hence, the method can also be used to detect noise in both spectral and spatial dimensions. The 'least' principal components will contain most of the noise present in the data.

- The main advantage of using our method to detect spectral and spatial features is the direct linkage between these dimensions by using the result in one dimension to calculate the other (see Equation 5.3). It is not possible to link components automatically by using separate analyzes and treating the spectra or images as separate dimensions.

Future work

Our method is the first step in creating a tool for the analysis of spectral datacubes using direct volume rendering. In future, 3D separation and clustering algorithms can be incorporated to improve the definition of the opacity function. Another improvement would be the automatic identification of features by adding additional information about spectral peaks. Component vectors could be matched using this database of score vectors to label the different volumetric regions in the visualization. Storing datasets, features and classifications could eventually evolve into an integrated system for feature recognition and analysis. Finally, the method can be

applied to other multidimensional scientific datasets which lend themselves for finding high dimensional patterns.

5.6 Summary and conclusion

This chapter shows how principal components can be used to create multi-dimensional transfer functions. Two types of spectral datacubes are visualized in 3D by direct volume rendering with these transfer functions to control opacity and highlight extracted features. This enables us to visualize the link between the spectral and spatial characteristics of a feature within the spectral datacube. The resulting volume can be viewed from different perspectives. Moreover, different features are highlighted, thereby providing a complete overview of the data. Consequently, selection and zooming operations can be applied to create partial views.

Chapter 6

Feature zooming

Chapter 5 described how PCA creates multi-dimensional transfer functions to highlight features in a 3D representation of spectral datacubes. These highlighted parts of a spectral datacube show the direct linkage between spectral and spatial properties of an extracted feature. Spectral datacubes resulting from imaging mass spectrometry contain too many variables to be displayed entirely in one direct volume rendering.

In this chapter, we present a zooming technique based on PCA to select regions in a datacube for enhanced feature extraction at the highest possible resolution. It enables us to select spectral and spatial regions at a low resolution and recursively apply PCA to zoom in on interesting, correlated features. The technique utilizes a higher signal-to-noise ratio in the data, without losing the high resolution characteristics. Less interesting and/or dominating features can be excluded in the spectral and spatial dimension. For these reasons, more features can be distinguished, and in greater detail as well. Analysts can zoom in on a feature of interest by increasing the resolution.

6.1 Introduction

Traditionally, analysts look at the sum of all spectral variables in a datacube to determine the presence of different chemical compounds. The peaks in intensity values within the spectral dimension, as shown in Figure 6.1a, are of particular interest. The m/z on the x-axis of the figure is the 'mass-to-charge ratio'. A knowledgeable mass spectrometrist can determine which ion corresponds with a certain spectral value. More recently, analysts also look at the sum of all spatial variables, or 'spectral image planes', as shown in Figure 6.1b. This example shows the spatial distribution of the summed mass spectra in a cross-section of a chicken embryo.

Figure 6.2 is an example of an extracted component image with the traditional application of PCA. Depending on PCA for the extraction of features has several weaknesses. A common weakness is the noise within the data. Noise occurs for various reasons. For example, counting statistics in the image detector rely on a small number of incident particles. Also, the ion source or detector can display instability. This noise remains present in the extracted components, thereby having a negative

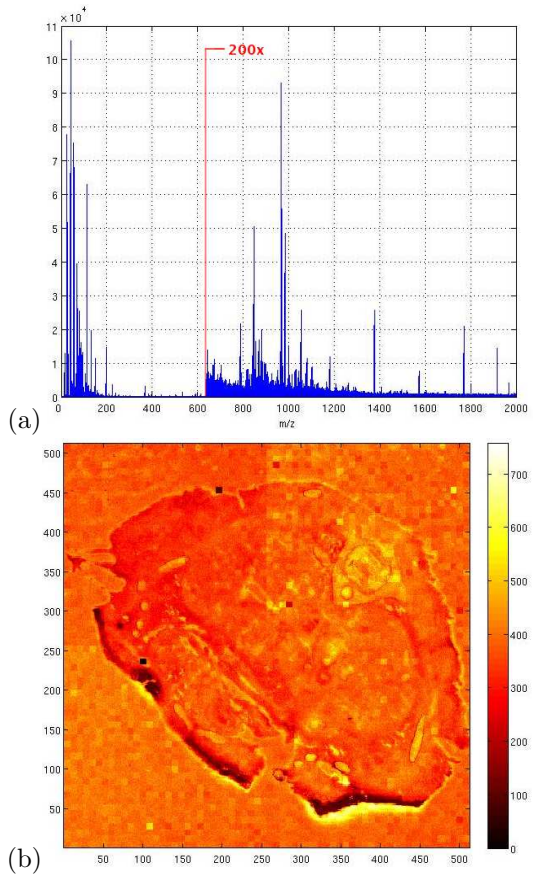


Figure 6.1: *(a) The sum of all* 512^2 *spectral variables with the side on the right zoomed* 200 *times and (b) the matching sum of all* $2 \cdot 10^6$ *image planes of a cross-section of a chicken embryo.*

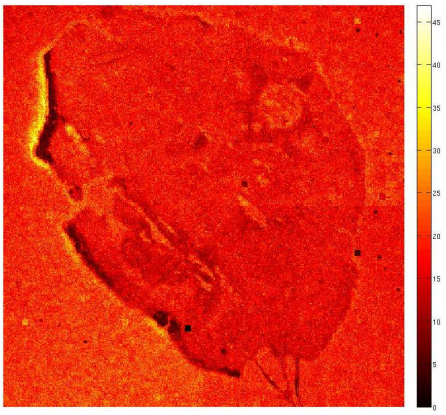


Figure 6.2: *Image component extracted traditionally with a dominating area in the top left corner and with a relatively low contrast.*

effect on peak separation. Methods for feature extraction usually try to reduce the effects of noise by applying complex filtering models in order to increase the signal-to-noise ratio. Due to the size of the imaging mass spectrometry data, filtering is not considered in this approach, because of memory and time restrictions. A second weakness of PCA is that it extracts features according to covariance. If there is an artifact in the data with a relative high variety of intensity values or a number of peaks with a large variance, it would dominate in the extracted components. Finally, the spectral dimension has to be reduced to be able to use PCA-based feature extraction. As a result, any high resolution characteristics in the spectral dimension are lost.

This study aims to improve feature detection as well as the level of detail of the results by a zooming strategy based on PCA as a multi-scale approach. Our recursive strategy consists of three steps. In the first step, the raw data is compressed to an appropriate size. In the second step, PCA is applied to extract features automatically. Interesting feature components are selected in the third step. The associated spectral and spatial variables are used to construct a new dataset at a higher resolution using the original data. Again, as in the first step, PCA is applied to the selected data until an interesting feature is isolated at the highest possible level of detail.

The proposed strategy exploits a basic data reduction technique. The number of variables is reduced by summing the intensity values of a consecutive number of variables. At first, peak information is neglected in the analysis. Eventually though, this strategy is still able to extract the high-resolution characteristics that are lost in the traditional application of PCA. The main difference with the traditional approach is that a larger bin size must be chosen in order to fully exploit the capacity to zoom and create more detail. Larger bins increase the signal-to-noise ratio in each variable. So binning can be applied to reduce the noise without the use of complex filters and data-dependent noise reduction techniques. At the same time, it compresses a datacube to a size that is computationally less restrictive to apply a feature extraction technique to. A final strength of our 'divide and conquer' approach is that by zooming in on interesting features, dominating artifacts can be left out in the newly constructed datasets. This increases the contrast in the resulting components after reapplying PCA.

In previous work [Bro05a; Bro07], we visualized extracted correlated features in 3D by combining spectral information with spatial locations. These features are parametrically visualized at the highest resolution possible, but they have to be identified first. With this new technique, we are able to spectrally and spatially zoom in on specific regions of the datacube and extract a larger amount of features with more precision and more detail. Therefore, the final visualizations will benefit, because they depend on the feature extraction step prior to the visualization step. Even with a technique to visualize mass spectrometry data at the highest resolution possible, the selectivity and quality of the feature extraction is still key.

6.2 Related work

The classical method for feature discovery in mass spectrometry data is by manually locating peaks in the mass spectrum. In imaging mass spectrometry, the ions that are removed from a surface are counted. Their spatial location is stored, together with the time it took to arrive in a detector. This time-of-flight is used as a spectral

variable or raw channel. Additional databases with known peak locations in mass spectra can be consulted to identify a selected peak. Often, only a small part of the data is selected for detailed analysis. Most of the data remains unutilized in the analysis. The spatial distribution of a peak is traditionally visualized by summing all intensity values in the spectral dimension to create an image of summed intensity values.

Univariate analysis [McC05] can be applied to imaging spectrometry data to image the distribution of ion counts in a small spectral window. PCA is attractive for our purposes, since it is fast, does not use complex models and is easily scalable to large inputs. In order to be able to apply PCA on mass spectrometry data, the number of spectral variables has to be reduced. A common reduction step is the summation of multiple spectral variables into one spectral bin [Pac04]. Binning is done, because:

1. rather than a feature being distributed among multiple spectral variables, it is now combined and can therefore be treated as a single variable in PCA;

2. it reduces the size of datacubes enough to be able to apply PCA without memory restrictions;

3. it improves the low signal-to-noise ratio in a datacube.

There is a balance between the number of spectral variables combined into one bin and the resolution of the features. If PCA is used at the highest resolution with a small bin size, fewer features can be distinguished. Once the bin-size increases, more features with higher contrast can be extracted. Unfortunately, an analyst is unable to separate the different spectral peaks combined into one bin. With a low signal-to-noise ratio, PCA is not able to separate those features with small variances from the many dominant peaks with large variances.

To overcome the problem of size and influence of noise, many researchers have developed advanced multi-scale compression techniques that use deconvolution filters in combination with wavelets [Sta98; Wol97]. Wickes et al. [Wic03] compared three spatial denoising algorithms on their performance with PCA. Claiming down-binning is the most effective technique compared to boxcar and wavelet filtering, this study still faces the same problem of initial spectral binning in order to apply the wavelet transform. After the transform, it is no longer possible to use the traditional PCA for feature extraction. For these reasons, we used the simple binning strategy to compress the data.

A new problem rises when several mass spectrometry datacubes are registered and combined into one new datacube [Bro06]. Because of the increased size, PCA can not be applied to this new dataset in the traditional way without reducing the spatial dimensions. It is useful to apply PCA to an enlarged datacube, because it includes more spectral variance in the analysis. Besides spectral binning, we also bin the data spatially in our approach and therefore are able to handle the combined datacubes. Although binning results in a decrease in detail, it has turned out to be very effective in increasing image contrast, especially for images with highly sparse features [Tyl03].

PCA is already hierarchically applied to different 3D datasets that contain point-clouds [Fra06; Kal05]. In these studies, PCA is used to replace a group of 'real' 3D points with an ellipsoid, thereby reducing the dataset. PCA is also used for data reduction in the field of neural networks in the so-called 'PCA-pyramids' [Wei96].

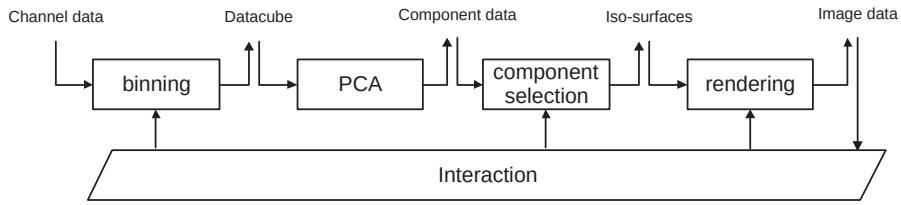


Figure 6.3: *The data-flow in visualization pipeline of the approach for zooming.*

This study can be compared to the well-known image pyramids, but it uses PCA as a reduction function in a neural network. These applications of PCA are intended as data reduction strategies instead of automatic feature selection schemes.

Other methods for noise reduction are mainly based on spectral or spatial filtering. According to Randolph [Ran05a], there is no optimal technique to filter mass spectrometry data due to the unknown and data-specific signal-to-noise ratio. Even if a point spread function could be estimated to apply a spatial deconvolution filter [Hut96], it is impossible to apply this filter to the millions of spectral planes in a mass spectral dataset in order to improve feature extraction. Keenan et al. [Kee04a] proposed a weighted variant of PCA to account for the Poisson noise present in mass spectrometry data. Although this study shows that the contrast in the results improved, it basically created a filter according to a model of the noise in the data in order to add a weight to the covariance matrix of the PCA. Other filtering methods like adaptive filtering techniques or simple Gaussian deconvolution on the individual image planes could also improve the results. However, we would like to emphasize the application of the widely accepted step of binning mass spectrometry data. Although most case studies ignore the selection of the appropriate size of the bins, the results mainly depend on this step. By implementing a variable bin size, and therefore different levels of detail, an analyst can select the most suitable level of detail for each component during the component exploration.

6.3 Method

An overview of the visualization pipeline in this approach for zooming is shown in Figure 6.3. The first step in the proposed method is the spectral and—optional—spatial binning of the raw channel data resulting of a measurement. Binning results in a multi-spectral datacube on which feature selection is applied by PCA or by a comparable decomposition method. PCA creates multiple components of which the resulting spatial distributions of the different extracted components are visualized in an overview. One or more components in the overview can be selected or excluded by an analyst. A new dataset is created with the contributing spectral and spatial bins of the selected components from the original data. After re-binning the new dataset with a larger bin size, PCA is reapplied. When a feature is isolated successfully, it can be visualized at the highest resolution.

The algorithm for zooming proceeds as follows:

1. Create a reduced spectral datacube on a low resolution by the spectral or spatial

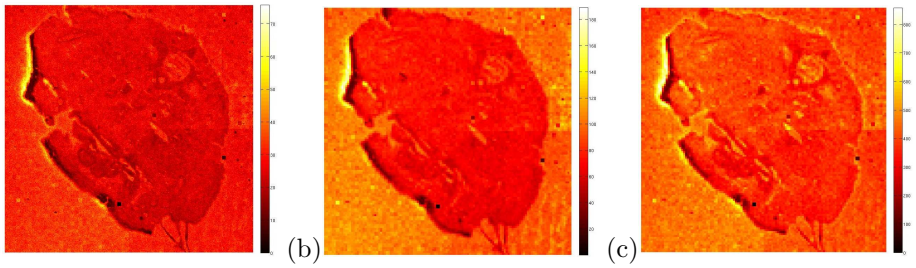


Figure 6.4: *Similar extracted components with enhanced contrast by (a) spectral binning with spectral bin size* 4096*, (b) spatial binning with spatial bin size* 16 *and (c) spectral and spatial binning with bin size* 4096 *and* 16 *respectively.*

binning of raw channel data.

2. Apply PCA on the low resolution datacube which results in a decomposition with a number of spectral and image components.

3. Display the resulting spectral and image components to a user for inspection.

4. A selection of spectral and/or image components is made that contain interesting features or undesired artifacts.

5. Create a new, rebinned datacube with higher resolution from the original channel data with the interesting features included or the excluded undesired artifacts.

6. Repeat step 2 to 5 until the extracted components contain interesting features on the highest resolution.

6.3.1 Binning and PCA

Binning is simply the act of grouping neighboring spectral variables by summing their intensity values into a single, new spectral variable. The same principle can be applied spatially by summing the intensity values of certain area to create a new spatial variable. This way, the signal-to-noise ratio of a spectrally and spatially binned datacube is increased, while the resolution is decreased. An example of this increase in contrast is shown in Figure 6.4. These three extracted components are all similar to the component in Figure 6.2. Figure 6.4a is binned spectrally, (b) spatially and (c) both spectrally and spatially. All image components show more contrast compared to the component in Figure 6.2.

