Hyperspectral Image Analysis for Questioned Historical Documents

by

Patrick Shiel BSc.

NUI MAYNOOTH

Ollscoil na hÉireann Má Nuad

A thesis presented in fulfilment of the requirements for the Degree of a Master Of Science (MSc), in Computer Science.

Supervisor: Dr. John Keating

Department of Computer Science

Faculty of Science

National University of Ireland, Maynooth

Maynooth, Co.Kildare, Ireland

July, 2010

Acknowledgments

I would like to express my sincere thanks to my supervisor, Dr. John Keating, for his constant guidance and insight. A special thank you to Amy, my parents - Joe and Caroline, and to my family and friends for their unwavering support and encouragement throughout.

Thank you to all at An Foras Feasa: The Institue for Research in Irish Historical and Cultural Traditions. This thesis has emanated from research conducted with their financial support.

Contents

List of Figures

Abstract

This thesis describes the application of spectroscopy and hyperspectral image processing to examine historical manuscripts and text. Major activities in palaeographic and manuscript studies include the recovery of illegible or deleted text, the minute analyses of scribal hands, the identification of inks and the segmentation and dating of text. This thesis describes how Hyperspectral Imaging (HSI), applied in a novel manner, can be used to perform quality text recovery, segmentation and dating of historical documents. The non-destructive optical imaging process of Spectroscopy is described in detail and how it can be used to assist historians and document experts in the exemption of aged manuscripts. This non-destructive optical method of analysis can distinguish subtle differences in the reflectance properties of the materials under study. Many historically significant documents from libraries such as the Royal Irish Academy and the Russell Library at the National University of Ireland, Maynooth, have been the selected for study using the hyperspectral imaging technique. Processing techniques have are described for the applications to the study of manuscripts in a poor state of conservation. The research provides a comprehensive overview of Hyperspectral Imaging (HSI) and associated statistical and analytical methods, and also an in-depth investigation of the practical implementation of such methods to aid document analysts. Specifically, we provide results from employing statistical analytical methods including principal component analysis (PCA), independent component analysis (ICA) and both supervised and automatic clustering methods to historically significant manuscripts and text

such as Leabhar na hUidhre, a 12th century Irish text which was subject to part-erasure and rewriting, a 16th Century pastedown cover, and a multi-ink example typical of that found in, for example, late medieval administrative texts such as Gttingen's kundige bok. The purpose of which is to achieve an overall greater insight into the historical context of the document, which includes the recovery or enhancement of faded or illegible text or text lost through fading, staining, overwriting or other forms of erasure. In addition, we demonstrate prospect of distinguishing different ink-types, and furnishing us with details of the manuscript's composition, all of which are refinements, which can be used to answer questions about date and provenance. This process marks a new departure for the study of manuscripts and may provide answer many long-standing questions posed by palaeographers and by scholars in a variety of disciplines. Furthermore, through text retrieval, it holds out the prospect of adding considerably to the existing corpus of texts and to providing very many new research opportunities for coming generations of scholars.

Chapter 1

Introduction

1.1 Document Analysis and Questioned Historical Documents

A document, in the context of this research, can be defined as a written, painted or printed record, presented on paper or parchment of any kind that provides and records the original or official form of any event or thing. Documents of this form have been used historically as the principle artefact of recording and preserving information and at present, written and printed document still remain as a primary method of information storage [SrH01]. As a result of the large quantity of existing documents and continual production of new ones, important questions have been generated as to effective preservation and recovery of these documents and the information which they contain. This has led to the development of new research domains dealing with the digital storage and computational recognition and interpretation of a

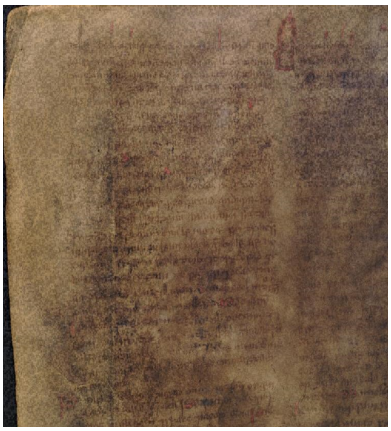
document's core elements, such as images, characters, text, handwriting, etc. These new research areas have also expanded to include the automatic analyses of the overall physical and logical structures of documents. Document Analysis is one such research field which aims at the automatic extraction of information presented on paper and initially addressed to human comprehension [MF08] but has expanded in recent times to include the computational recognition and analysis of documents. The input of any computational document analysis or recognition system is the document itself and the output of the majority of document analysis and recognition systems is a symbolic digital representation of the original document that can be processed by computers. Document analysis refers typically to computer aided analysis of documents to answer particular questions referring to input document. The set of questioned documents exist as a subset of all documents, where the history of the document is unknown or the document is suspected of being fraudulent [Lin06b]. The target document set and primary focus of this research are the set of questioned historical documents; a close relative to the set of questioned documents. A questioned historical document contains many similar properties to that of questioned documents such as an unknown historical context or questioned features in need of examination or extraction. However, there exists a vital difference between a questioned historical document and a questioned document. The validity (i.e. whether the document is fraudulent or not) of historical documents targeted by this research is never in question. A questioned historical document (QHD) in this research then refers to historical documents containing specific difficulty

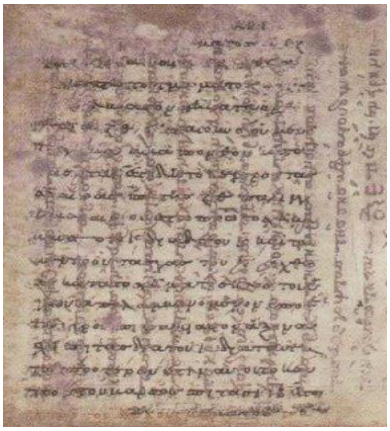
impeding information retrieval, such as document degradation, obscuration or obliteration of text or features, or containing a document specific feature which would benefit from more detailed examination. This research focuses on the application of document analysis systems including document image processing, document models, and signal processing to provide physical and logical analysis of questioned historical documents, with the ultimate objective of a high-level understanding of their semantic content.