Different methods for decomposition or factor analysis can be used for this feature visualization. PCA still has satisfying results with respect to the computational complexity, discrimination between extracted components and ability to identify correlations as well as anti-correlations between spectral and spatial dimensions. PCA is applied to the datacubes to select and extract the most important correlated spectral bins with the summed spectral peaks.

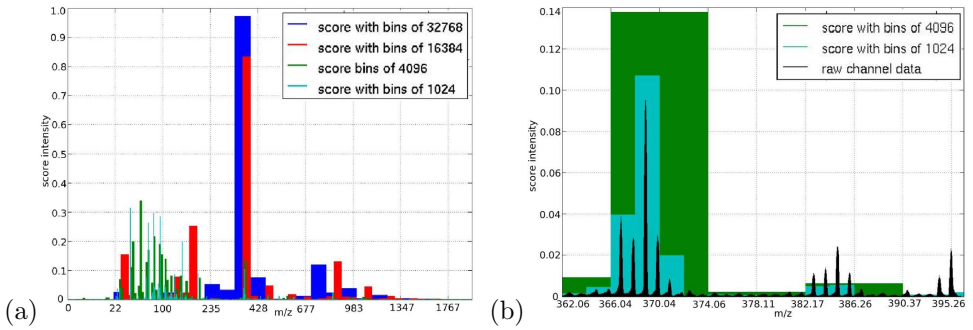


Figure 6.5: *(a) Four selected components with the same spectral peak of interest, each bin with a different number of spectral variables. (b) Close-up of the components in the large red bin with* 16384 *spectral variables seen in (a) and the raw channel data in black.*

6.3.2 Selection and zooming

In the previous section, we showed that larger spectral and spatial bins increase the separability and contrast in the resulting components. After binning and feature extraction, the next step in the pipeline of our approach is selection of the resulting spectral scores and image components that contain features of interest (see Figure 6.3). An analyst can select interesting spectral or spatial regions. PCA can be applied to these regions once again, but this time at a higher spectral and spatial resolution, because smaller bin sizes are used.

Spectral dimension

Some examples of extracted spectral components are shown in Figure 6.5a. The blue spectral score is one of many extracted components, of which only the positive part is displayed is this histogram. The smaller red bin (in the middle) highlights the same feature of interest at a similar spectral location. The bin containing the peak with the highest contribution to that particular component is selected again. Because of the reduced bin size, the component is now much more specific. Multiple red bins contribute to the selected component. Once the bin size becomes smaller, the contribution from the other bins increases. This phenomenon becomes more apparent in the green and cyan score vectors. These are still the closest possible representations of the same feature. However, we now see the dominating contributions to the resulting component of the multiple bins on the left of Figure 6.5b in much more detail.

After identifying a feature of interest at a low resolution, the bins in this score vector are selected by a threshold set by a user. Only those bins with a high intensity score are selected to construct the new dataset to which PCA is reapplied. Figure 6.5b shows a close-up of a spectral area of interest with 16384 spectral variables. The two score vectors are plotted behind the raw data to show how a spectral selection can be made at the highest resolution. If zooming was not applied, it would have been a time-consuming job to find the small peak of interest based on the traditional bin size of 1024 spectral variables. Its contribution might even have remained hidden by the dominating other bins in the score vector.

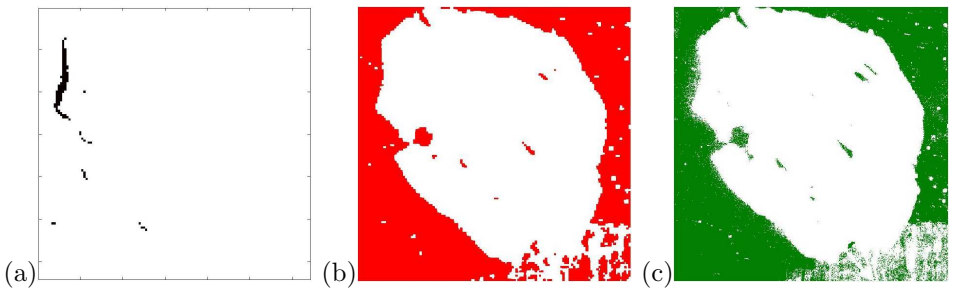


Figure 6.6: *Region selection by applying a threshold to image components with (a) a low resolution (*128×128*) image component to exclude a dominating artifact on the sample surface, (b) a different component to exclude areas outside the object of interest and (c) the same, more cluttered region selected at a high resolution (*512×512*).*

Spatial dimension

A similar procedure can be followed for spatial zooming. PCA is first applied to a spatially binned datacube at a relatively low spatial resolution. Image components resulting from PCA are organized and displayed. This enables a scientist to include or exclude selections of features for the application of PCA on a spatially zoomed part of the datacube. Certain image components have both a positive and negative loading. An appropriate threshold is applied to the intensity loading of a component image. Small intensity loadings contain mostly noise and are therefore excluded from the selection.

The thresholded images act as a mask to include or exclude certain spatial areas. A better selection of areas can be established when a spatially binned area is used as shown in Figure 6.6. Figure 6.6a contains a small area with a dominating, but less interesting artifact that reoccurs in almost each extracted component. The red area in Figure 6.6b is an image component selected by a user and contains a section outside the area of interest. This spatial selection can be excluded from further analysis to reduce clutter in the final results. The contour of this image component has less cluttered boundaries compared to the green area in Figure 6.6c, which is a similar image component extracted at a full spatial resolution. This example shows that a low spatial resolution can be used to create better defined boundaries without having to apply any image segmentation algorithms.

6.4 Results

The data used in this example were measured using TOF SIMS. The sample is a thin cross-section of a chicken embryo. The cross-section is 8×8 mm in size and contains a spectral mass window from $\sim m/z\,1 - 2000$. The dataset consists of four separate measurements, each with a spatial dimension of 256×256. We registered and combined the four measurements into one spectral datacube with a spatial dimension of 512×512 and a spectral dimension of $\sim 2 \cdot 10^6$ intensity values. In this example, the MatLab$^{\mathrm{TM}}$ environment is used with a sparse implementation of the eigenvalue decomposition.

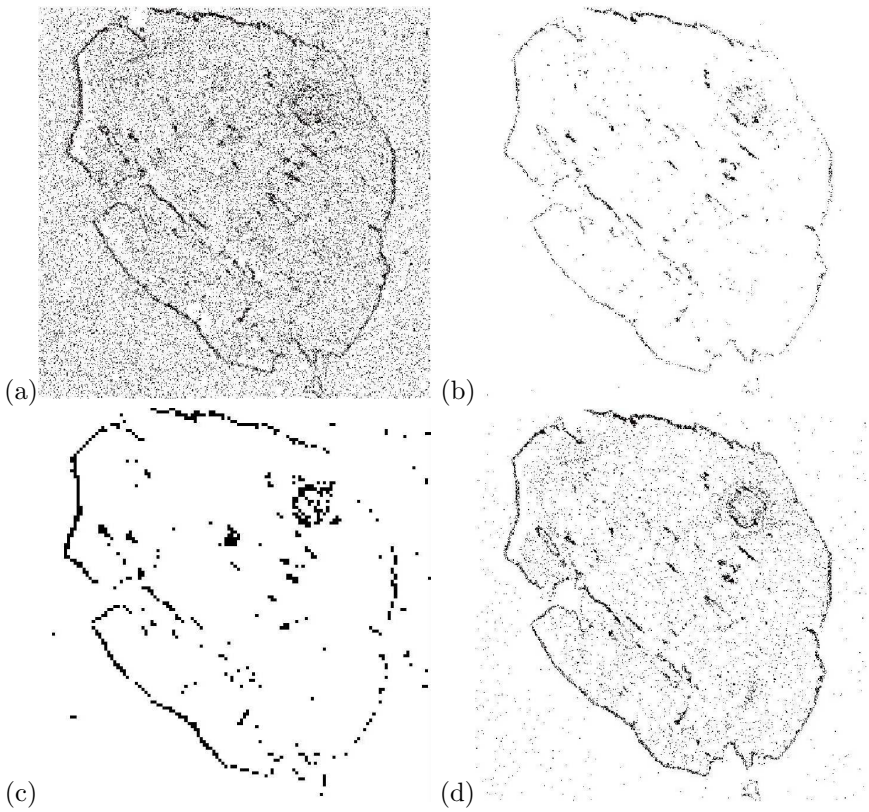


Figure 6.7: *Different extracted cholesterol image components with (a)* 1024 *spectral variables in one bin, (b)* 32768 *spectral variables in one bin, (c)* 1024 *spectral variables and a low spatial resolution and (d) a zoomed selection of spectral bins with* 1024 *spectral variables at the high spatial resolution.*

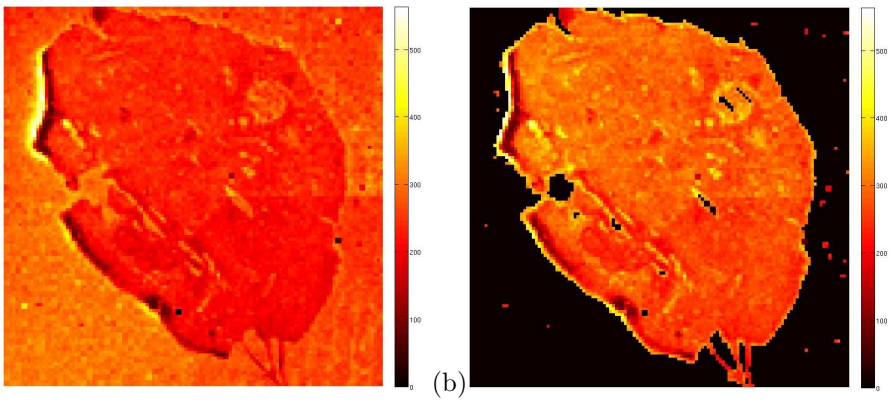


Figure 6.8: *(a) Image component by application of PCA to the complete datacube and (b) improved selectivity and contrast of a similar extracted component with a selected region of interest.*

6.4.1 Spectral zooming

In many cases, interesting features are cluttered in image components, mainly by the low signal-to-noise ratio in the part of the mass spectrum with the higher mass-to-charge ratios. Our method is able to extract more discriminating features, with higher contrast in the image components. A good example of an interesting feature is the distribution of cholesterol within the embryo itself. Figure 6.7a shows the twenty-fifth principal component in the sorted list of eigenimages of the traditional application of PCA on 1024 spectral variables in one bin.

Figure 6.7a is the closest representation to be classified as the cholesterol distribution in the sample, but it contains considerable noise indicated by the dispersity of the black dots. The black dots represent the presence of cholesterol with a threshold of $max(image)/2$ applied to the values of each image component. Cholesterol is distributed mostly towards the edges of the embryo, and a little is situated around the organs within the cross-section. Figure 6.7b shows the corresponding component with spectral bins that are thirty-two times larger. This is component number eight and shows significantly more contrast towards the edges of the sample. Once the datacube is binned spatially as well, and each square of 4×4 locations is combined into one, the resulting cholesterol component has another increase in contrast (see Figure 6.7c). The final distribution of cholesterol after spectral zooming is shown in Figure 6.7d. This image has the same resolution as the image in Figure 6.7a, but the cholesterol feature can be extracted and visualized with more detail.

6.4.2 Spatial zooming

The selectivity of feature extraction and therefore the contrast of the image is enhanced if the selection mask of Figure 6.6b is applied. This way, the area and noise outside the cross-section of the embryo are omitted when PCA is reapplied. Both spectral scores in Figure 6.8a and 6.8b have similar spectral peaks. The main difference being that two peaks are excluded from the spectral score of Figure 6.8b. The

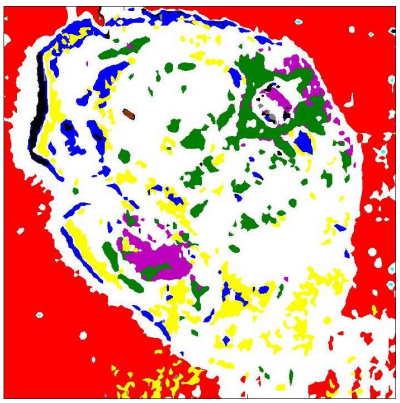


Figure 6.9: *A low resolution overview of the components after PCA on a spectrally and spatially binned datacube. Each colored shape represents an extracted feature that can be selected or excluded for zooming.*

spatial distribution of those peaks is mainly located in the omitted area.

It is fairly easy to create two-dimensional contours of low resolution principal components. Hence, it is possible to show one overview of the sample without using additional segmentation algorithms. An analyst can use our approach to spectrally and spatially zoom into extracted features using the automatically generated 2D overview in Figure 6.9. The green shape represents the negative part of the second principal component. This feature depicts most of the hard tissues inside the embryo and includes for instance bone. The red area is the positive part of the second component and clearly covers the area outside of the embryo itself. The blue areas were identified as cholesterol, according to the peaks in the spectral scores, which are located between $m/z\,368.2$ and $m/z\,369.13$.

A final representation of a selected feature can be made when zooming at the highest resolution, using the parametric visualization technique of Broersen et al. [Bro07]. This technique creates a 3D representation of a selected feature. The cholesterol distribution, for example, could not automatically be extracted and visualized by the traditional application of PCA. Within the components extracted in a traditional manner, the contribution of this peak was too small to be noticed at once. With the increased selectivity of the zooming approach, it is possible to locate the peak among the components extracted automatically.

6.5 Discussion

In this chapter, we proposed a method for selective zooming in spectral datacubes by recursive application of PCA. By combining two well-known techniques, binning and PCA, we generally increase the signal-to-noise ratio spectrally and spatially. Less noise improves the feature extraction with PCA in two ways. Firstly, more distinguishing features can be extracted from a multi-spectral datacube. Secondly, the spatial contrast or detail in these features is higher. By selecting a component of interest, an analyst can spectrally and spatially zoom into these improved results

(the spatial distribution and the spectral window in which the feature is represented). Zooming is accomplished by removing uninteresting parts and increasing the resolution in an interesting area. After increasing the spectral and spatial resolution, PCA is reapplied. It is possible to mask irrelevant areas in the datacube and to increase contrast by using component shapes. Additionally, binning and selective zooming reduce the number of variables. This creates the possibility to apply PCA to several spectral datasets which are stitched together. Initially, the combined dataset is too large to apply PCA to. However, the zooming functionality overcomes this problem of size. Hence, it takes full advantage of the increase of spectral and spatial information in the combined dataset.

In future work, zooming may even be enhanced by experimenting with other data-specific spatial filtering techniques. The noise in a binned multi-spectral datacube could be reduced, for instance, by applying Gaussian convolution or other conventional smoothing techniques. This operation would be computationally expensive at a high resolution datacube, but takes less effort when applied in combination with our zooming approach. Another addition to this approach could be the use of a different PCA-based feature extraction technique. A similar technique is the PARAFAC model of Harshman [Har70]. Kiers [Kie91] has shown that PARAFAC can be considered a constrained version of the two-way PCA. PARAFAC uses fewer degrees of freedom to fit the data on a simple model and can put constraints on the resulting factors, for instance a non-negativity constraint. This increases contrast between extracted features, but against an increased computational cost [Kle07]. Again, with our method, the number of spectral planes and locations—and therefore computational costs—are reduced. If PARAFAC, instead of PCA, is used for feature extraction the execution time is increased roughly with a factor thousand, depending on the data and the number of extracted components. With our zoomed approach the bin size increases with a factor of thirty-two, the execution time for feature extraction is cut down by two-third.

6.6 Summary and conclusion

This chapter describes an approach for feature based zooming on mass spectrometry datacubes. The approach primarily utilizes the data reduction technique of binning, which is commonly used in imaging mass spectrometry. This approach is primarily designed to enable feature exploration in fused imaging mass spectrometry datasets after registration with the approach described in Chapter 4. The combined spectral datasets are too large in size to be explored and visualized using commonly feature extraction and visualization techniques. Analysts are able to select important features or deselect unimportant features. Again, feature extraction is applied to the dataset, which is binned to make visualization on a higher resolution possible. By selectively removing spectral and/or spatial regions with noise or uninteresting features, new features can be found with greater accuracy.

Chapter 7

High-resolution feature visualization

Chapter 6 introduced an approach for zooming in on imaging spectrometry data using feature extraction. After zooming in on a feature on the highest possible resolution, it is necessary to control the representation of the feature to be able to use it in analysis. Without such control, the high-resolution spectral data has a signal-to-noise ratio, that is too low to be insightful for analysts. A 3D representation of the spectral data with high spectral resolution creates well-defined feature borders and is useful to gain more insight into the noise present in a measurement.

In this chapter, we present a parametric visualization technique, which allows an analyst to examine spectrally and spatially correlated patterns on the highest possible resolution. The extracted features are represented as abstract geometric shapes using three parameters to allow for data exploration. The first parameter thresholds the spectral contribution at which an extracted component is visualized. The level of detail of the shapes is controlled by a second parameter. A third parameter determines at which density-level the extracted feature is represented. With this method, the visualization of extracted features includes less noise. Moreover, by introducing various levels of detail the full spectral resolution can be utilized.

7.1 Goal

It is our intention to create exploratory visualization techniques with as few as possible data-specific denoising or complex clustering methods. At the same time, we want to be able to visualize features of these enormous datacubes at the full spectral and spatial resolution. In the most simple case of exploration, a spectral window is selected by hand using the histogram in Figure 7.1a. Here, all intensities at a single location are summed to create one image. Other methods almost always use a limited set of spectral planes compared to the amount of planes imaging mass spectrometry supplies. Figure 7.1a shows the sum of all spectral profiles in a datacube. However, this view had to be simplified by combining several neighboring spectral levels into one bin.

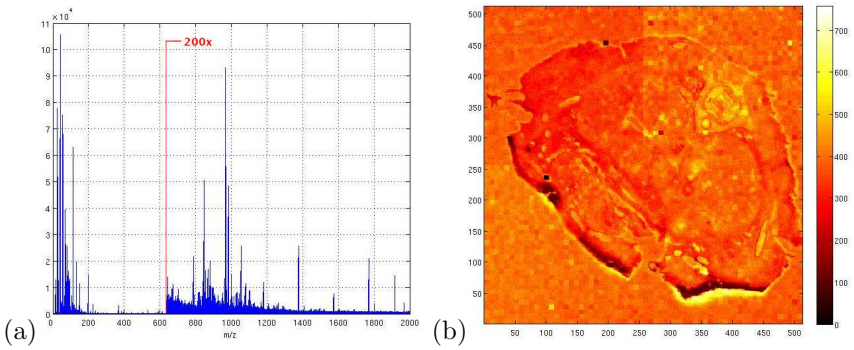


Figure 7.1: *(a) The sum of all binned spectral profiles with left part zoomed* 200 *times and (b) the matching sum of image planes.*

7.2 Parametric feature visualization

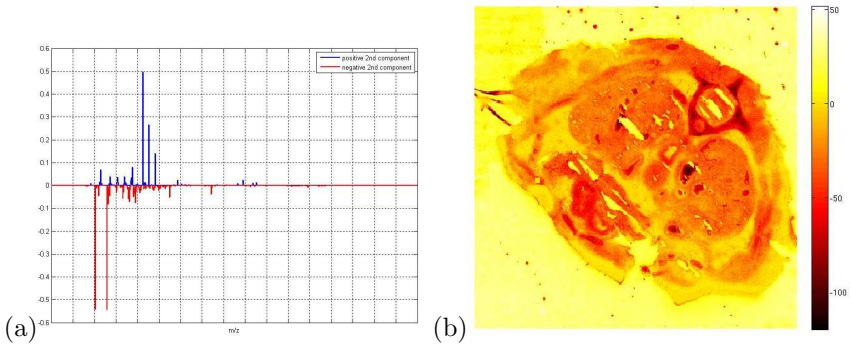
After detecting features in imaging MS data, results are commonly visualized as a spectrum or spectral image. Features in mass spectral data are distributed among multiple spectral channels. A spectral image is conventionally visualized by taking the sum of a range of spectral channels, thus visualizing one (or more) spectral bin(s). The separate spectral channels in these images can not be distinguished from each other. High-resolution topological information is removed although it could be useful in the interpretation and analysis of the data. The resulting spectral images do not contain shapes with well-defined edges. These shapes represent the distribution of a molecular substance. The spatial density distribution can be estimated better if the high-mass information is not removed. This could result in boundaries which are defined better, thereby distinguishing a feature more clearly.

The spatial distributions of features in different spectral ranges have to be compared to find similar or correlated patterns. These patterns can indicate connections between different molecular substances. When the amount of a particular substance increases, while another substance decreases correspondingly, it can be said that these substances are anti-correlated (i. e., having a negative correlation). This phenomenon is also useful for the analysis of interaction between molecular substances. The average signal intensity of the correlated features can vary, as well as the number of spectral peaks within a feature. Therefore, an analyst should be enabled to adjust the representation of the visualized features.

We present a new visualization technique enabling the user to:

- use the highest possible resolution instead of a spectrally binned resolution;
- extract features as 3D shapes with boundaries which are defined better;
- visualize spectrally and spatially correlated and anti-correlated patterns;
- parametrically explore multiple features within the same view.

Our feature visualization is controlled by three parameters. The first parameter α is set as a threshold for the spectral contribution of an extracted feature. This way, only



(a) (b)

Figure 7.2: *(a) The spectral profile of a second principal component and (b) the matching second image component.*

the spectrally correlated windows with the highest contributions to that extracted feature are included in the visualization. This enables a user to remove smaller peaks and noise cluttering the visualization. A second parameter β controls the level of detail of a 3D feature. A feature can be represented as a simplified smooth 3D shape with high resolution, thereby containing more details of the structure. The third parameter γ is used to determine at which level of density in the data a geometric shape is created. A family of iso-surfaces can be created to explore areas with different densities in the feature. Extracted iso-surfaces contain less noise compared to 2D contours extracted from an image component. Therefore, these three parameters enable us not only to display multiple correlated features, but to represent them as 3D geometrical shapes, while containing less noise than a traditional 2D view.

7.2.1 Extraction and visualization

Various visualization techniques have been proposed to inspect datacubes. The most basic technique uses a TIC image (Figure 7.1b), in which a side-by-side view of spectral and spatial domains can be analyzed. However, it is left to the user to identify which chemical compounds are present in the datacube and whether or not their spatial distributions are correlated. There are some complex fuzzy logic segmentation algorithms [Wol99] as well, but these can only be applied to a limited number of spectral windows. Also, a few implementations exist to visualize 3D spectral imaging data in the spectral or spatial domain. Visualization implies coping with a number of difficulties. First, the 2D spatial information with added 1D spectral information can not be treated the same as 'real' 3D volumes, resulting for instance from a CT or MRI scan. To overcome this problem, most techniques apply feature extraction using factor analysis first, for instance Kenny et al. [Ken97] or Keenan [Kee05]. Feature extraction is closely related to compression or dimension reduction techniques. It targets the removal of redundant data or data that mostly contains noise. Unfortunately, both studies are unable to use the full available spectral resolution in their final visualizations. The second difficulty is to find the most appropriate technique for feature extraction or dimension reduction, which is specific for each spectral dataset. A third problem is the ever increasing size and resolution of the datasets. This problem makes

both visualization and feature extraction more difficult even with increasing computational power. For instance, Haigh et al. [Hai97] use correlation partitioning on five spectral channels turning to dimension reduction techniques. Other visualization techniques use between 100–300 spectral channels to visualize spectral data with flat image overlays with color weighting envelopes [Jac05]. Others such as Polder and van der Heijden [Pol01], apply volume visualization techniques. These techniques work with a spectral dimension within visible light. The spectral dimension depicting the wavelength has continuous intensity values. The spectral dimension resulting from mass spectrometry can have $\sim 2 \cdot 10^6$ spectral channels and can be considered as a cloud of single 3D points. Because mass spectrometry datasets consist of a cloud of data-points with an increased amount of spectral channels, the aforementioned methods of feature visualization can not be applied.

In almost all attempts to explore and visualize the enormous datasets resulting from mass spectrometry, multivariate statistical analysis tools are used. Most tools tend to focus on denoising [Wic03] or specific 1D filtering techniques [Kee04b]. The tool AXSIA (used in Smentkowski et al. [Sme04]), for instance, statistically aggregates spectral profiles to identify features in the data, but the results are still shown as separate spectral profiles and summed spatial distributions. AXSIA claims to decompose the datacube more intuitively by disallowing negative spectral contributions. Many successful multivariate tools [Kle07; Pac04] for spectral feature selection and unsupervised exploration use PCA, which is less time-consuming. Nevertheless, there are several disadvantages. Although with PCA correlations between spectral peaks and their spatial distribution can be studied in a single view, feature extraction can not be parametrically controlled. Furthermore, any noise inside the spectral bins is included in the resulting volume rendering. Combining high-resolution spectral channel data into one bin in order to apply PCA, results in the loss of spectral information. A final disadvantage is that it is impossible to select a spatial region or specific spectral window inside an extracted principal component for further examination.

In our approach we focus on visual parametric exploration of the datacube. Although with PCA, correlations between chemical components can be found unsupervised, much spectral information is lost when visualized in the traditional two dimensions. We use the full spectral resolution in feature visualization to reduce noise as much as possible without having to focus on advanced and computationally expensive algorithms. It highlights positively correlated features, as well as their negatively correlated counterparts in one parametrically simplified view.

In our method, the features are extracted in a four-step process. First, principal component analysis is used to discriminate specific components present in the datacube according to their spectral correlation. Then, the most important spectral windows are parametrically selected to exclude smaller spectral contributions containing more noise. In the third step, the selected windows are convolved into continuous scalar fields to be able to extract appropriate iso-surfaces from those regions where the data is most dense. In the final step, correlations between extracted features are visualized at their 2D locations at the highest spectral resolution. The adjustment of the parameters α, β, and γ (defined in Section 7.2) allows the user to interactively analyze and highlight the spatial and spectral distributions of the chemical elements and molecules on the surface of the material.

7.2.2 Principal Component Analysis

Although different methods for decomposition or factor analysis can be used for our approach, we use PCA [Jol02]. It has satisfying results with respect to speed, discrimination between extracted components and identification of correlations as well as anti-correlations between spectral and spatial dimensions. We decided not to normalize or auto-scale in our technique, because Keenan and Kotula [Kee04b] showed that with mass spectrometry, common preprocessing steps such as normalization or auto-scaling can lead to less satisfactory results.

As shown in Chapter 3, PCA is applied to the datacubes to extract the most important correlated spectral profiles. This way, the thousands of spectral profiles are decomposed and compressed into a few main components that capture the main characteristics of the data. These components especially contain spectral peaks that are correlated. When sorted according to their eigenvalues, the first few components describe the most variance in the spectral data and therefore have the most contrast in the peak intensity.

Our study offers an approach to visualize these components in more detail by isolating positive and negative spectral peaks. Each component is used as a new base to project the original datacube as in Equation 3.7. This results in a matrix P with spectral loading vectors, which can be interpreted as spectral components. Here, each peak in a spectral component represents the contribution of a specific ion. An example of an extracted component is shown in Figure 7.2a, in which positive peaks are blue and negative peaks are red. In this component, the positive and negative parts are anti-correlated. The transposed datacube D can be multiplied by the spectral component matrix P to obtain the spatial distributions of these spectral correlations as in Equation 3.7. Each row in the resulting matrix Y contains an unfolded image component containing the spatial contributions of each profile in P.

All positive and negative values in a profile in P contribute to a component, even when the values are close to zero. However, the higher a—positive or negative—contribution to a component is, the more important it is considered to be. A threshold parameter α is defined to reduce the number of spectral bins that are used in the feature visualization. The thresholded contributions P_α^+ for positive and P_α^- for negative are given by

$$P_\alpha^+[m] = \begin{cases} m, & if\ m \geq \alpha \\ 0, & otherwise \end{cases} \tag{7.1}$$

$$P_\alpha^-[m] = \begin{cases} -m, & if\ m \leq -\alpha \\ 0, & otherwise \end{cases} \tag{7.2}$$

Often, a good initial choice for α is the highest possible value. As a consequence, only those peaks with the highest positive or negative contribution are left for further processing. This way, small or less important contributions containing more noise remain hidden at first. When the value of α is lowered, more correlated spectral bins are added to the visualization. Although these bins do not contribute as much to a principal component, they could contain correlated spatial or spectral characteristics. For example, when using an $\alpha = 0.3$ in the spectral profile of Figure 7.2a, those peaks with the largest contribution remain (in this case three peaks, as shown in Figure 7.3).

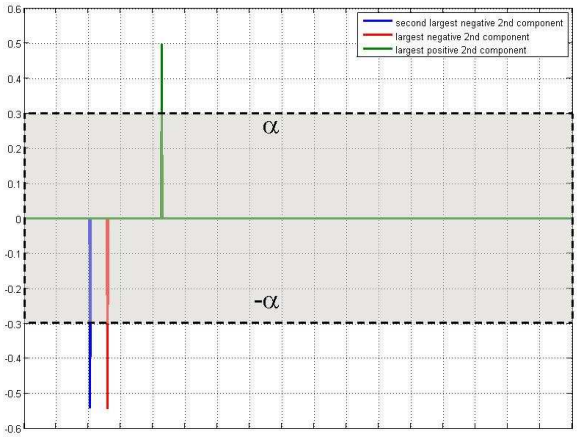


Figure 7.3: *Selection of spectral peaks outside the grey area with $\alpha = 0.3$ from the spectral profile of the second principal component from Figure 7.2a.*

This figure shows two negative contributions and one positive contribution. If necessary, a user can lower α and add more contributing correlated spectral windows to the resulting visualization.

7.2.3 Convolution

So far, PCA has been applied to a binned datacube. The bins with the highest contributions in a principal component were selected using a threshold α. The technique from Chapter 6 is used to zoom in on an interesting feature by reducing the size of the bins. After this, the selected principal components with the binned spectral profiles are then used to extract feature data from the original unbinned datacube. The resulting 3D clouds of data-points with high-resolution ion-counts do not reveal a clear structure. Most intensities ($\sim 99\%$) have either value one ($\sim 9\%$) or zero ($\sim 90\%$). To be able to visualize more structural details within the cloud, a 3D convolution filter transforms the datacube into a scalar field with continuous values. This low-pass frequency filter blurs the volume in such a way that those regions with a high concentration of data values can be represented by an iso-surface. Choosing smaller kernel sizes, more fine-scaled anomalies will appear in a scale-space [Wit83; Koe84] representation of the extracted features.