1.2 Field Research Problems in Historical Document Examination

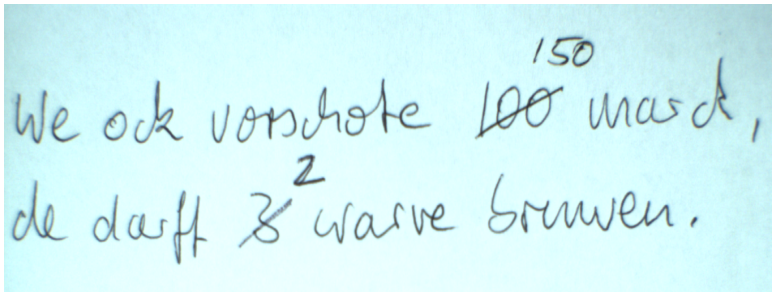
Revealing the entire contents of historical documents is an important aid to scholars that are interested in dating historical documents, establishing origin, or reading older and historically relevant writings they may contain [TSMB04b]. However, due to degradation of historical documents interesting document features are often hidden or barely detectable using conventional analysis methods such as visual inspection or standard imaging techniques. The condition of these historical documents ranges from those that are fully legible to those which can only be read in part, and their legibility is determined by the manner in which they were preserved and treated throughout the ages. In some cases this deterioration of historical documents is caused by processes such as fading over time or staining; in others, the text may have been interfered with in some way in order to edit or revise previous work. Major activities in palaeographic and manuscript

studies include the recovery of illegible or deleted text, the minute analyses of scribal hands, the identification and differentiation of different inks and the segmentation of dating of text. This has a commonality with the engagements been undertaken by questioned document experts trying to identify different inks with a view to scribe attribution, dating of writings, and the detection of amendments or edits to an original text. Providing access to these sometimes fragile historical documents can also present a challange to document archivists who's overall goal can be to preserve the document in its current state. For instance, figure 1.1(a) is an extract from the *Liber Flavus Fergusiorum* [uaryb], a 15th Century manuscript written on vellum by numerous scribes containing mostly religious content. The manuscript is in fairly good condition overall but with some details faded, over time and due to a previous old fashioned chemical attempt to read faded text by a process of staining the text. As a result of this staining the legibility of the text is increased momentarily, and then followed by a quick depression in the clarity of the text. Figure 1.1(c) is a handwritten simulation of a 15th Century manuscript (titled "kundige bok" [Rehng]) containing a legal text that is characterised by many revisions over a period of 50 years. Different layers of amendments and edits represent the various stages and the development of the town law over the time. Segmentation of these layers is crucial for understanding the text, aiding in the association of different scribes to sections of the text, and for the possible relative dating which can be imperative in historical studies. Palimpsests documents, which are documents that have been written more than once with the earlier writing incompletely erased,





(a) Faded, Removed Text - *Liber* (b) Over Writing - *Archimedes Flavus Fergusiorum, Royal Irish Palimpsest Extract* [STB07] *Academy, Dublin.*



(c) Layered Text, Scribe Identification required - *Kundige Bok Handwritten Re-creation*

(d) Text Recovery required - *16th Century Book Cover*

Figure 1.1: Examples of common difficulties in Historical Questioned Document Examination

are of great interest to historians and scholars also as they may contain important hidden text underneath the visible text. In medieval times it was common practise to erase previous writings by means of washing or scraping in order prepare the palimpsests for new text (figure 1.1(b)), leaving a slight remain of the original historically significant text. Collectively these practical examples, and questioned documents in general, produce obstacles for both historical and modern document analysts such as text recovery, bleed through, watermark analysis, and differentiation of dissimilar inks. The central research questions that are posed in the examination of historical and questioned documents include:

- Is it possible to recreate an accurate digital representation of a document a by non-destructive means for conservation and examination purposes?

- Can the individual scribal hands in a document be identified and their respective work be attributed leading to a relative chronological ordering of sections of documents?

- To what extent is textual recovery possible from documents that contain text that is faded (due to time, staining, or deliberate removal)?

- Can we digitally reverse signs of degradation and study the effect of environmental conditions on the object to reveal previously unseen sections of the document?

1.3 Digital Representations of Documents for Analysis - Related Obstacles

Document access and preservation goals are usually interrelated, since access to scholarly materials depends upon their being fit for use over time [Rie08]. As a result there exists a need by both archivists and scholars for a robust simultaneous method of conservation of and provision of access to documents. Typically, this is achieved through digitisation. Digitisation is a reformatting process of converting or transferring real world information and objects into digital format. This process, combined with the Internet's ability to connect, allow access to many digital representations of documents, but is not a perfect one in terms of the preservation of documents. It has been argued by [Smi99] that much is gained by digitising, but permanence and authenticity, at this juncture of technological development, are not amongst these gains. In contrast to the previous reformatting method of Microfilm, these digital forms rely on many technological appliances such as hardware and software to visualise the digital material. However, with advances in technology since Smith's publication the preservation of historical documents by digital means is fast becoming the method of choice, thus presenting the question - what is the best method of digitally representing a historical document for storage and future analysis? Digital images are widely used as the fundamental unit of conservation, storage, duplication of a digitised document due to the dual property of both being large digital datasets and easily interpretable by visual inspection, and as such provide a base for data anal-

ysis of digital representations of documents. Difficulty arises as the process of acquiring digital representations of documents is often a subjective one targeting a specific direction or purpose. A generic digitisation process must generate and store as much information and detail as possible so as to provide a comprehensive base for future analysis of many different types. In order for the materials to be digitised, they must be converted using a method to capture the material digitally (e.g., scanning or digital imaging) without altering the information that the material contains. Generally, this process can be divided up into several steps: cataloging, image capture, image analysis and recognition [Lin06a]. Digital imaging involves capturing information directly from a physical scene by camera, scanner or other imaging device, and refers typically to the capture of electromagnetic information.

> "An image is the optical representation of an object illuminated by a radiant source. Thus, the following elements are used in an image formation process: object, radiating source and image formation system. The mathematical model underlying the image formation depends on the radiation source (e.g. visible light, X-Ray, ultasound), on the physics of the radiation object interaction and on the acquisition system used" [Pit00].