The second parameter, called β, controls the size of the kernel and therefore the level of detail of the smoothed 3D feature. For smoothing, a standard Gaussian isotropic convolution kernel h_β is chosen in such a way that β is the variance of the Gaussian kernel, as defined in

$$h_\beta[x] = (2\pi\beta)^{-n/2} \cdot e^{-\frac{\|x\|^2}{2\beta}} \tag{7.3}$$

where $\|x\|$ is defined as the length of multidimensional vector x and n is the dimensionality of vector x. The kernel h_β has the same value for β in both spectral and spatial dimensions to keep the representation of the density distribution the same in all three dimensions. To be able to apply the convolution filter to 3D datacubes in

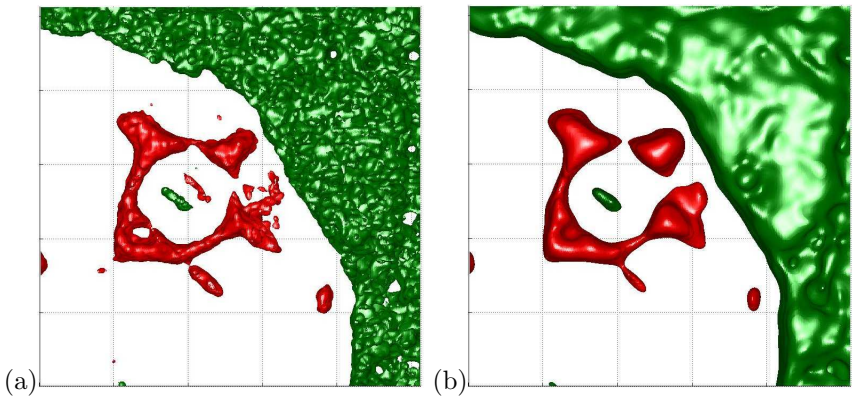


Figure 7.4: *The iso-surfaces of two anti-correlated features within the second principal component with (a) $\beta = 16$ and (b) $\beta = 32$.*

an interactive visualization, the Discrete Fourier Transform (DFT) can be used, as in Nussbaumer [Nus81] and Geusebroek et al. [Geu02]. According to the convolution theorem, a convolution in a spatial domain is equivalent to multiplication in the frequency domain. The 3D convolution filter can now be defined in the frequency domain using the discrete Fourier transform in

$$F[k] = \sum_{n=0}^{N-1} f[n] \cdot e^{-2\pi ikn/N} \tag{7.4}$$

Here are $n = (n_x, n_y, n_m)$ and $k = (k_x, k_y, k_m)$. n and k defined as the three-dimensional vectors of indices of the selected datacube $N = (N_1, N_2, N_3)$ to simplify the equation. After the Fourier transformations of datacube $f[n]$ and the filter $h_\beta[x]$, the results are multiplied, as in

$$G[k] = F[k] \otimes H_\beta[k] \tag{7.5}$$

After this, the inverse discrete Fourier transform results in a convolved datacube

$$g[n] = \frac{1}{\prod_{l=1}^{3} N_l} \sum_{k=0}^{N-1} G[k] \cdot e^{2\pi ink/N} \tag{7.6}$$

It is now possible to extract iso-surfaces from the high density regions. These iso-surfaces represent a high concentration of a specific chemical element at a certain location without losing high-resolution spectral information. Figure 7.4 (an enlarged version of the top right part of Figure 7.1b) shows how β influences two extracted features.

7.2.4 Correlated geometric shapes

Each extracted iso-surface represents the spectral and spatial distribution of elements or molecules in the datacube. These iso-surfaces can be visualized as different geometric shapes. PCA enables us to add more information to 3D shapes. Information

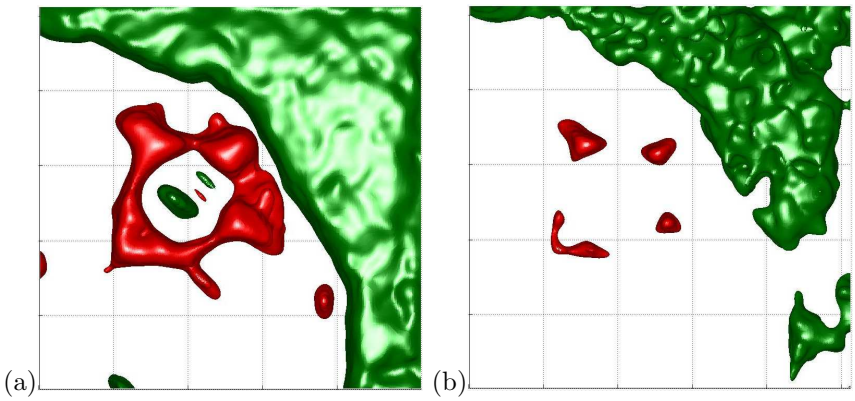


Figure 7.5: *The iso-surfaces of two anti-correlated features within the second principal component with (a) the red shape on $\gamma = 0.4$, a green shape on $\gamma = 0.2$ and (b) the red shape on $\gamma = 0.6$, a green shape on $\gamma = 0.4$.*

about the (positive and negative) correlation between peaks and regions could be used for the visualization of these shapes. The largest peaks in a spectral component are selected by α. After this, β is used to control the detail and convert the 3D clouds of data-points into a scalar field with continuous values. A third parameter γ is set to extract and show iso-surfaces of the selected features in the created 3D space $g[n]$. A histogram is created with the values in $g[n]$ of the extracted features, as shown in Figure 7.6a. γ is defined as an intensity value (or iso-value) on the horizontal axis of this histogram. The low intensities on the left side of the horizontal axis represent those points in the areas of $g[n]$ with a low density. The points in the areas with a high density are on the right of this intensity scale. Each iso-surface is extracted by selecting a particular iso-value γ. The effects of choosing different values for γ are shown in Figure 7.5. Figure 7.5a contains 3D feature shapes with low values for γ and thus represent those areas in the datacube with lower densities. The areas with higher densities are represented in Figure 7.5b and have higher values for γ.

7.3 Results

In this chapter, we refer to the dataset acquired in Section 6.4. All its spectral intensities are summed in a single image (Figure 7.1b). This makes it impossible to distinguish between different values in a spectral profile and their corresponding specific spatial contribution. In this type of representation, interesting features like heart, blood vessels, bone structures or distribution of cholesterol remain hidden or are poorly visible at best. It is hard to distinguish the cross-section itself from the material in which it is embedded. In our approach, α is used to reduce the amount of spectral noise in the selection of spectral windows. A second parameter β enables a user to view the resulting features on different levels of detail. The highest level of detail shows the original cloud of points from one particular spectral window, but the iso-surface of the unconvolved data does not reveal clear coherent information in the cloud. The information of the spectral structure becomes more apparent when

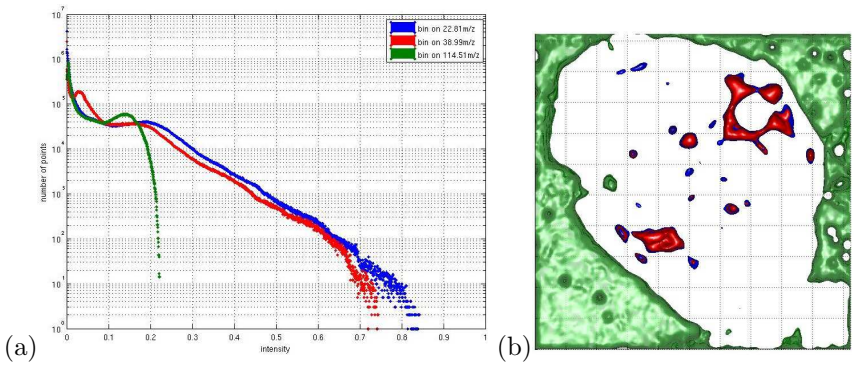


Figure 7.6: *(a) The histogram with the densities of three analytes in the second principal component from which an appropriate value for γ can be selected and (b) the extracted correlated shapes with $\alpha = 0.3$, $\beta = 16$ and $\gamma = 0.35$ for the blue and red shape, $\gamma = 0.15$ for the green shape.*

smaller values are chosen for β. Figure 7.4 shows the iso-surfaces of two anti-correlated features within the second principal component. It shows the same part of the cross-section of the backbone with the red shape representing sodium and the green shape representing indium. Clearly, more structural details can be seen in the image on the right of Figure 7.4. For instance, small red regions appear beside the backbone that can be identified as blood vessels. The neural tube (represented by the hole on the left-bottom part of the cross-section of the backbone) is visible in the image on the right, whereas it is closed on the left. Different values for β can be used to find a balance in the complexity of the structure of the iso-surfaces and the desired level of detail. The size of the extracted shapes can be controlled by the third parameter γ. Figure 7.6a shows the histogram with the densities of three analytes within the second principal component. In Figure 7.5 we show that different values for γ can be used in order to find an appropriate level of density to display the component. Regions with the highest data density are selected by choosing higher values for γ.

Using all three parameters at the same time results in the visualization of the cross-section in Figure 7.6b. In this example, we used the second principal component only, as it displays a clear distinction between bone tissue and the material in which the cross-section is embedded. Again, the red and blue shapes are elements that are correlated and the green shapes represents the anti-correlated material outside the embryo. The holes in the green shape are caused by fragments of other elements. If desired, they can be deselected. The irregularities on the green surface are due to noise artifacts in the sample itself. An expert is able to interpret the distribution of elements in this visualization. For instance, the blue element (potassium) shows a similar distribution as the red element (sodium). Both are present in bone-tissue and blood. The large red shape on top of the figure can be identified as a cross-section of the backbone. The large red shape on the bottom can be identified as the heart. Different principal components can be used to create multiple views of the distribution of correlated features within the same datacube. For instance, if other components contain elements or molecules present in the heart but not in the bone (or vice versa) they can be classified and separated as different types of tissue.

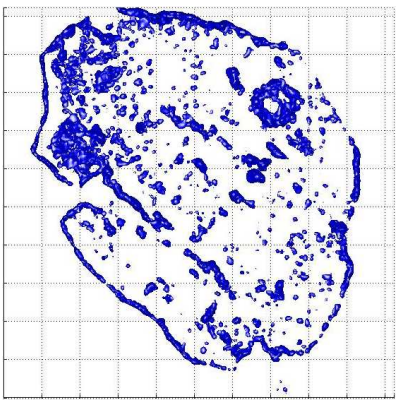


Figure 7.7: *Isolated cholesterol distribution at the highest resolution between* m/z 368.2 *and* m/z 369.13 *found by PCA-based zooming described in Chapter 6.*

The PCA-based zooming technique introduced in Chapter 6 enables us to automatically extract a component containing cholesterol distribution in the cross-section. This feature could not be detected without the zooming approach. A resulting visualization of the cholesterol distribution at the highest resolution can be found in Figure 7.7. The blue areas, mainly situated on the boundaries of the cross-section of the embryo, represent high concentrations of cholesterol. This high-resolution feature is represented by 3D shapes instead of the raw data-points. As a result, the boundaries of those areas with a high density of the cholesterol feature are visualized with sharply defined borders.

When PCA is applied using the highest possible spectral resolution, it can be used to identify different peaks. These peaks could remain hidden in one merged peak when observing the summed spectrum. For instance, Figure 7.8 shows how a single peak in the summed spectrum of a measurement (b) actually consists of three separate peaks as shown by the principal component in (a). There are several reasons why these peaks are represented as a single peak. There could be analytes with different m/z values. Also, there can be noise in the measurement due to differences in height on the surface. Finally, instrumental noise can be present because of optics: this is a so-called 'ringing'-effect, due to optics within a mass spectrometer, which manifests itself as a difference in height. With our method, this noise can be visualized as a high-resolution representation of a peak. Figure 7.8c and d (resulting from the droplet dataset in Figure 5.7) are examples of high-resolution visualizations of one peak. These examples demonstrate how a 3D view provides more insight in the 3D distribution of peaks in a spectral datacube. Figure 7.8c shows the 3D representation of the spectral and spatial dimension and Figure 7.8d the spectral dimension with another spatial dimension. The red iso-surface has a higher density than the blue iso-surface. Both views clearly show the additional value of a high-resolution visualization for identifying anomalies in spatial structures of peaks. After identification, these effects can be removed by correcting these differences in height and therefore deconvolving the data in the spectral dimension.

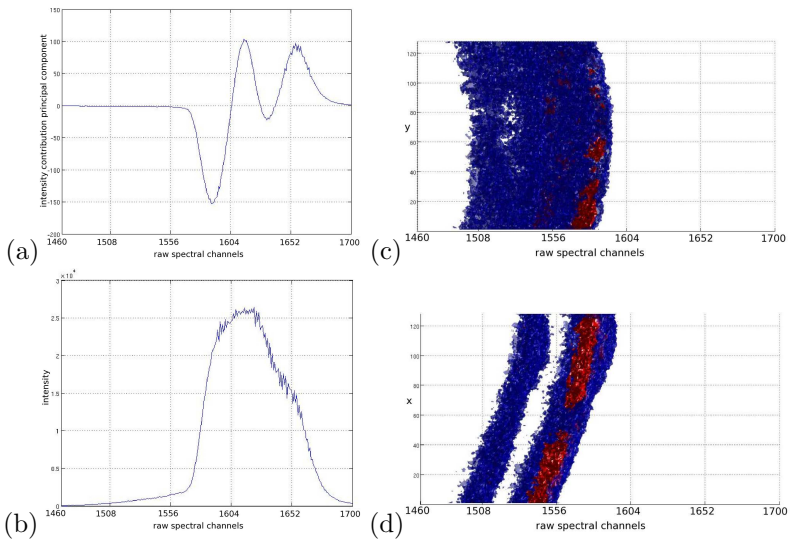


Figure 7.8: *(a) PCA component that separates three peaks from (b) one peak in the summed spectrum using a high-resolution visualization with (c) the spectral dimension against y and (d) the spectral dimension against x.*

7.4 Discussion and future work

The proposed method of parametric visualization of high-resolution correlated features has a number of advantages compared to the classical method of manual exploration. First, the extracted correlated and anti-correlated patterns are made distinctive through different colors. The threshold α is used to reduce the spectral noise present in a single feature. Second, we make use of a dataset at its full spectral resolution, while other tools for visualization use a binned dataset. Third, our method displays the extracted features with better defined boundaries, because the image is extracted in three dimensions rather than in just 2D. Therefore, our shape extraction contains less spectral noise. Our shapes contain only those 3D regions which have the highest density in contrast to low-density regions with lower signal-to-noise ratio. These geometric shapes are used to conveniently select one or more spectral windows or even just a single spatial region of interest. Finally, our visualization is parametrically controlled in such a way that an analyst is in control of feature extraction and can set any desired level of detail.

With our approach, some aspects have to be taken into consideration. Our visualization depends on the effectiveness of PCA. We chose PCA for feature extraction, because it has already proven itself in this field of application. Other methods for decomposition can be used as well, but—due to the enormous size of the datacubes and the distribution of peaks among multiple spectral levels—it is not yet possible to apply the algorithm on a dataset at its full resolution. So, even while depending on the effectiveness of PCA, our method creates a more satisfying representation of the distribution of mass spectral components than other known methods. Our zooming technique can also be used on features extracted with PCA-based methods (see

Section 3.3). In fact, the application of PCA-based techniques is similar to PCA regarding filtering by binning resulting in spectral and spatial components. However, with PCA the computational load is considerably smaller.

Furthermore, our method is able to provide an intuitive interpretation of the negative scores that result from PCA [Sme04]. Moreover, the presence of negative scores even contributes to our visualization, because it allows for the display of anti-correlated features. It is easy to experiment with different 3D filtering techniques, for instance anisotropic convolution, which can provide less smoothed boundaries in the extracted shapes. Eventually, we would like to add a functionality to this parametric visualization method which is able to select the most appropriate values for the three parameters we introduced. These values have to be independent from the methods for decomposition, use of different convolution kernels and most importantly, the type of spectral dataset used.

7.5 Summary and conclusion

This chapter explains how features are parametrically visualized at the highest possible resolution. Three parameters control the spectral contribution, the level of detail and the level of density on which an extracted feature is represented. This visualization has feature shapes with well-defined borders and provides more insight into the influences of noise on a mass spectral measurement. It is possible to distinguish different peaks according to their difference in density and spatial position, which would not be possible in a separate spectral or spatial view. For a clear distinction, different colors are assigned to the positively or negatively correlated spectral peaks.

Chapter 8

Conclusions and future research

Imaging mass spectrometry is a powerful technique to measure the spatial distribution of molecular content on surfaces of biological samples. Modern developments in spectrometry instrumentation allow for data acquisition in continually higher mass and spatial resolution, resulting in datasets which have become very large and rich of detail. This thesis has explored the usage of features for the visualization of such datasets. The research questions in Section 1.5 were formulated as:

- How can PCA be used for robust feature detection in large imaging spectrometry datasets?
- How can features be used to improve registration, zooming, and visualizations of spectral datasets?

We have shown that PCA can be used for robust feature detection in mass spectral datasets. However, due to the low signal-to-noise ratio and the number of variables in these datasets, PCA can not be applied to the raw channel data of a mass spectral measurement. Therefore, a binning function has been used for preprocessing in order to combine several spectral channels into a single new spectral variable. We have shown that binning generally increases the signal-to-noise ratio in the data and reduces the size of the a mass spectral dataset which increases sensitivity of the feature detection by PCA.

8.1 Conclusions

PCA has several advantages. It does not require different parameters based on a chemical model of a specific dataset or based on the purpose and goal of the analysis. By combining PCA with binning, interesting features can be found. PCA can highlight features in original data-points for quantification, whereas the resulting principal components are not suited for quantitative interpretations. Both positive and negative parts in the principal components provide information about the correlation of different spectral profiles. Just one single parameter is needed in the process of binning to be able to determine the level of detail in the feature detection. Therefore,

PCA is ideal for the discovery of unknown features and the exploration of spectral datacubes without well-defined spectral peaks.

We have shown that features can improve registration, zooming, and visualizations of spectral datasets. Feature-based registration is more robust compared to registration of complete datacubes or registration of TIC images. The extended datacubes have more spectra, which improve signal-to-noise ratio and therefore feature detection. Feature-based zooming fully utilizes all available data. For example, it is possible to zoom in on a specific area with improved detection of features. Features can be visualized by direct volume rendering or in a high-resolution parametric 3D representation. Both visualization techniques provide a direct linkage between the spectral and spatial properties of a feature. Direct volume rendering gives an analyst better insight in the relations between features as well as their spectral and spatial distributions. High-resolution parametric 3D representations create distinct boundaries and provide more detailed characteristics of the surface of a feature.

All methods are implemented in the MatLab^{TM}-environment. This environment serves as a rapid prototyping environment for each method in this thesis. It offers a large variety of optimized implementations for (sparse) data-structures, mathematical operations, and visualization techniques. The flow of the data in the visualization pipeline (see Figure 1.2) can be easily adjusted in the MatLab^{TM}-scripts that implement our methods. These scripts are cross-platform implementations that can be used in batch processing for the automatic registration of multiple datacubes. Several mass spectrometrists use these scripts to create enlarged spectral datacubes, extract, and visualize features. All results in this thesis were obtained in close collaboration with these experts. The methods can be extended accordingly to the evolving methods for acquisition of mass spectral data.

8.2 Directions for future research

There are several directions in which this work can be continued and improved.

Detection

The binning function combines several spectral channels into a single new spectral variable by a fixed spectral window size. This binning function can be improved when the spectral window is dynamic—rather than fixed—according to a certain transfer function. An appropriate transfer function could be determined according to the differences in height of the surface [McD03]. This way, a deconvolution filter is created in which none of the peaks are broadened by the variation in height of the sample surface. This will improve the signal-to-noise ratio in the data.

Another improvement in the detection of features is possible when the spatial neighborhoods of spectra are included in PCA. Normally, PCA distinguishes variables and samples in a dataset. Therefore, relations between closely located spectral variables or spatial samples are not considered in a solution, i. e., spectral datacubes are not handled as images but a disarranged list of spectral measurements. It is not possible to incorporate spatial information in the implementation of PCA, but filtering could create the same desired effect. An appropriate 3D convolution filter will smooth data-points and include the intensities of neighboring spectra for improved

application of PCA. It is not complicated to smooth a small part of a datacube, but smoothing increases the number of data-points with non-zero values compared to the original sparse datacube. Alternative solutions have to be found for processing these non-sparse datacubes to be able to apply PCA.

A final improvement to feature detection can be made by the method used for the fusion of several datacubes. Spectral datasets have to be combined after finding the correct offset between different datacubes. Traditional methods use the spectra of the first measured datacube for the overlapping region, because these spectra are measured from the undamaged surface of the sample. The quality of a second measurement of the same surface area will be less, since the surface is affected by the first measurement. An alternative method for combining two datasets could use the information in the overlapping region to enrich these spectra in this area.

Visualization

Two improvements can be considered in the visualization of features. Additional uncertainty information can be visualized in the overlapping regions of several measurements. Two measurements of the same overlapping area can provide statistics on the data distribution in that area. With these statistics, uncertainty information about the data distribution can be determined and added to a visualization as an overlay on these spectral and spatial areas. This way, the signal-to-noise ratio could be visualized.

In addition, the visualization of features can be improved by applying additional statistics on the extracted features. Instead of visualizing spectral and spatial information, other (statistical) characteristics could provide more insight in the extracted feature. For example, the density distribution within a feature shape can be measured and visualized. Alternatively, the characteristics of two feature shapes could be compared. Such comparison could result in a metric to quantify the correlation between feature shapes.

Interpretation

The final interpretation of the visualized features is done by an analyst. Interpretation could be made less complex when a (statistical or model-based) weighting function is used on the spectral variables before PCA is applied. Depending on the type of measurement and goal(s) for analysis, an analyst would be able to control the contribution of the spectral variables in a solution. Characteristic noise in a measurement could be reduced (as in Lee et al. [Lee08]) and emphasis can be put on small, but important peaks. This is similar to our zooming approach. In zooming, features are either selected or deselected. With weighting, uninteresting features do not have to be removed completely, but their influence could be limited and controlled appropriately to an experiment.

Interpretation of spectral measurements could also be enhanced by combining measurements of different modality of the same sample. Different imaging techniques (e. g., microscopy, radiography, thermography) provide different characteristics of the same sample [Ery07]. If these datasets can be registered correctly, all datasets could be used for enhanced feature detection. These multi-modal features provide a more complex but enhanced view of the composition of a sample material. Besides having

more spatial distributions, signals of different modality would improve interpretation of features.

Another improvement for interpretation of features is the use of features already identified in previous analyses. The characteristics of these features can be used for a more robust or automatic classification. Quantitative information of the intensities of a combination of several peaks within a feature can be stored and used for feature detection in other—similar—spectral datasets. Feature detection would not be limited to the variation in intensity values, since other properties of previously detected features could be used as well. An analyst would be able to visualize and compare the same identified feature within several datasets without comparing all extracted components for similarities. This improvement would be useful in the comparison of samples in, for instance, biomarker detection.

Bibliography

[Aeb03] R. Aebersold and M. Mann. Mass spectrometry-based proteomics. *Nature*, vol. 422(6928):pp. 198–207, 2003. 15

[Alt07] A. F. M. Altelaar. *Biomolecular Imaging Mass spectrometry: mapping molecular distributions in cells and tissue sections*. Ph.D. thesis, University of Utrecht, Mar 2007. 14

[And00] C. A. Andersson. *Exploratory Multivariate Data Analysis with Applications in Food Technology*. Ph.D. thesis, The Royal Veterinary and Agricultural University, 2000. 25

[Bai00] Z. Bai, J. Demmel, J. Dongarra, A. Ruhe, and H. van der Vorst, editors. *Templates for the Solution of Algebraic Eigenvalue Problems: A Practical Guide*. SIAM, 2000. 49

[Ben87] A. Benninghoven, F. G. Rüdenauer, and H. W. Werner. *Secondary ion mass spectrometry: basic concepts, instrumental aspects, applications and trends*. John Wiley & Sons, New York, 1987. 2

[Ben94] A. Benninghoven. Surface analysis by Secondary Ion Mass Spectrometry (SIMS). *Surface Science*, vol. 299–300:pp. 246–260, 1994. 15

[Bro97] R. Bro. Parafac, tutorial and applications. *Chemometrics and Intelligent Laboratory Systems*, vol. 38(2):pp. 141–171, 1997. 28, 39, 44, 49

[Bro98] R. Bro. *Multi-way Analysis in the Food Industry*. Ph.D. thesis, University of Amsterdam, 1998. 44

[Bro99] C. D. Brown and P. D. Wentzell. Hazards of digital smoothing filters as a preprocessing tool in multivariate calibration. *Journal of Chemometrics*, vol. 13:pp. 133–152, 1999. 21

[Bro03] R. Bro and A. K. Smilde. Centering and scaling in component analysis. *Journal of Chemometrics*, vol. 17:pp. 16–33, 2003. 40

[Bro05a] A. Broersen and R. van Liere. Transfer functions for imaging spectroscopy data using principal component analysis. In K. Brodlie, D. Duke, and K. Joy, editors, *Eurographics / IEEE VGTC Symposium on Visualization*, pp. 117–123. Jun 2005. 9, 85

[Bro05b] A. Broersen, R. van Liere, and R. M. A. Heeren. Comparing three PCA-based methods for the 3d visualization of imaging spectroscopy data. In J. J. Villanueva, editor, *IASTED International Conference on Visualization, Imaging, & Image Processing*, pp. 540–545. IASTED Benidorm, Spain, ACTA Press, Sep 2005. 9

[Bro06] A. Broersen and R. van Liere. Feature based registration of multispectral data-cubes. In J. J. Villanueva, editor, *IASTED International Conference on Visualization, Imaging, & Image Processing*, pp. 543–548. IASTED Palma de Mallorca, Spain, ACTA Press, Aug 2006. 9, 86

[Bro07] A. Broersen, R. van Liere, and R. M. A. Heeren. Parametric visualization of high resolution correlated multi-spectral features using PCA. In K. Museth, T. Möller, and A. Ynnerman, editors, *Eurographics / IEEE VGTC Symposium on Visualization*, pp. 203–210. May 2007. 9, 85, 93

[Bro08a] A. Broersen, R. van Liere, A. F. M. Altelaar, and R. M. A. Heeren. Automated, feature-based image alignment for high-resolution imaging mass spectrometry of large biological samples. *Journal of the American Society for Mass Spectrometry*, vol. 19(6):pp. 823–832, Jun 2008. 9, 64

[Bro08b] A. Broersen, R. van Liere, and R. M. A. Heeren. Zooming in multispectral datacubes using PCA. In K. Börner, M. T. Gröhn, J. Park, and J. C. Roberts, editors, *Electronic Imaging*, vol. 6809, p. 68090C. SPIE-IS&T, Jan 2008. 9

[Bus05] I. Bustinduy, F. J. Bermejo, T. G. Perring, and G. Bordel. A multiresolution data visualization tool for applications in neutron time-of-flight spectroscopy. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 546(3):pp. 498–508, Jul 2005. 33

[Car70] J. D. Carroll and J. J. Chang. Analysis of individual differences in multidimensional scaling via an n-way generalization of Eckart-Young decomposition. In *Psychometrika*, vol. 35, pp. 283–319. 1970. 43, 55

[Car03] M. Carpenter, M. Melath, S. Zhang, and W. E. Grizzle. Statistical processing and analysis of proteomic and genomic data. In *Proceedings of the Pharmaceutical SAS Users Group*, pp. 545–548. 2003. 21

[Cat66] R. B. Cattell. The scree test for the number of factors. *Multivariate Behavioral Research*, vol. 1(2):pp. 245–276, Apr 1966. 42

[Cha99] P. Chalermwat. *High Performance Automatic Image Registration for Remote Sensing*. Ph.D. thesis, Philosophy at George Mason University, Fairfax, Virginia, 1999. 55, 57, 60

[Che03] G. Chen. *Real-time wavelet compression and self-modeling curve resolution for ion mobility spectrometry*. Ph.D. thesis, Ohio University, Mar 2003. 24

[Chi01] S. Chitroub, A. Houacine, and B. Sansal. Principal component analysis of multispectral images using neural network. In *ACS/IEEE International Conference on Computer Systems and Applications*, pp. 89–95. 2001. 25

[Cly06] M. A. Clyde, L. L. House, and R. L. Wolpert. *Nonparametric Models for Proteomic Peak Identification and Quantification*, chap. 15, pp. 293–308. Cambridge University Press, 2006. 20

[Coo07] K. R. Coombes, K. A. Baggerly, and J. S. Morris. *Pre-Processing Mass Spectrometry Data*, chap. 4, pp. 79–99. Kluwer, 2007. 23

[Dav07] R. A. Davis, A. J. Charlton, J. Godward, S. A. Jones, M. Harrison, and J. C. Wilson. Adaptive binning: An improved binning method for metabolomics data using the undecimated wavelet transform. *Chemometrics and Intelligent Laboratory Systems*, vol. 85:pp. 144–154, 2007. 22

[Dju02] S. Djurcilov, K. Kim, P. Lermusiaux, and A. Pang. Visualizing scalar volumetric data with uncertainty. *Computers and Graphics*, vol. 26(2):pp. 239–248, Apr 2002. 73

[Dre88] R. A. Drebin, L. Carpenter, and P. Hanrahan. Volume rendering. In J. Dill and J. Dill, editors, *ACM SIGGRAPH Computer Graphics*, vol. 22, pp. 65–74. Aug 1988. 73

[Dro03] I. Drori and D. Lischinski. Fast multiresolution image operations in the wavelet domain. *IEEE Transactions on visualization and computer graphics*, vol. 9(3):pp. 395–411, 2003. 20

[Dry98] I. L. Dryden and K. V. Mardia. *Statistical shape analysis*. John Wiley and Sons, Jul 1998. 67

[Ery07] O. L. Eryilmaz and A. Erdemir. Investigation of initial and steady-state sliding behavior of a nearly frictionless carbon film by imaging 2- and 3-D TOF-SIMS. *Tribology Letters*, vol. 28(3):pp. 1023–8883, Dec 2007. 109

[Fle06] J. S. Fletcher, A. Henderson, R. M. Jarvis, N. P. Lockyer, J. C. Vickerman, and R. Goodacre. Rapid discrimination of the causal agents of urinary tract infection using ToF-SIMS with chemometric cluster analysis. *Applied Surface Science*, vol. 252(19):pp. 6869–6874, 2006. 24

[Fon97] L. Fonseca and M. Costa. Automatic registration of satellite images. In *Brazilian Symposium on Graphic Computation and Image Processing*, pp. 219–226. 1997. 56

[Fra06] J. Fransens and F. van Reeth. Hierarchical PCA decomposition of point clouds. In *Third International Symposium on 3D Data Processing, Visualization, and Transmission*, pp. 591–598. 2006. 86

[Fuh02] A. Fuhrmann, B. Özer, L. Mroz, and H. Hauser. VR2 interactive volume rendering using PC-based virtual reality. Tech. Rep. 14, TR-VRVis, Mar 2002. 73

[Geb98] M. S. klein Gebbinck. *Decomposition of mixed pixels in remote sensing images to improve the area estimation of agricultural fields*. Ph.D. thesis, Katholieke Universiteit Nijmegen, Nov 1998. 14

[Geu02] J. M. Geusebroek, A. W. M. Smeulders, and J. van de Weijer. Fast anisotropic gauss filtering. In *European Conference on Computer Vision, Copenhagen, Denmark, Part I*, vol. 2350, pp. 99–112. Springer Berlin / Heidelberg, May 2002. 101

[Gon03] R. Gonzalez, R. Woods, and S. Eddins. *Digital Image Processing Using MATLAB*. Prentice Hall, 2003. 57

[Gra06] D. J. Graham, M. S. Wagner, and D. G. Castner. Information from complexity: Challenges of tof-sims data interpretation. *Applied Surface Science*, vol. 252(19):pp. 6860–6868, 2006. 3