Due to the wealth of variables and information (discussed in detail later in this thesis) that can be acquired using different modern imaging devices, the process of capturing digital images is, again, usually application centered and designed for purpose. The post-acquisition analysis and recognition of these images involves the application of computer algorithms to perform im-

age analysis and manipulation. Digital Image processing has been an area of research for quite some time, producing computational algorithms to perform classification [HC94], feature extraction [VA04] and data projection for a new visualisation of the data amongst other duties. Real world inspection of obstacles contained in historical documents such as fading of important features, or patterns (e.g. mould, staining) interfering with features of interest can provide difficult practical problems to document analysts, even those equipped with advanced image processing techniques and algorithms. Mapping practical problems inherent in historical document to a imaging technique or processing algorithm which will supply an answer is not always obvious. For example, a digital image of a faded historical text can yield feature (faded text) pixels of similar value to that of background or noise pixels in the image. Differentiation of faded text pixels from that of noise or background pixels can prove to be a difficult task for both visual inspection and a processing algorithm. Throughout the process of digitisation and analysis of historical document, many computational and analytical questions are presented, such as:

- What physical attributes and information of historical documents are to be captured to:

    1. provide a dataset which is both accurate and useful for computational analysis

    2. provide answers to historical questions related to the context of the historical document

- What methods of image acquisition, image formation and document modelling are needed to solve particular historical inspection and context related problems?

- Once a digital image representation has been formed as what analysis methods are necessary for the identification of features of interest, or to provide virtual restoration?

- Can we create a generic process that can provide an automatic segmentation of the features contained in these digital representations of historical documents?

1.4 Aims of This Thesis and Research

This thesis aims to address the field research problems outlined in the previous sections (1.2, 1.3) with the application of modern scientific techniques and emerging forensic technology (i.e. Hyperspectral imaging [Cha03] and associated image processing techniques [Cha07]) in a novel manner. This research pays particular attention to the analysis of historical documents and the application of hyperspectral imaging to multi-layered manuscripts to help solve these problems and overcome the uncertainty of the textual development. It focuses on the possibility of chronologically ordering entries to a multi-layered text where paleographic means alone would fail. The thesis will also examine the techniques ability to enhance the visibility of faint or obscured features, to distinguish and recognize materials, and to detect signs of degradation on documents using hyperspectral imaging and modern

software methods.

1.5 Background and Methodology

In the last decade Hyperspectral Imaging (HSI) and analysis of the resultant hyperspectral images has been an area of active research amongst professional analysts and scientists in a wide variety of fields or study, from microscopic analysis of biological materials [BPD07] to military purposes [BBD$^+$06], but most notably in remote sensing study [HG08, VA04]. Hyperspectral Imaging is a non-destructive optical technique that gathers electromagnetic radiation from across a large portion of the electromagnetic spectrum outside the capabilities of conventional imaging. Hyperspectral sensors can be used to measure many physical light characteristics of an object such as reflectance, transmission, absorption and luminescence 3.2. HSI sensors measure the intensities of these properties at specific wavelengths in regions of the spectrum ranging from ultraviolet, through visible, to infrared [Cha07]. The information obtained is in the form of an image cube, with the third dimension specified by spectral wavelengths. Each pixel in the image cube is a column vector of components representing intensity recorded over a specific wavelength region. These column vectors are a data representation of the electromagnetic energy measured at specific wavelengths over for each pixel. This essentially reveals a "spectral signature" for each pixel's performance over the electromagnetic region. Document examination performed based on the hyperspectral analysis of theses representative spectral curves discloses the possibility to read various different layers of a manuscript in a manner

not possible to the human eye and a more detailed approach to analysing elements of its composition. As such it presents the possibility of retrieving text that has been lost through fading, staining, overwriting or other forms of erasure. In addition, it offers the prospect of distinguishing different ink-types, and furnishing us with details of the manuscript's composition, all of which are refinements, which can be used to answer questions about documents such as date and provenance. The application of hyperspectral to the area of historical document examination is a novel and specialised area of research, incorporating both machine vision and image processing to provide information about the physical characteristics of questioned documents and historical manuscripts. HSI, together with modern two-dimensional spectral analysis software and three dimensional image and visualisation software provides modern researchers working in the field of document analysis with opportunities for forensic document examination that were previously unavailable [KAP+08]. Due to the non-destructive nature of hyperspectral imaging, laboratory instruments specifically developed for the analysis of historical paintings and documents have been employed in the field of conservation of artistic and historic objects [HAS03b]. Recently, due to the successful application of multispectral imaging in artistic object examination fields, new spectral imagers featuring significantly more spectral bands have been developed, resulting in a new method for historical and questioned documents, namely quantitative hyperspectral imaging [PSK+08]. Methodologically, there are two main fields of applications of this technique: (i) the extraction of relevant historic, diplomatic and paleographic information from

documents and (ii) the investigation of the impact of environmental conditions on document condition and of degradation effects on writing materials and substrates. In particular, reflectance curves found in different sections of the manuscripts can be compared with each other in order to determine whether different types of inks had been used during text composition or to identify modifications that occurred during the manuscripts' history. The approach for document examination taken in this thesis can be considered to be a two part system:

1. **Spectrometry and Hyperspectral Image Acquisition:** Utilising the full capabilities of the Forensic-XP Imaging Spectrometer to acquire hyperspectral images in both high spatial and spectral resolution of historical documents and manuscripts.

2. **Spectral Analysis and Data Processing:** Employ the use of two and three dimensional software designed for purpose; to perform spectral analysis of the data cube obtained during the initial stage, then to provide a means to apply statistical and image transform techniques to enhance the preliminary results obtained from imaging alone.

The initial stage involves the acquisition of the image data-cube using the Forensic-XP imaging spectrometer. Fundamental analysis of the hyperspectral data includes comparison of the spectral signatures of each pixel. More sophisticate mathematical feature extraction and classification techniques can then be applied to enhance features of particular interest, remove document degradation and give greater insight into a documents history.