[Gut54] L. Guttman. Some necessary conditions for common-factor analysis. *Psychometrika*, vol. 19:pp. 149–162, 1954. 42

[Hai97] S. Haigh, P. G. Kenny, R. H. Roberts, I. R. Barkshire, M. Prutton, D. K. Skinner, P. Pearson, and K. Stribley. Automatic and interactive correlation partitioning compared : Application to TiN/ Ti/ SiO. *Surface and Interface Analysis*, vol. 25:pp. 335–340, 1997. 98

[Har70] R. A. Harshman. Foundations of the PARAFAC procedure: Models and conditions for an explanatory multimodal factor analysis. In *UCLA Working Papers in Phonetics*, vol. 16, pp. 1–84. 1970. 43, 55, 94

[Har84] R. A. Harshman. *How can I know if it's real? A catalogue of diagnostics for use with three-mode factor analysis and multidimensional scaling*, chap. Appendix A, pp. 566–591. Praeger, 1984. 11

[Här03] W. Härdle and L. Simar. *Applied multivariate statistical analysis*. Springer-Verlag, 2003. 23

[He96] T. He, L. Hong, A. Kaufman, and H. Pfister. Generation of transfer functions with stochastic search techniques. In R. Yagel and G. Nielson, editors, *IEEE Visualization*, pp. 227–234. 1996. 73

[Hil06] M. Hilario, A. Kalousis, C. Pellegrini, and M. Mu. Processing and classification of protein mass spectra. *Mass Spectrometry reviews*, vol. 25:pp. 409–449, Feb 2006. 18, 23

[Hut96] H. Hutter, C. Brunner, S. Nikolov, C. Mittermayer, and M. Grasserbauer. Imaging surface spectroscopy for two- and three-dimensional characterization of materials. *Fresenius' Journal of Analytical Chemistry*, vol. 355(5-6):pp. 585–590, Jun 1996. 25, 87

[Iba03] L. Ibanez and W. Schroeder. *ITK software guide*, chap. 8, pp. 215–313. Kitware, Inc., Aug 2003. 55

[Jac05] N. P. Jacobson and M. R. Gupta. Design goals and solutions for display of hyperspectral images. In *IEEE Transactions on Geoscience and Remote Sensing*, vol. 43, pp. 2684–2692. 2005. 98

[Jol02] I. T. Jolliffe. *Principal Component Analysis*. Springer series in statistics. Springer-Verlag, second edn., 2002. 99

[Kaa01] A. Kaarna and J. Parkkinen. Transform based lossy compression of multispectral images. *Pattern Analysis & Applications*, vol. 4(1):pp. 39–50, Mar 2001. 29

[Kai58] H. F. Kaiser. The varimax criterion for analytic rotation in factor analysis. In *Psychometrika*, vol. 23, pp. 187–200. 1958. 43

[Kal05] A. Kalaiah and A. Varshney. Statistical geometry representation for efficient transmission and rendering. *ACM Transactions on Graphics*, vol. 21(2):pp. 348–373, Apr 2005. 86

[Kar88] M. Karas and F. Hillenkamp. Laser desorption ionization of proteins with molecular masses exceeding 10 000 daltons. *Analytical chemistry*, vol. 60(20):pp. 2299–2301, 1988. 14

[Kee04a] M. R. Keenan and P. G. Kotula. Accounting for Poisson noise in the multivariate analysis of ToF-SIMS spectrum images. *Surface and Interface Analysis*, vol. 36:pp. 203–212, 2004. 41, 87

[Kee04b] M. R. Keenan and P. G. Kotula. Optimal scaling of TOF-SIMS spectrum-images prior to multivariate statistical analysis. *Applied Surface Science*, vol. 231-232:pp. 240–244, 2004. 98, 99

[Kee05] M. R. Keenan. Maximum likelihood principal component analysis of time-of-flight secondary ion mass spectrometry spectral images. *Journal of Vacuum Science & Technology A: Vacuum, Surfaces, and Films*, vol. 23(4):pp. 746–750, Jul 2005. 38, 97

[Ken97] P. G. Kenny, P. F. Newbury, D. L. Mountain, D. Whitehouse, S. A. Haigh, M. Prutton, R. H. Roberts, I. R. Barkshire, and M. J. G. Wenham. Compression, visualization and segmentation techniques for 3D spectrum-images from multispectral analytical electron microscopy. In H. R. Arabnia, editor, *International Conference on Imaging Science, Systems and Technology*, pp. 355–363. CSREA, 1997. 33, 97

[Kes03] N. Keshava. A survey of spectral unmixing algorithms. *Lincoln Laboratory Journal*, vol. 14(1):pp. 55–78, 2003. 12, 18, 24

[Kie91] H. A. L. Kiers. Hierarchical relations among three-way methods. *Psychometrika*, vol. 56(3):pp. 449–470, 1991. 27, 44, 94

[Kie94] H. A. L. Kiers. Simplimax: Oblique rotation to an optimal target with simple structure. *Psychometrika*, vol. 59(4):pp. 567–579, Dec 1994. 43

[Kin98] G. Kindlmann and J. Durkin. Semi-automatic generation of transfer functions for direct volume rendering. In *IEEE Symposium on Volume Visualization*, pp. 79–86. 1998. 73

[Kle07] L. A. Klerk, A. Broersen, I. W. Fletcher, R. van Liere, and R. M. A. Heeren. Extended data analysis strategies for high resolution imaging MS:new methods to deal with extremely large image hyperspectral datasets. *International Journal of Mass Spectrometry*, vol. 260(2-3):pp. 222–236, Feb 2007. 9, 94, 98

[Kni01a] J. Kniss, G. Kindlmann, and C. Hansen. Interactive volume rendering using multi-dimensional transfer functions and direct manipulation widgets. In *IEEE Visualization*, pp. 255–262. Oct 2001. 73

[Kni01b] J. Kniss, G. Kindlmann, and C. Hansen. Multi-dimensional transfer functions for interactive volume rendering. *IEEE Transactions on Visualization and Computer Graphics*, vol. 8(4):pp. 270–285, Jul 2001. 73

[Koe84] J. J. Koenderink. The structure of images. *Biological Cybernetics*, vol. 50(5):pp. 363–370, Aug 1984. 100

[Lan00] D. Landgrebe. *Information Processing for Remote Sensing*, chap. Information Extraction Principles and Methods for Multispectral and Hyperspectral Image Data, pp. 2–30. World Scientific Publishing Co., Inc., 2000. 30

[Las98] P. Lasch, W. Wäsche, W. J. McCarthy, and D. Naumann. Imaging of human colon carcinoma thin sections by FT-IR microspectrometry. In *Infrared Spectroscopy: New Tool in Medicine*, vol. 3257, pp. 187–197. 1998. 41, 73

[Lee08] J. L. S. Lee, I. S. Gilmore, and M. P. Seah. Quantification and methodology issues in multivariate analysis of ToF-SIMS data for mixed organic systems. *Surface and Interface Analysis*, vol. 40(1):pp. 1–14, Jan 2008. 25, 109

[Leh96] R. B. Lehoucq and D. C. Sorensen. Deflation techniques for an implicitly restarted arnoldi iteration. *Siam Journal on Matrix Analysis and Applications*, vol. 17:pp. 789–821, 1996. 49

[Len04] R. Lenz and T. H. Bui. Recognition of non-negative patterns. In *Proceedings International Conference on Pattern Recognition*, vol. 3, pp. 498–501. Aug 2004. 42

[Lev88] M. Levoy. Display of surfaces from volume data. *IEEE Comput Graph Appl*, vol. 8(3):pp. 29–37, 1988. 73

[Lev05] I. W. Levin and R. Bhargava. Fourier transform infrared vibrational spectroscopic imaging: Integrating microscopy and molecular recognition. *Annual review of physical chemistry*, vol. 56:pp. 429–474, 2005. 14

[Lho01] J. B. Lhoest, M. S. Wagner, C. D. Tidwell, and D. G. Castner. Characterization of adsorbed protein films by time of flight secondary ion mass spectrometry. *Journal of Biomedical Materials Research*, vol. 57(3):pp. 432–440, 2001. 30

[Lin05] L. Linsen, J. Locherbach, M. Berth, J. Bernhardt, and D. Becher. Differential protein expression analysis via liquid-chromatography/mass-spectrometry data visualization. In *IEEE Visualization*, pp. 447–454. Oct 2005. 71

[Lis05] J. Listgarten and A. Emili. Statistical and computational methods for comparative proteomic profiling using liquid chromatography-tandem mass spectrometry. *Molecular & Cellular Proteomics*, vol. 4:pp. 419–434, 2005. 17

[Loh98] M. T. Lohnes. *Multivariate approaches to qualitative and quantitative analysis in chemistry.* Master's thesis, Daihousie University, Jun 1998. 12

[Maa07] L. J. P. van der Maaten, E. O. Postma, and H. J. van den Herik. Dimensionality reduction: A comparative review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. draft:pp. 1–35, 2007. 25, 29

[Mar05] K. Martin, L. Ibáñez, L. Avila, S. Barré, and J. Kaspersen. Integrating segmentation methods from the insight toolkit into a visualization application. *Medical Image Analysis*, vol. 9(6):pp. 579–593, 2005. 77

[McC05] G. McCombie, D. Staab, M. Stoeckli, and R. Knochenmuss. Spatial and spectral correlations in MALDI mass spectrometry images by clustering and multivariate analysis. *Analytical chemistry*, vol. 77(19):pp. 6118–6124, Oct 2005. 86

[McD03] L. A. McDonnell, T. H. Mize, S. L. Luxembourg, S. Koster, and G. B. Eijkel. Using matrix peaks to map topography: Increased mass resolution and enhanced sensitivity in chemical imaging. *Analytical Chemistry*, vol. 75(17):pp. 4373–4381, 2003. 32, 68, 108

[McD05] L. A. McDonnell, S. R. Piersma, A. F. M. Altelaar, T. H. Mize, P. D. E. M. Verhaert, J. van Minnen, and R. M. A. Heeren. Matrix-enhanced secondary ion mass spectrometry imaging of brain tissue. *Journal of Mass Spectrometry*, vol. 40:pp. 160–168, 2005. 46

[McD07] L. A. McDonnell and R. M. A. Heeren. Imaging mass spectrometry. *Mass Spectrometry Reviews*, vol. 26(4):pp. 606–643, Aug 2007. 2, 14

[Moi02] J. le Moigne. Multi-sensor image registration, fusion and dimension reduction. *Online journal of space communication*, vol. 3, 2002. 55

[Mor05] J. S. Morris, K. R. Coombes, J. Koomen, K. A. Baggerly, and R. Kobayashi. Feature extraction and quantification for mass spectrometry in biomedical applications using the mean spectrum. *Bioinformatics*, vol. 21(9):pp. 1764–1775, Feb 2005. 23

[Mül06] W. Müller, T. Nocke, and H. Schumann. Enhancing the visualization process with principal component analysis to support the exploration of trends. In *Asia-Pacific Symposium on Information Visualisation*, vol. 60 of *ACM International Conference Proceeding Series*, pp. 121–130. 2006. 30

[Mur00] S. Muraki, T. Nakai, and Y. Kita. Basic research for coloring multichannel MRI data. In *Proceedings of the 11th IEEE Visualization 2000 Conference*, pp. 187–194. IEEE Computer Society, 2000. 73

[Nas06] J. M. P. Nascimento. *Unsupervised Hyperspectral Unmixing*. Ph.D. thesis, Universidade Tecnica de Lisboa, 2006. 25

[Nus81] H. J. Nussbaumer. *Fast fourier transform and convolution algorithms*. Springer series in information sciences. Springer-Verlag Berlin / Heidelberg / New York, 1981. 101

[Ooi06] W. S. Ooi and C. P. Lim. Fuzzy clustering of color and texture features for image segmentation: A study on satellite image retrieval. *Journal of Intelligent and Fuzzy Systems*, vol. 17(3):pp. 297–311, 2006. 57

[Pac99] M. L. Pacholski and N. Winograd. Imaging with mass spectrometry. *Chemical Reviews*, vol. 99:pp. 2977–3005, 1999. 15

[Pac04] M. L. Pacholski. Principal component analysis of TOF-SIMS spectra, images and depth profiles: an industrial perspective. *Applied Surface Science*, vol. 231-232:pp. 235–239, 2004. 86, 98

[Pau06] V. P. Pauca, P. Piper, and R. J. Plemmons. Nonnegative matrix factorization for spectral data analysis. *Linear Algebra and its Applications*, vol. 416(1):pp. 29–47, Jul 2006. 28

[Pfi01] H. Pfister, W. Lorensen, C. Bajaj, G. Kindlmann, W. Schroeder, L. Sobeierajski Avila, K. Martin, R. Machiraju, and J. Lee. The transfer function bake-off. In *IEEE Computer Graphics and Applications*, pp. 16–22. May 2001. 73

[Piw01] J. M. Piwowar, C. P. Derksen, and E. F. LeDrew. Principal components analysis of the variability of Northern Hemisphere sea ice concentrations: 1979-1999. In *Proceedings, 23rd Canadian Symposium on Remote Sensing / 10e Congrès de L'Association québécoise de télédétection, Ste.-Foy PQ*, pp. 619–628. Aug 2001. 73

[Pla04] A. Plaza, P. Martínez, R. Pérez, and J. Plaza. A quantitative and comparative analysis of endmember extraction algorithms from hyperspectral data. *IEEE Transactions on geoscience and remote sensing*, vol. 42(3):pp. 650–663, Mar 2004. 29

[Pla06] C. Plant, M. Osl, B. Tilg, and C. Baumgartner. Feature selection on high throughput SELDI-TOF mass-spectrometry data for identifying biomarker candidates in ovarian and prostate cancer. In *Proceedings of the Sixth IEEE International Conference on Data Mining - Workshops*. 2006. 20

[Pol01] G. Polder and G. W. A. M. van der Heijden. Visualization of spectral images. In Y. Censor and M. Ding, editors, *Visualization and Optimization Techniques*, vol. 4553, pp. 132–137. 2001. 73, 98

[Pra91] W. K. Pratt. *Digital Image Processing*. John Wiley and Sons, 2nd edn., 1991. 57

[Pre02] K. J. Preacher and R. C. MacCallum. Exploratory factor analysis in behavior genetics research: factor recovery with small sample sizes. *Behavior Genetics*, vol. 32(2):pp. 153–161, Mar 2002. 28

[Pru99] M. Prutton, D. K. Wilkinson, P. G. Kenny, and D. L. Mountain. Data processing for spectrum-images: extracting information from the data mountain. In *Applied Surface Science*, pp. 1–10. Elsevier, 1999. 32

[Ran05a] T. W. Randolph and Y. Yasui. Multiscale processing of mass spectrometry data. *Biometrics*, vol. 62(2):pp. 589–597, Jun 2005. 21, 23, 87

[Ran05b] V. Rankov, R. Locke, R. Edens, P. Barber, and B. Vojnovic. An algorithm for image stitching and blending. In J. Conchello, C. Cogswell, and T. Wilson, editors, *Proceedings of SPIE*, vol. 5701, pp. 190–199. Dec 2005. 56

[Ric07] K. Richter. *Application of imaging TOF-SIMS in cell and tissue research*. Ph.D. thesis, Göteborg University, 2007. 15

[Rit04] M. Ritter, H. Hutter, and M. Grasserbauer. Maximum entropy deconvolution of secondary ion mass spectra with a measured response. *Fresenius' Journal of Analytical Chemistry*, vol. 349(1-3):pp. 186–190, Dec 2004. 20

[San02] O. D. Sanni, M. S. Wagner, D. Briggs, D. G. Castner, and J. C. Vickerman. Classification of adsorbed protein static ToF-SIMS spectra by principal component analysis and neural networks. *Surface and Interface Analysis*, vol. 33:pp. 715–728, 2002. 25