1.6 Conclusion

This introductory chapter serves to outline some of the practical problems faced by document analysts, and in particular, those working with historical documents. Inherent characteristics in historical documents often illustrate the historical context of the document itself but impose certain examination obstacles to document analysis. For example, amendments or alterations to historical documents can give researcher a good insight into the historical background or situation associated with the document but at the cost of removal or obliteration of information contained within the document. Uncovering removed of obscured features from historical documents yields even more perspective into the surrounding context of the document. It is the aim of this research to investigate the performance of computationally modelling and digitising historical documents and performing detailed digital processing methods on the subsequent digital representations to provide solutions to these obstacles. An introduction to the imaging method, Hyperspectral imaging, and a methodology for its novel application to the analysis of historical documents is delivered through this chapter, which will be the method applied for supplying digital forms of a historical documents for conservation and analysis purposes. The following chapter represents a literature review of past and current prominent research questions in historical document analysis and related works. A critical review of these problems and methodologies applied to provide their solutions will be given. In chapter 3 the science of hyperspectral imaging and its practical implementation for document analysis is detailed. The chapter initially discusses the fundamen-

tal background properties of light, digital imaging and the electromagnetic spectrum. Following this, the spectroscopy equipment is illustrated and the image acquisition process is explained. The spectrometer's different modes of operation to capture different properties of an object are explained such as reflectance, luminescence, absorption and transmission. Chapter 4 examines the segmentation of hyperspectral images using image and signal processing methods. The research in this chapter is at an algorithmic level exploring possible methods for hyperspectral image segmentation and the extraction or enhancement of features of interest using both statistical and image transformation means. In the final chapter, we analyse our method of automatic document segmentation to perform specific analysis and its development, and finally we summarise the novel techniques presented in this thesis along with their results and draw a conclusion about the benefits and limitations of hyperspectral imaging for historical document examination.

Chapter 2

Hyperspectral Imaging for Questioned Historical Documents

2.1 Introduction

Visual inspection of documents can often be rewarding for document analysts as it is a fast and inexpensive method for examining the current state of a document as a whole, detecting critical areas for further analysis, identifying discolouration, staining, and where corrosion occurs, and other degradations impairing the legibility of the document. From the visual inspection of documents, a qualitative impression of the overall condition of document can be formed. However, the rating of a documents condition based on visual inspection is a subjective process. Due to the limitations of the human eye,

optical inspection of documents is restricted to the spectral range in the visible region of the electromagnetic spectrum (approx. 400 - 700nm). Thus, visual inspection utilises only a small portion of the available visual information that can be retrieved from materials. Most conventional methods of imaging historical and questioned documents are restricted to this small portion of information, again in the visible range. The majority of the currently used digital imaging techniques are restricted to visible colour imaging, typically recording three independent channels of information from red, green, and blue wavelengths from the visible range of the electromagnetic spectrum respectively. Often, it is seen in scientific imaging applications that large amounts of data representing complex systems can only be represented by visualisation, as images. Light, and specifically its properties such as reflectance, are the fundamental units used in capturing and measuring an objects state, although almost every physically measurable property can be used as a basis for image creation. Multivariate data, data obtained by measuring a number of different quantities simultaneously, has been used in image creation combining one or more measurable properties into a visual image. Using modern spectral imaging equipment it is possible to acquire reflectance spectra from much more channels of information taken from a wider range of the spectrum in an effort to attain a qualitative analysis of a documents properties and assessment of their relation using as much of this visual information as possible, while also supplying a digital representation of the document for preservation and analysis. This chapter examines the origin and development of hyperspectral imaging (HSI) and processing in

the context of the analysis of questioned historical documents. It provides an extensive literature review and a critical analysis of the role HSI can provide in the analysis of questioned and degraded documents for contemporary and historical manuscripts and current related work being carried out in the area. It examines what advantages are provided by digitising historical documents and how hyperspectral images and subsequent image processing can attempt to answer the research questions outlined previously. Finally some conclusions of the current state of the art thinking surrounding the topic are drawn.

2.2 Hyperspectral images for Representation and Preservation

The preservation of document collections is a key activity for the study of materials of archaeological, cultural and historical value, monitoring and evaluation of conservation treatments and digital imaging for documentation and archiving [FK06]. Digital preservation of documents refers to the series of managed activities necessary to ensure continued access to digital materials for as long as necessary [Coa09]. It is the activities required to keep materials in usable form for a long period of time; the actions required to maintain access to digital materials beyond the limits of media failure or technological change. Digital imaging plays an important role in the preservation many different types of documents and non-destructive optical techniques have always belonged to the most important investigation methods applied in paper and

writing durability research [SaAS04]. Many different factors influence and determine the most appropriate digital format(s) to be used for representing and preserving historical documents. To access the suitability of digital imaging for historical document preservation we must examine the requirements and use-cases of a digital representation of these documents. In the context of text documents, the appropriateness of major formats (TIFF, PDF, XML) for representation and preservation have been evaluated as part of a comprehensive study carried out by the Library of Congress, USA [AF06]. A suitable method for digital preservation of historical documents depends on the kind of information being preserved, the original format of the object, future uses of the digital representations, and the expectation of future users. The suitability of these formats for document preservation was evaluated under the following technical criteria - disclosure, degree of adoption to which the format is already in use, transparency or the degree to which the digital representation is open to direct analysis, self-documentation, external dependencies such as specific hardware or software for rendering, impact of patents and technical protection mechanisms. Much of these issues are relevant to the assessment of the suitability of hyperspectral imaging, and specifically hyperspectral images, to the preservation of historical documents. Similar research on effective methods of digitally preserving archaeological objects and complex artwork [VRR+07] found that the generation of digital models of these objects for documentation purposes requires a technique with the following properties - accuracy, portability, low cost, fast acquisition and flexibility. The criteria for the creating of digital forms of artwork also have bearing on