[San04] S. dos Santos and K. Brodlie. Gaining understanding of multivariate and multidimensional data through visualization. *Computers & Graphics*, vol. 28(3):pp. 311–325, Jun 2004. 7

[Sca93] I. Scarminio and M. Kubista. Analysis of correlated spectral data. *Analytical Chemistry*, vol. 65(4):pp. 409–418, 1993. 45

[Sme04] V. S. Smentkowski, J. A. Ohlhausen, P. G. Kotula, and M. Keenan. Multivariate statistical analysis of time-of-flight secondary ion mass spectrometry images: looking beyond the obvious. *Applied Surface Science*, vol. 231/232:pp. 245–249, 2004. 98, 106

[Sme07] V. S. Smentkowski, S. G. Ostrowski, E. Braunstein, M. R. Keenan, J. A. Ohlhausen, and P. G. Kotula. Multivariate statistical analysis of three-spatial-dimension TOF-SIMS raw data sets. *Analytical Chemistry*, vol. 79:pp. 7719–7726, 2007. 33

[Sta98] J. L. Starck, F. Murtagh, and A. Bijaoui. *Image Processing and Data Analysis: The Multiscale Approach*. Cambridge University Press, 1998. 86

[Tim01] M. E. Timmerman. *Component analysis of multisubject multivariate longitudinal data*. Ph.D. thesis, University of Groningen, 2001. 43

[Tor89] J. M. Torson. Interactive image cube visualization and analysis. In *Proceedings of the 1989 Chapel Hill workshop on Volume visualization*, pp. 33–38. ACM Press, 1989. 73

[Tra05] T. N. Tran. *Including Spatial Information in Clustering of Multi-Channel Images*. Ph.D. thesis, Radboud Universiteit Nijmegen, Nov 2005. 24

[Tyl03] B. Tyler. Interpretation of TOF-SIMS images: multivariate and univariate approaches to image de-noising, image segmentation and compound identification. *Applied Surface Science*, vol. 203-204:pp. 825–831, 2003. 86

[Tyl06] B. Tyler. Multivariate statistical image processing for molecular specific imaging in organic and bio-systems. *Applied Surface Science*, vol. 252(19):pp. 6875–6882, Jul 2006. 27

[VH04] F. Vega-Higuera, N. Sauber, B. Tomandl, C. Nimsky, G. Greiner, and P. Hastreiter. Automatic adjustment of bidimensional transfer functions for direct volume visualization of intracranial aneurysms. In R. L. Galloway, Jr., editor, *Medical Imaging 2004: Visualization, Image-Guided Procedures and Display*, vol. 5367, pp. 275–284. Jun 2004. 73

[Vic02] J. C. Vickerman and D. Briggs, editors. *ToF-SIMS: Surface Analysis by Mass Spectrometry*. IM Publications and SurfaceSpectra, 2002. 14

[Vog04] F. Vogt, S. Banerji, and K. Booksh. Utilizing three-dimensional wavelet transforms for accelerated evaluation of hyperspectral image cubes. *Journal of Chemometrics*, vol. 18(7–8):pp. 350–362, Feb 2004. 21

[Wag04] M. S. Wagner, D. J. Graham, B. D. Ratner, and D. G. Castner. Maximizing information obtained from secondary ion mass spectra of organic thin films using multivariate analysis. *Surface Science*, vol. 570(1-2):pp. 78–97, Oct 2004. 28

[Wal03] M. E. Wall, A. Rechtsteiner, and L. M. Rocha. *Singular Value Decomposition and Principal Component Analysis*, chap. 5, pp. 91–109. Kluwer:Norwell, MA, 2003. 41, 73

[Wee02] J. van der Weerd. *Microspectroscopic analysis of traditional oil paint*. Ph.D. thesis, University of Amsterdam, Dec 2002. 14

[Wei96] A. Weingessel, H. Bischof, and K. Hornik. Hierarchies of autoassociators. In *Proceedings of the 13th International Conference on Pattern Recognition*, vol. 4, pp. 200–204. 1996. 86

[Wic03] B. T. Wickes, Y. Kim, and D. G. Castner. Denoising and multivariate analysis of time-of-flight SIMS images. *Surface and Interface Analysis*, vol. 35:pp. 640–648, 2003. 22, 68, 86, 98

[Wit83] A. P. Witkin. Scale-space filtering. In *International Joint Conference on Artificial Intelligence*, pp. 1019–1021. 1983. 100

[Wol97] M. Wolkenstein, H. Hutter, S. G. Nikolov, and M. Grasserbauer. Improvement of SIMS image classification by means of wavelet de-noising. *Fresenius' Journal of Analytical Chemistry*, vol. 357(7):pp. 783–788, Mar 1997. 21, 86

[Wol99] M. Wolkenstein, T. Stubbings, and H. Hutter. Robust automated three-dimensional segmentation of secondary ion mass spectrometry image sets. *Fresenius' Journal of Analytical Chemistry*, vol. 365(1-3):pp. 63–69, Sep 1999. 21, 97

[Zit03] B. Zitová and J. Flusser. Image registration methods: a survey. In *Image and Vision Computing*, vol. 21, pp. 977–1000. 2003. 56

List of Figures

1.1 A schematic spectral and spatial view. 3
1.2 The dataflow in the visualization pipeline of this approach. 7

2.1 Three stages in a spectral analysis. 13
2.2 A single spectrum and a single image in the spectral datacube. 14
2.3 Schematic representation of an imaging TOF SIMS instrument. 15
2.4 Hierarchical taxonomy to classify methods for feature extraction. . . . 19
2.5 Binning with equal-with bins with size w. 22
2.6 PCA by projecting two correlated variables to a new dimension. . . . 26
2.7 Examples of spectra with raw channel data. 31
2.8 Examples of spectral images from binned channels. 32
2.9 Example of a 3D spectral datacube. 33
2.10 Example of a spectral and spatial feature component. 34

3.1 1D, 2D and 3D views of the synthetic spectral datacube. 45
3.2 Spectral and spatial summation of a spectral datacube. 46
3.3 Spectral and spatial components of three methods for decomposition. . 47

4.1 A microscopic and spectrometric image. 54
4.2 One image component and a spectrally matching image component. . 56
4.3 Shifting two image components to create a MSE landscape. 57
4.4 Different texture measurements applied to the partial windows. 58
4.5 Entropy space and fitted entropy space of all regions. 59
4.6 The relative locations of the crystal and the kneecap dataset. 61
4.7 Mean Squares landscapes from the first and second component. 62
4.8 $\frac{MS}{entropy}$ search space of the first and second principal component. 62
4.9 Enhanced separability and contrast by PCA using a larger dataset. . . 65
4.10 MSE landscape for a registration using a intermediate result. 66

5.1 The summation of all spectra with the TIC image of the pond snail. . 72
5.2 The first score vectors with the accompanying eigenimages. 76
5.3 The spectral datacube with a resulting transfer function. 77
5.4 The second, third and fourth component mappings. 78
5.5 Two component mappings compared with the VolView representation. 78

5.6 A representation of the spectral datacube of the embedded hair. . . . 79
5.7 A representation of the spectral datacube of the PVP droplet series. . 80

6.1 The summed spectral and spatial view of a chicken embryo. 84
6.2 Traditionally extracted image component. 84
6.3 The data-flow in visualization pipeline of the approach for zooming. . 87
6.4 Similar extracted components with enhanced contrast. 88
6.5 Spectra with a different number of variables in each bin. 89
6.6 Region selection by applying a threshold on image components. 90
6.7 Different extracted cholesterol image components. 91
6.8 Improved selectivity and contrast in an image component. 92
6.9 A low resolution overview of the components after PCA. 93

7.1 Spectral and spatial summation of a spectral datacube. 96
7.2 The spectral and image component of the second component. 97
7.3 Selection of spectral peaks using α. 100
7.4 The iso-surfaces of two features with different values for β. 101
7.5 The iso-surfaces of two features with different values for γ. 102
7.6 The histogram and shape representation of three analytes. 103
7.7 Isolated cholesterol distribution at the highest resolution. 104
7.8 PCA component that separates three peaks. 105

Glossary of Terms

3D opacity map A datavolume in which a certain opacity value is assigned to each cell, page 75.

abundance Proportion in which an element is present compared to all other elements, page 12.

analyte Substance or chemical constituent that is determined in an analytical procedure, page 103.

binning Down-binning combines serveral data-points into one variable or 'bin' without the use of a complex model. The intensity values are added together to create a new variable. A single parameter sets the size of the bin, page 21.

CA Cluster Analysis is mainly applied to imaging spectrometry data to group similar spectra that are within each other's neighborhood, page 24.

chemometrics Refers to multivariate analysis in the context of applications in the chemistry. It comprises the application of multivariate statistics, computational methods and also includes the representation and display of the extracted information, page 17.

convolution Filters or transforms data by applying a local function on several data-points that are located together. The local function (or density function) is based on a local properties of the data distribution instead of a complete model of the data. All data-points are transformed according to this local function. A suitable model can transform the dataset into one with less local noise, page 20.

cost function The cost function describes the difference between a temporal classification and the actual data. This cost function is minimized to be able find the best mapping between data and a resulting classification. A cost function could be complex or simple, depending on the type of input, page 25.

CT Computed Tomography is a medical method for imaging by sections, originally known as the a 'EMI scan' and later as a computed axial tomography (CAT) scan, page 73.

deconvolution Reversed convolution by making an estimation of a point-spread function which describes a local model of (variance in) the data distribution, page 20.

DVR Direct Volume Rendering is a technique to display a 2D projection of a 3D volume in which every voxel has an opacity and a color, page 71.

endmembers Constituent spectra that correspond to known chemical compounds or materials in a spectral image. They are the most elementary building blocks that define a 'pure' material within a spectrum, page 12.

$f(m)$ Function describing a spectrum that assigns a intensity value to each mass m, page 2.

$F(x, y, m)$ Function describing a 3D spectral datacube that assigns a intensity value to each mass m on each location (x, y), page 2.

$F'_m(x, y)$ Function describing an image with a spatial distribution of intensity values of a spectral band m, page 2.

FA Factor Analysis or common factor analysis is a dimension reduction technique in which variables are re-expressed by orthogonal projection on new variables according to their common variability, page 22.

FT-IR Fourier Transform InfraRed imaging spectroscopy is a technique that employs the transmission of infrared light through the surface of a material. Light from the spectrometer illuminates the sample. Reflected or transmitted light is projected on the camera. During each step of the interferometer, the camera acquires and averages 200 images to gain signal-to-noise. This way, an interferogram is obtained for each of the 4096 camera pixels. These can be Fourier transformed resulting in 4096 infrared spectra. Using the IR spectrum, chemical bonds and the molecular structure of organic compounds can be identified. Surface areas as small as 10-15 microns can be detected, page 14.

GB GigaByte means either exactly one billion bytes or approximately 1.07 billion bytes, page 54.

Hermitian Square matrix with complex entries which is equal to its own conjugate transpose, page 49.

isotropic Uniformity in all directions, for instance when an image is circularly symmetric in 2D, page 100.

landmark Significant and easily identifiable point of correspondence on each object that matches between and within populations and is useful to represent shape changes, page 67.

LDI Laser Desorption and Ionization time of flight imaging is a soft ionization technique in mass spectrometry, page 50.

matrix material Crystallized molecules that surround and support an embedded analyte material to protect it from being destroyed by the laser beam and to facilitate vaporization and ionization in chemical analysis, page 16.

MCA Multiple Correspondence Analysis is a generalized approach that assigns distances between nominal or ordinal data measurements. It utilizes a model, for instance chi-square, to base it distance functions on. It provides tools for analyzing the associations in contingency tables. Simple indices are assigned to each pair of row and column variables to show the relation between the both of them, page 27.

MCR Multivariate Curve Resolution is a group of techniques which intend the recovery of the response profiles (spectra, pH profiles, time profiles, elution profiles, etc.) of more than one component in an unresolved and unknown mixture when no prior or little information is available about the nature and composition of these mixtures, page 27.

MDS Multi-Dimensional Scaling visualizes distances or similarities between variables in a graphical map, page 34.

MRI Magnetic Resonance Imaging is a technique used in radiology to visualize the structure and function of the body using a magnetic field, page 73.

MSE Mean Squared Error is a pixel-wise similarity measure between two images, page 56.

multicollinearity Refers to a situation in which two or more explanatory variables in a multiple regression model are highly correlated, page 18.

NCC Normalized Cross-Correlation is a pixel-wise similarity measure between two images of the same modality, page 57.

NN Neural Networks can be designed to perform all tasks of filtering, selecting and create a classification. Before this, a neural network needs training data, a number of output variables and an objective or optimization function, page 25.

noise Any unwanted signal interfering with the clarity and intelligibility of desired signals, page 16.

PARAFAC Parallel Factor Analysis is a multiway decomposition method and a generalization of PCA on higher order arrays. The number of components in a model have to be estimated and there is a residual term with the difference between the model and the measured data, page 28.

PCA Principal Component Analysis is a dimension reduction technique in which variables are re-expressed by orthogonal projection on new variables according to the total variance, page 23.

PP Projection Pursuit maximizes a chosen projection index, for example the normal distribution, the entropy, kurtosis etc., page 28.

qualitative view Focuses on classifying a sample in an analysis as either having a specific chemical attribute or not, page 12.

quantitative view Focuses on modeling a measured response in an analysis based on the amounts, concentrations, or other physical or chemical properties, page 12.

RGB Red, Green, Blue describes a device-dependent color space that can reproduce a range of colors when these three additive primary colors added together, page 2.

S/N Signal-to-Noise ratio is the power ratio between a signal and the level of background noise corrupting the signal, page 16.

datacube Spectral datacubes are used in the field of imaging spectrometry, in which a spectrally-resolved image dataset can be represented as a 3D volume, page 2.

SVD Singular Value Decomposition is a factorization method that results in a diagonal matrix ande two unitary matrices, page 27.

TIC Total Ion Count refers to an image with all ions counts summed to create a spatial distribution of intensity values, page 32.

TM Thematic Mapper is a sensor observing the earth in the Landsat program. These sensors have seven spectral bands: three in visable wavelenghts and four in infrared most of which have a spatial resolution of 30 meter, page 14.

TOF SIMS Time-of-Flight Secondary Ions Mass Spectrometry is a method in which a surface is bombarded with a primary, pulsed beam of ions. Emitted ions are accelerated and imaged with sophisticated ion optics on a detector. This results in the emission of secondary ions, which are accelerated and subsequently mass analyzed to generate surface mass spectra. Secondary ions with different masses and charges have different velocities and are separated based on their flight time through the system. The technique therefore provides very detailed elemental and chemical structure information. Using the latest TOF SIMS instruments, surface areas as small as 3-5 microns can be detected, page 14.

u Unified atomic mass unit, or Dalton (Da) or, universal mass unit, is an unit of mass used to express atomic and molecular masses. It is the approximate mass of a hydrogen atom, a proton, or a neutron, page 2.

X_{mn} Matrix with the coefficients of the pure spectral profiles in the columns, where m are the spectral variables and n the number of distinct compounds, page 12.