a suitable digital format for representation and preservation of historical documentation. An appropriate digital image for representation must capture as much relevant information as possible for not only for the current or immediate analysis purpose but it also must contain a large dataset for future users and uses. Spectral imaging technology is a new tool in the field of historical preservation, but has provided promising results in the analysis of paintings [CLP+99, BCLP98] and written documents [BPP+03]. Spectral imaging addresses the digital analysts need for quantity of information obtained by simultaneously recording both spectral and spatial information about an object in high resolution. Spectral images contain a series of digital images acquired at specific wavelengths from a wide region of the electromagnetic spectrum, including wavelengths from the ultraviolet, visible and infrared regions, and as such provide a this large information dataset for digital analysis. Spectral imaging has been used for the conservation and study of high profile historical documents, for the conservation and reading of manuscripts and documents which have degraded naturally over time by enhancing the contrast between under-written and over-written text[SaAS04], most notably in the imaging of the Archimedes Palimpsest and the study of the Dead Sea Scrolls contributions. The Archimedes Palimpsest is a medieval manuscript, which initially existed as a tenth century copy of otherwise unknown work of Archimedes describing the history of mathematics. Some 200 years later the earlier text was erased, each page was cut in half and then overwritten as a prayer book. The palimpsest now contains erased texts that were written several centuries earlier still including two treatises by Archimedes that can

be found nowhere else and so has been the subject of conservation, imaging and scholarship, in order to better read the texts[WN]. The palimpsest was the subject of an extensive imaging study and conservation undertaken at the Walters Art Museum in Baltimore, due to the document degrading considerably due to mould. Spectral imaging and computer processing of the digital images produced substantiated a vital part of the conservation of the manuscript. The Dead Sea Scrolls, discovered between 1947 and 1956, are a collective of approximately 900 documents of great religious and historical significance containing historical text including the oldest known surviving copies of Biblical and extra-biblical text from the Hebrew Bible. In 1991, the Israel Antiquities Authority (IAA)[Aut07] employed the existing state of the art imaging technologies to publish high-resolution images of all the Dead Sea scrolls to the public, but also to reach an understanding of how to optimally preserve them. In 2007, an international committee of document experts evaluated the most advanced imaging technologies to monitor the well-being of the scrolls, and to expand access to the scrolls while preventing further damage from physical exposure. Due to its non-invasive and precise manner, spectral imaging was used for the long term preservation and monitoring of the scrolls; creating both high resolution colour and infra-red images of the scrolls in their entirety. Thus, preventing the need to re-expose the document and provide access to a comprehensive databank of the scrolls. In both cases, multispectral imaging, combined with digital image analysis has been used as the preservation format, the source dataset of which further processing and analysis has been performed, and a technique to discern

previously unreadable sections of text from the manuscripts[BS96].

2.3 Spectroscopy for Historical Document Analysis

Spectroscopy is the study of the interaction between radiation and matter, covering many forms of radiation including electromagnetic radiation (light) and partical and sound emissions. This research will focus on light spectroscopy, which is the study of the relationship between specifically electromagnetic radiation and matter. Imaging Spectrometry has been defined as the acquisition of image data in many contiguous spectral bands [TB92]. The terms multispectral imaging and imaging spectrometry are often used interchangeably, both referring to the acquisition of a series of digital images at a multitude of optical wavelengths. Hyperspectral imaging (HSI) can been seen as an extension of multispectral imaging, expanding and improving the capabilities of multispectral imaging by taking advantage of hundreds of contiguous spectral channels of information, for the purpose of uncovering materials that cannot be resolved with multispectral imaging. Hyperspectral Imaging, in varying forms, has been employed in a many instances for feature enhancement, such as visible light (VIS) reflectance spectroscopy, visible / infrared (VIS-IR) spectroscopy, ultraviolet visible (UV-VIS) spectroscopy and fluorescence spectroscopy; each technique utilising one or more different properties of light and the electromagnetic spectrum to augment a particular object or feature of interest in a sensed image. Most notably, HSI and these

spectroscopy techniques have been employed in the area of remote sensing of geospatial data in recent years, to great benefit. Many factors can hinder the detection and classification of features in a geospatial image such as noise, low resolution, and even a low number of representative pixels for the feature of interest. VIS and VIS-IR spectroscopy can improve on conventional imaging in this respect. Quantitative measurements of spectral characteristics of materials acquired by HSI sensors produce spectral signatures defining the chemical composition of each material in a scene. HSI has been widely used in the determination of geological minerals and materials and the calculation of their respective concentrations as in [DB91]. This approach to feature detection has also proven to be a valuable tool for document analysts[KAP+08] for the extraction of relevant historic information from documents and study of environmental conditions on document condition and of degradation effects on writing materials and substrates. Klein et al. provide an excellent description of the basic concepts, working principles, construction and performance of a HSI device specifically developed for the analysis of historical documents. The device described by Klein et al. has been used to record the variation of spectral reflectance on a historic 17^{th} Century map, and also used the instrument to compare the local variation of the yellowness index of reference papers stored in a bound volume, and loose sheets. Ongoing HSI investigations allow the detection and visualisation of differences in ageing processes on a document, and are particularly useful when taken, for example, before and after an exhibition, whereby it is possible to investigate the effects of exhibiting and handling on document yellowing [PSK+08].

HSI investigations carried out by Padoan et al the first manuscript written in Dutch Language (14^{th} Century) aimed to provide an answer to whether a coat of arms, drawn on the lower part of the first page could have been ascribed to the famous 15^{th} Century Flemish bibliophile *Lodewijk van Gruuthuse* or whether the element may have been added at the same time as the 17^{th} Century text surrounding it [PSK+08]. It was concluded that the HSI derived image shows no correlation between the coat-of-arms and the surrounding text, while strong similarities were observed within areas where corrections were made in the 15^{th} Century. There have also been several other recent reports of promising results from the application of the HSI analysis of paintings and conservation of works of art [FK06], paper discolouration and detection of iron-gall inks [HAS03a]. Based on a false colouration of VIS-IR spectroscopy (another technique which finds its origins in the remote sensing field [Tuc79]) has shown that HSI and VIS-IR spectroscopy is a valuable tool for comparing iron-gall inks, and distinguishing various materials in a non-destructive manor. The enhancement of historical documents and painting has also benefited from hyperspectral methods. Due to differences in spectral profile of features in a hyperspectral image, hyperspectral imaging combined with false colouration can be used to augment and extract specific features and diminish noise or other unwanted features. As demonstrated in [KAP+08] using the *Map of Syracuse* as its subject, spectroscopy is used to diminish the effects of corrosion and improve the observable fine detail in the painting. Spectroscopy measuring fluorescence (light energy emitted due to absorption of light of a different wavelength) has also proved to be useful

in the examination of an objects state, mainly in medical imaging and the assessment of foodstuffs and other organic material [WMW98]. The application of fluorescence spectroscopy to document analysis is an emerging area which holds potential in particular for the analysis of paper and ink. Tilley [Til00] provides an excellent discussion on the relationship between light, the optical properties of materials and colour, and is particularly useful for those interested in using HSI to distinguish different materials, for example, inks on paper. HSI observations of historical documents written on paper, parchment, etc. are complex spectral combinations of the reflectance of a collection of materials that have different temporal degradation properties. Vaarasalofs [Vaa99] provides a useful discussion on the optical properties of paper, which loses its optical properties as time passes [dR92]. More recently, De la Rosa and Bautista[lRB05] have discovered that to find the presence and concentration of different colourants or components in the paper it is only necessary to know the spectra and fluorescence lifetimes (at 337.1nm). They indicate that these kinds of measurements could be useful for studying the papers long-term stability and how ageing affects it, and is particularly important in the preservation of paper-based historical records (Committee on Preservation of Historical Records 1986). Light spectroscopy has also been particularly useful in the examination of paper ageing, for example, in accelerated ageing experiments [Ban02] and on the evaluation of whiteness and yellowness [Smi97]. A particularly interesting example of this high-precision, skilled and manual work is reported by Quandt[Qua91] who conducted a detailed physical analysis and treatment of a late 13th Century copy of the

Etymologies of Isidore of Seville. The manuscript, although damaged over the centuries, retained most of its original medieval binding structure, and it is reported that part of the project included the compilation of technical evidence that would lead to accurate localisation of the text and a reconstruction of the binding history of the volume. One aspect of this work included the removal of apparently blank pastedown from the upper board; however, disassembly revealed that the pastedown contained writing (cursive Latin text) on the verve. Quandt concludes that manuscript fragments, such as those used as pastedowns, are potentially important, as they may serve to document the origin of the manuscript and its medieval binding[Qua91].From these examples it can be concluded that hyperspectral imaging in its many forms has been proven as a non-destructive tool for the analysis of many physical objects. Most notably, its main contribution seems to be directed at problems that involve resolving features only partially, or sometimes not at all, perceivable to that naked eye. Its application to the field of document analysis is still in development stages but has been very successful in the uncovering of text, and for the analysis, quantification and the enhancement or isolated of features in an image. This, however, is often achieved using a specific component in the hyperspectral armoury (e.g. utilisation of fluorescence spectroscopy for text recovery) and a one-to-one mapping of hyperspectral imaging techniques to the solution of document analysis problems is not always obvious. Hyperspectral imaging approaches this by creating a large dataset for analysis purposes, recording vast amounts of information from wide reaching portions of the electromagnetic spectrum, providing a docu-

ment analyst with information regarding an object's behaviour across this wide region. Other factors such as illumination source, type of emission being examined (eg. reflectance, transmittance, fluorescence, absorption) also have a bearing on what information is acquired and available to the analyst. As such, currently the application of hyperspectral imaging for the examination of historical documents will require some manual influence or examination to solve document specific questions.

2.4 Hyperspectral Image Processing for Document Enhancement and Recognition

HSI sensors gain spectral information in the form of a series of monochromatic images representing the intensity of a particular property of light at a specific wavelength for each spatial pixel in a two dimensional image. Fundamentally, this data can be thought of as a large matrix of intensities or numerical values. The collection of these intensity values, representing each monochromatic image in the captured series, can be formatted as a three dimensional data matrix for computational processing purposes. Without any form of manipulation these HSI devised digital images are a good method of digitally preserving historical documents and manuscripts. This three dimensional data also provides ample spectral data to perform post processing for more detailed analysis. Many motivations exist for the manipulation of hyperspectral images in many different fields and applications. Signal and hyperspectral image processing techniques have, again, been most prominently

exploited in the realm of remote sensing, but they have also exhibited potential in documenting restoration and improvement. When an image is scanned using standard conventional imaging systems three views (or channels of information) can be obtained using red, blue, and green channels. Hyperspectral scans of the same document will permit the use of information acquired using non-visible wavelengths of light combined with these visible ones to produce a comprehensive dataset. Some of the improvements which can be made to historical documents with the aid of hyperspectral processing include the minimisation of signs of aging or degradations, bleed though removal, the detection of underlying patterns or watermarks and blind source separation techniques for detecting hidden texts and textures in document images. Processing the different colour components makes it is possible to extract some of the overlapped patterns, and in some cases, achieve a complete separation of each of the patterns. In the field of historical document examination, the post processing of hyperspectral images has been a vital tool in the attempt to uncover hidden text and textures [TSMB04b]. In this approach, Tonazzini et al. have modelled a document as a series of superimposed patterns and have implemented a linear mixture model to describe the relationship between the patterns. The problem of detecting the patterns that are barely perceivable in the visible colour image is thus formulated as the one of separating the various patterns in the mixtures. A successful discrimination of barely perceivable features has been achieved based on statistical methods including independent component analysis. This discrimination presents a method for extracting partially hidden texts, and the removal of the effects of the seeping