Index

Symbols

$F'_m(x, y)$ 2, 13
$F(x, y, m)$ 2, 13
$F[x, y, m]$ 15, 25
P 13, 26, 74, 99
X_{mn} 12, 25, 26
Y 13, 26, 41, 43, 44, 74, 75, 99
$f(m)$ 2, 13
$f[m]$ 12
1D 19, 97
2D 19, 38, 39, 71, 93, 97
3D 2, 32, 72, 74, 97, 108

A

α 96
analysis
  cluster 24
  correspondence 23
  factor 22, 24, 88
  independent component 25, 73
  independent factor 25
  linear discriminant 24
  multiple correspondence 27
  principal component 23
analyte 103, 104
AVIRIS 14
AXSIA 98

B

baseline correction 40
β 97
bin 86, 87, 95
binning 21, 39, 107
bucketing . . . . . *see* binning

C

calibration 20
Cattell's scree test 42
chemical compound 1, 2, 97
chemometrics 17
chicken embryo 83
collinear . . . . . *see* multicollinearity, 26
color 32, 71
  band 2
components 6
computation 48
  memory 40, 49
  time 50
constraint 18, 27, 42, 44, 51
convolution 20
  kernel 100
CT 73, 97

D

decomposition 43
  eigenvector 41
  PARAFAC 28, 43, 79, 94
  singular value 27
deconvolution 20, 87
dimension reduction
  definition 3
direct volume rendering 71
distribution
  χ^2 . . . . . 23, 27
  Poisson 16, 41
divide and conquer 67, 85
down-binning . . . . . *see* binning

E

eigenimages 42, 74, 92

eigenspectra 41, 74
entropy 55
ε ..... *see* Root Mean Squared Error

F
factor analysis ..... *see* analysis
factors 22, 27, 79
feature 4–6, 18, 37
  definition 4
  space 3
Fourier transform
  discrete 101
FT-IR 14, 74, 76

G
γ 97
Gaussian ..... *see* distribution
  kernel 100
Guttman-Kaiser criterion 42

H
height map 68
Hermitian
  non 49
histogram 57

I
image
  abundance 12, 60
  pyramids 87
  TIC 32, 53, 54, 97, 108
iso-surface 73, 101
iso-value 102

L
landmarks 67
LDI 50
Lymnaea Stagnalis 46, 75

M
mass spectrum 1, 14
mass-to-charge ratio 39
matrix
  diagonal 42
  identity 42
  loading 26, 27, 42
  score 26, 42, 48
matrix material 16, 46
maximum likelihood 27
MDS 34
mean squared error 56
mean-centering 40, 74
memory 49
MRI 73, 97
multicollinearity 18, 28
multivariate curve resolution 27
mutual information 58

N
neural network 25, 86
noise 16
  calibration 16
  chemical 16
  electrical 16
  Gaussian 44
  shot 16
  signal-to-noise ratio 16, 88
normalized cross-correlation 57

O
opacity map 75

P
PARAFAC ..... *see* decomposition
partial least squares 28
PCA 6, 26, 55, 88, 107
  pyramids 86
performance 48
pipeline 7, 89
Poisson ..... *see* distribution
preprocessing 19

R
raw channel 31, 39, 86, 87, 107
registration 17
regression
  multiple lineair 28
  principal component 28
RGB 2
RMSE ..... *see* Root Mean Squared Error
Root Mean Squared Error 45

S
S/N ..... *see* signal-to-noise ratio
scaling 40, 74
signal-to-noise ratio ... 31, 39, 48, 67, 107

SIMS 15, 53
singular value decomposition 44
spectral
    band 2, 13
    channel 2, 15, 73
    datacube 2, 14
    endmember 12, 18, 27
    image 2
    peak 1, 96
    profile 2, 12, 72, 73, 107
    signal 4
    tiles 53
spectrometry 13
    imaging 2
    MALDI 14, 40
    mass 1
    TOF SIMS 14, 40, 50, 60, 75, 90
spectroscopy . . . . . 13, *see* FT-IR

T
TIC 46
time-of-flight 46, 85
TM 14
transform
    Fourier 20
    Hotelling 41
    Karhunen-Loève 41

V
VolView 77
voxel 71

W
wavelet
    compression 21
    discrete transform 20

Summary

Imaging mass spectrometry is an innovative technique that combines high-resolution microscopic imaging tools with analytical capabilities of spectrometry. It is a powerful tool to determine the spatial distribution of chemical compounds on complex surfaces, for example, for microscale analysis of cells and tissue in biological samples. The result is a large spectral datacube: a three-dimensional (3D) dataset in which surface position and mass spectral distribution are represented. Analysts try to discover 'features': correlations in spectral profiles with a recognizable spatial distribution. Techniques for feature extraction and visualization are developed to improve the exploratory analysis of spectral datacubes. The topic of this work is the design and implementation of feature extraction and visualization techniques in high-resolution imaging spectrometry data. Principal Component Analysis (PCA) is interactively used as a governing approach for feature detection. A wide range of visualization techniques are implemented based on extracted features.

The thesis is organized as follows. In Chapter 2 (*Spectral analysis: a survey*), we provide a brief background survey on spectral analysis. The analysis in the proposed approach is divided into three stages: data acquisition, feature extraction and feature visualization. For each stage, a detailed description of currently applied methods is given. The methods most appropriate for this qualitative approach of analysis are chosen as a specific subset. PCA in combination with a binning function is most suited for extracting features from imaging mass spectral data. Both methods increase the signal-to-noise ratio and reduce the amount of data from imaging mass spectrometry.

Chapter 3 (*PCA-based feature extraction*) compares the quality of three different PCA-based methods for detecting and extracting features from spectral datacubes. We discuss preprocessing of mass spectral data, PCA, additional rotational optimization by VARIMAX, and the PARAFAC method for factor regression. The results are compared quantitatively and qualitatively, together with some performance characteristics. For the quantitative comparison, we used a RMSE metric to compare the methods with ground truth spectra under various noise conditions. For the qualitative comparison, we used three criteria to judge the quality of features in the resulting visualizations. These criteria were applied to interpret the visualizations of features.

In Chapter 4 (*Feature-based registration*), a robust method for automatic feature-based registration is developed. The reduction of uncorrelated noise provided by PCA allows high-resolution imaging mass spectrometry datasets to be automatically

aligned and combined for high-resolution analysis of large areas. The results clearly show that the entropy-weighted, mean squared error landscape of chemically matched component images can be used to automatically align high-resolution imaging mass spectrometry datasets. Several spectral datacubes are combined to provide better detection and extraction of features.

In Chapter 5 (*Feature visualization*), a visualization technique is described that utilizes principal components to create transfer functions for volume rendering of a spectral datacube. Two types of spectral datacubes are visualized in 3D by direct volume rendering with these transfer functions to control opacity and highlight extracted features. This enables us to visualize the link between the spectral and spatial characteristics of a feature within the spectral datacube. Applications demonstrate the additional value of these visualizations.

Chapter 6 (*Feature zooming*) presents a technique for spectral and/or spatial zooming of extracted features. This technique is especially useful for spatially extended datasets. The combined spectral datasets are too large in size to be explored and visualized using commonly feature extraction and visualization techniques. Analysts are able to select important features or deselect unimportant features for further analysis on different levels of detail. Moreover, features with unwanted artifacts can be removed to reduce noise.

Chapter 7 (*High-resolution feature visualization*) provides an approach to parametrically visualize features in 3D and at the highest resolution possible. Three parameters control the spectral contribution, the level of detail and the level of density on which an extracted feature is represented. This visualization has feature shapes with well-defined borders and provides more insight into the influences of noise on a mass spectral measurement. It is possible to distinguish different peaks according to their difference in density and spatial position, which would not be possible in a separate spectral or spatial view. An application shows how resulting features are visualized and interpreted.

The developed tools generate new possibilities to handle, explore, and visualize the large imaging mass spectrometry datasets. A sensitive, selective, and robust approach for feature extraction enables detection and classification of features in different proteomics applications. Multiple feature shapes with high-resolution characteristics can be compared and examined on different levels of detail. These visualizations can provide more detailed molecular insight in the biochemistry of surfaces and improve classification of peptides and proteins.

Samenvatting (Dutch summary)

Plaatsopgeloste massa spectrometrie is een innovatieve techniek die gereedschappen voor microscopische beelden met hoge resolutie combineert met de analytische mogelijkheden van spectrometrie. Het is een krachtig middel om de ruimtelijke verdeling te bepalen van chemische componenten op complexe oppervlakten, bijvoorbeeld voor de analyse van cellen en weefsel in biologische specimen. Het resultaat is een grote spectrale datakubus: een driedimensionale (3D) dataverzameling waarin de oppervlaktepositie en de verdeling van spectrale massa zijn vertegenwoordigd. Onderzoekers proberen 'kenmerken' te ontdekken: correlaties in spectrale profielen met een herkenbare ruimtelijke verdeling. Er zijn technieken ontwikkeld voor het extraheren en visualiseren van kenmerken om de verkennende analyse van spectrale datakubussen te verbeteren. Dit werk richt zich op het ontwerpen en implementeren van technieken voor de extractie en visualisatie van kenmerken uit plaatsopgeloste spectrometrische data met hoge resolutie. Principale Componenten Analyse (PCA) wordt interactief gebruikt als een leidende benadering voor het detecteren van kenmerken. Op basis van deze geëxtraheerde kenmerken is een brede reeks visualisatietechnieken geïmplementeerd.

Dit proefschrift is op de volgende manier georganiseerd. In Hoofdstuk 2 (*Spectrale analyse: een overzicht*) wordt een kort overzicht gegeven van de achtergrond van spectrale analyses. In de voorgestelde benadering is de analyse in drie fasen verdeeld: dataverwerving, het extraheren van kenmerken en de visualisatie van kenmerken. De methoden die op dit moment worden toegepast, worden voor elke fase gedetailleerd beschreven. De methoden die het meest toepasselijk zijn voor deze kwalitatieve benadering van een analyse worden als deelverzameling gekozen. PCA in combinatie met een verdelingsfunctie is het meest toepasselijk voor het extraheren van kenmerken uit plaatsopgeloste data met spectrale massa's. Beide methoden vergroten de signaal/ruisverhouding en reduceren de hoeveelheid data afkomstig uit de plaatsopgeloste massa spectrometrie.

Hoofdstuk 3 (*PCA-gebaseerde extractie van kenmerken*) vergelijkt de kwaliteit van drie verschillende PCA-gebaseerde methoden voor het detecteren en extraheren van kenmerken uit spectrale datakubussen. We bespreken het voorbewerken van data met spectrale massa, PCA, extra optimalisatie door rotatie met VARIMAX en de PARAFAC methode voor de regressie in factoren. De resultaten zijn kwantitatief en kwalitatief vergeleken, samen met enkele kenmerken van het prestatievermogen.

We hebben een RMSE-metriek gebruikt voor de kwantitatieve vergelijking van de methoden, waarbij verschillende spectra gebruikt worden waarvan de hoeveelheid ruis bekend is. In de kwalitatieve vergelijking hebben we drie criteria gebruikt om de kwaliteit van de kenmerken in de resulterende visualisaties te beoordelen. Deze criteria zijn toegepast om de visualisaties van de kenmerken te kunnen interpreteren.

In Hoofdstuk 4 (*Registratie gebaseerd op kenmerken*), is een robuuste methode ontwikkeld voor automatische registratie op basis van kenmerken. PCA zorgt voor een afname van niet-gecorreleerde ruis, waardoor plaatsopgeloste dataverzamelingen met spectrale massa automatisch uitgelijnd kunnen worden en gebruikt kunnen worden voor de analyse van grote gebieden op een hoge resolutie. De resultaten laten duidelijk zien dat het landschap bestaande uit gemiddelde kwadraten met een entropieweegfactor, gemaakt door beelden met chemisch gelijke componenten, gebruikt kan worden om plaatsgebonden dataverzamelingen met spectrale massa op een hoge resolutie automatisch uit te lijnen. Meerdere spectrale datakubussen zijn gecombineerd om een betere detectie en extractie van kenmerken te kunnen krijgen.

In Hoofdstuk 5 (*Visualisatie van kenmerken*) wordt een visualisatietechniek beschreven die principale componenten gebruikt om transferfuncties te generen voor het afbeelden van het volume van een spectrale datakubus. Twee typen spectrale datakubussen worden gevisualiseerd in 3D door directe weergave van volumes met deze transferfuncties. Hiermee kan de mate van ondoorschijnendheid worden gecontroleerd en kunnen de uitgelichte kenmerken worden benadrukt. Hierdoor ontstaat de mogelijkheid om de spectrale en ruimtelijke eigenschappen van een kenmerk direct af te beelden in de spectrale datakubus. Applicaties demonstreren de toegevoegde waarde van deze visualisaties.

Hoofdstuk 6 (*Vergroten van kenmerken*) presenteert een techniek voor het spectraal en/of ruimtelijk vergroten van geëxtraheerde kenmerken. Deze techniek is vooral nuttig voor dataverzamelingen die ruimtelijk vergroot zijn. De gecombineerde spectrale dataverzamelingen zijn te groot om ze met de gebruikelijke methoden voor de extractie en visualisatie van kenmerken te kunnen afbeelden en bestuderen. Onderzoekers kunnen belangrijke kenmerken selecteren of onbelangrijke kenmerken schrappen voor verdere analyse op verschillende detailniveaus. Bovendien kunnen kenmerken met ongewenste artefacten verwijderd worden om de ruis te verminderen.

Hoofdstuk 7 (*Visualisatie van kenmerken op hoge resolutie*) stelt voor om kenmerken in 3D te parametrisch te visualiseren op de hoogst mogelijke resolutie. Drie parameters sturen de spectrale bijdrage, het detailniveau en het dichtheidsniveau waarop een geëxtraheerd kenmerk is weergegeven. Deze visualisatie geeft kenmerken een vorm met goed gedefinieerde grenzen en geeft meer inzicht in de invloeden van ruis op een meting met spectrale massa's. Het is mogelijk om verschillende pieken te onderscheiden aan de hand van het verschil in dichtheid en ruimtelijke positie. Dit zou niet mogelijk zijn in een gescheiden spectrale of ruimtelijke afbeelding. Een applicatie laat zien hoe de resulterende kenmerken worden gevisualiseerd en geïnterpreteerd.

De ontwikkelde middelen genereren nieuwe mogelijkheden om grote plaatsopgeloste dataverzamelingen afkomstig van massa spectrometrie te hanteren, onderzoeken en visualiseren. Een gevoelige, selectieve en robuuste benadering voor de extractie van kenmerken maakt het mogelijk om kenmerken te detecteren en classificeren in verschillende applicaties van proteomics. De vormen van meerdere kenmerken met karakteristieken op hoge resolutie kunnen vergeleken en bekeken worden op verschil-

lende detailniveaus. Deze visualisaties kunnen een gedetailleerder moleculair inzicht geven in de biochemie van oppervlakten en verbeteren de classificatie van de peptiden en proteïnen.

Curriculum Vitae

Alexander Broersen was born in March 1977 in Hoorn, The Netherlands. In June 1995, he received his 'Atheneum' diploma at the 'Regionale Scholengemeenschap' in Enkhuizen. The same year he started studying computer science at the University of Twente in Enschede, where he received his Master of Science degree in September of 2003. His graduation project was performed under the supervision of prof.dr.ir. A. Nijholt, and involved designing a virtual piano playing environment.

In February of the year 2004, he joined the visualization and 3D user interfaces theme of the Center for Mathematics and Computer Science (CWI) in Amsterdam. There he performed the Ph.D. research (in Dutch: *Onderzoeker in Opleiding*) at the CWI as a member of the Virtual Laboratory for e-Science (VL-e) project. Within this project, he collaborated closely with the FOM Institute for Atomic and Molecular Physics (AMOLF) on new visualization techniques and tools for analysis of imaging mass spectrometry data. The work was supervised by prof.dr.ir. R. van Liere and prof.dr. R. M. A. Heeren and resulted in this thesis.

Since November of 2008, he is working as a post-doctoral researcher at the Laboratory for Clinical and Experimental Image Processing at the Leiden University Medical Center (LUMC). His work involves the automatic diagnostic vascular analysis by comparing computed tomography angiography with corresponding intravascular ultrasound datasets.