of ink from the reverse side. Another similar method utilizes blind source separation (BSS) techniques for the removal of a particular document degradation known as bleed through [TSMB04c]. Bleed through is a particular type of document degradation which is the presence of patterns interfering with the main text due to seeping of ink from the reverse page side. This type of decomposition is prominent in historical texts and as a result strategies have been developed by analysts for the purpose of its removal [TSMB04c]. This strategy uses a colour de-correlation method to model a document image as a linear combination of three independent patterns: the main foreground text, the bleed through, and the background pattern. In general, it was found that the mixture coefficients are not known, and the separation problem becomes an instance of the blind source separation problem [BAMCM97] where the information in the observed signals exists in 'mixed signals'. It has been shown using the processing technique, independent component analysis (ICA), this problem can be minimised and impressive results can be obtained for the removal of bleed through for manuscripts. Hyperspectral image processing can also aid in the reading of palimpsests and documents which have degraded naturally over time by enhancing the contrast between under-written and over-written text. Hyperspectral imaging techniques have proven to be a powerful tool in the scientific analysis and documentation of paintings in the field of Art Research and Conservation. Analysis undertaken by researchers at the Rochester Institute of Technology and the Johns Hopkins University have produced previously unseen parts of famous palimpsests with great success and Archimedes' palimpsest is now readable using digital processing

of ultraviolet and visible light[TSMB04a]. Another feature of hyperspectral data is that images scanned in the infrared spectrum can be used to detect water markings [PSK+08], underlying patterns and images in historical documents. Principle Component Analysis (PCA) is another statistical technique with can be applied to hyperspectral data as a means to isolate, uncover and enhance non-obvious features in document images. PCA [Jol02, GY95] is principally used to condense a high dimensional dataset of interrelated variables to a dataset of lower dimensions retaining as much as possible of the variation between the variables as possible. This process can also be used to identify the most representative elements of the data, called principal components, which is of use in hyperspectral image analysis. These principal components are selected and ranked by their relative variance, judging the most important representative element to be the most variant [HBB+06]. In [Att05], the PCA algorithm has been successfully applied in the analysis of historical parchments from the Walters Art Museum, Baltimore. The algorithm has aided in the enhancement of faded text and the isolation of independent inks. It was found that the first principle component revealed a similar view to that of visual inspection by the naked eye. The second component significantly enhanced the legibility of the main text contained in the manuscript, which had been faded in parts, by displaying the maximum contrast between the ink and paper. The third principal component distinguished the individual inks in manuscripts containing more than one type of ink by highlighting differences in there spectral response at wavelengths outside the visible range. As outlined in the previous section, hyperspec-

tral images contain provide a method of digital preservation for historical documents but also offer an inherent wealth of data for analysis purposes. The post processing of hyperspectral images has seen much development in the areas of remote sensing of geographical images [GG07, VA04, HC94] for purposes as feature extraction, target discrimination, classification and quantification. Obstacles occurring in historical documents for document analysts and scholars can often be classified under these terms and have similar solutions and as such the post processing of hyperspectral images of historical document can yield useful results.

2.5 Review and Conclusions

Review and Conclusion Hyperspectral imaging for document analysis is an emerging field of study but has already proved useful in aiding many high profile document examination projects. HSI is a non-invasive, non-destructive imaging technique used to capture the optical properties of objects at a series of contiguous wavelength bands [Cha07]. The series of these hyperspectral bands are typically in the hundreds, and contain information from a large portion of the electromagnetic spectrum, encompassing information outside the capabilities of the human eye and other conventional imaging systems. Historical documents, due to their very nature, provide analysts and scholars with difficult problems. The question of how to digitally represent and preserve the current state of historical documents is likely the most prominent obstacle encountered by historical analysts and document conservators. In this chapter we have examined hyperspectral imaging as an answer to

this question. The capability of the technique to provide an accurate and true representation of an objects state digitally has put it at the forefront of preservation mechanisms for historical documents. Its non-destructive nature is a vital component to its success as means of preservation. Current conventional digital imaging techniques combined with the World Wide Web is quickly becoming the standard for digital representation of documents and strikes a great balance between providing accesses fragile digital documents and preserving their integrity. Hyperspectral imaging can provide a superior method of digital representation of historical documents due to fact that it captures more information about a documents current state and its inherent non-invasiveness. The spectroscopy component of the hyperspectral process also provides further potential for the tool as a method of historical document analysis. Hyperspectral sensors provide document examiners with the capability to inspect a document in greater detail and at more viewpoints than that of RGB digital imaging or visual inspection. The information obtained is optical energies measuring a specific property of light such as reflectance, transmittance, or fluorescence. Unlike the human eye, which just sees reflected light, Hyperspectral imaging advances these capabilities and provides the ability to measure physical attributes such as fluorescence which can aid in the text recovery, transmittance for watermark detection and enhancement as well as absorption. Certain objects leave unique 'fingerprints' across the electromagnetic spectrum. The implementation of spectroscopic measures to historical document investigations has yielded new mechanisms for digitally identifying individual inks and text recovery. Visual inspection of

faded text involved the eye capturing reflectance information from a combination of wavelengths of light from the visual portion of the electromagnetic spectrum. The text appears faded due to a poor contrast between the reflectance of the ink and the reflectance of surrounding document at these visible wavelengths. Hyperspectral sensors combine this information with many more wavelengths of light including optical information from ultraviolet and infrared regions, often yielding a new perspective on faded ink. Particular inks have been shown to have a considerable response in infra-red regions [Att05] of the spectrum and result in the ability of creating a significant contrast between ink and page to recover faded text. Hyperspectral sensors also have the capacity to scrutinise an objects spectral response at a specific wavelength, filtering out all others which can allow a document examiner to examine a particular feature in the document in greater detail enabling identification of the materials that make up a scanned object. Fine tuning of the examination wavelength can identify noteworthy changes in the relationship between features in a document, possibly differentiating inks or features, or to study the effect of environmental conditions on the document to minimising the effects of degradation. We have seen that the spectroscopy capabilities of hyperspectral imaging have yielded the answer to many historical questions but the tool has more to contribute to the field of document analysis. The dataset acquired by HSI sensors is useful for computational analysis and with the application of statistical and image processing methods it can provide answers to historical questions related to the context of the historical documents. Document modelling and statistical analysis methods

such as Independent Component Analysis or Principle Component Analysis have been employed for the identification of features of interest or to provide virtual restoration. This is mainly achieved through a segmentation process; dividing the image up into the most statistically variant components and then enhancing features or components of interest. Throughout the hyperspectral process, from image acquisition to spectroscopic examination to post-processing, there is a requirement for an examiners input and inspection as very little of the document segmentation process is automatic. Hyperspectral imaging and associated algorithms provide a skilled document expert with a new digital format for a comprehensive method for preservation and for providing access, and also a dataset from which more detailed computational examination can be performed.

Chapter 3

Applied Spectroscopy Methods for Text Recovery and Ink Identification

3.1 Introduction

This chapter describes the application of spectroscopic imaging to historical documents for text recovery purposes. The non-invasive imaging method of analysis can be used to measure and distinguish fine differences in spectral reflectance and fluorescence properties of document features. A fundamental objective in the reading of text is to separate foreground text from the background document. This is not always a simple process when dealing with historical document images as faded text over time can have a similar reflectance to the background in the visible region of the electromagnetic

spectrum. Scribe attribution or the ability to associate a particular section of text text in a historical document to an author is often an area of particular interest for historical document experts and is particularly challenging when both inks in a document appear identical when visually inspected under natural light. We shall also demonstrate a spectroscopic method of distinguishing individual inks that may appear similar. Two historically significant examples of faded text have been selected as a problem set for the application of faded text enhancement methods. We have also created a simulation of a well known historical manuscript containing two apparently identical inks for the process of distinguishing inks. The chapter describes the optical properties and the imaging system being used, and then provides a detailed account of the spectroscopic processes applied to achieve the goals of faded text recovery and differentiation of similar inks.

3.2 Optical Properties for Hyperspectral Document Analysis

Light Spectroscopy Light Spectroscopy is the study of light that is emitted by or reflected from objects or materials [BM97]. Light, a form of electromagnetic energy, consists of many basic particle-like packets, called *photons*, which contain energy and momentum. Light and other forms of electromagnetic radiation are commonly expressed in terms of their wavelength where each photon of light has a wavelength determined by its energy level. The wavelength of the emitted light depends on how much energy is released and

different type of atoms will release different sorts of light photons. When a single frequency light comes in contact with an object several different phenomena may be observed. The incident light may be:

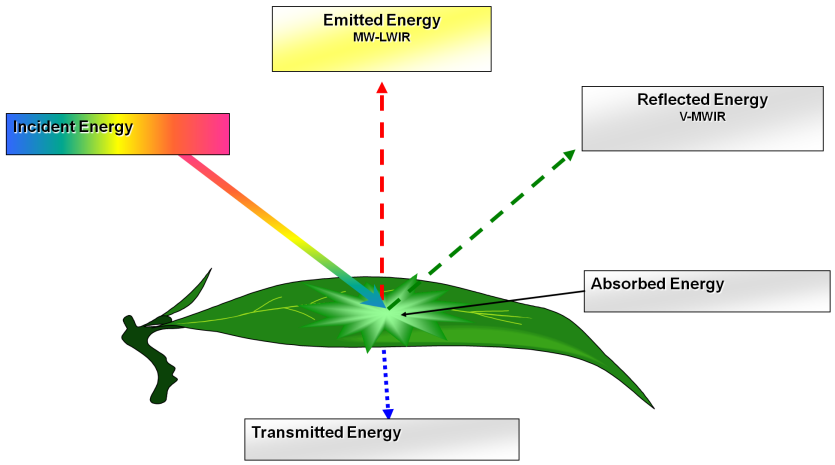


Figure 3.1: Optical Properties recorded in Hyperspectral Imaging

- Reflected - Reflectance is the fraction of incident light radiation that is reflected by a surface. The reflectance spectrum is the plot of the reflectivity as a function of wavelength.

- Scattered - Scattering occurs in certain forms of electromagnetic energies including light, where part of the energies are forced to deviate from a straight trajectory by non-uniformity in the material through which they pass.

- Transmitted (and Refracted) - Light transmission is the percentage of incident light that passes through a material.

- Absorbed - Absorption of electromagnetic energy describes the energy

is transformed to other forms of energy, for example, to heat when light comes in contact with objects or materials.

Single frequency incident light is uncommon, however. Normally, incident light such as visible light (solar radiation) or light emitted from incandescent light bulbs contains many different light frequencies, and when light of this type strikes an object, the object will selectively reflect, absorb or transmit certain frequencies. Another optical phenomenon that can occur and be measured using HSI equipment is the emission of light as a result of part of the incident energy being absorbed, called fluorescence. This absorbed energy causes excitement of the molecules within the object and then emits energy in the form of lower energy light usually within the visible range of the spectrum. Generally, hyperspectral imaging involves the quantitative computation of the spectral characteristics of materials producing measurements from a contiguous portion of the light spectrum. A plot of an objects reflec-

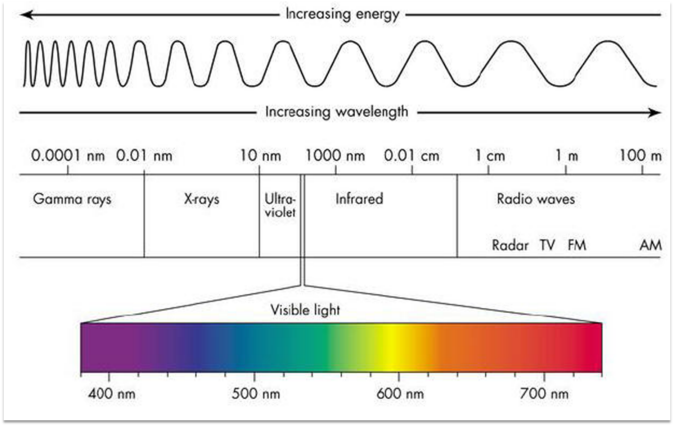


Figure 3.2: Electromagnetic spectrum

tivity as a function of wavelength may also be helpful in distinguishing inks

and pigments [Att05]. This defines the composition of the material through its spectral response over the electromagnetic region (termed spectral signature. Hyperspectral imagers can be used to record these optical properties across a large portion of the spectrum, resulting in a comprehensive digital description of an object - i.e. a spectral signature for each pixel representing the intensity of the specific optical property being recorded at a series of wavelengths within the spectrum range.

3.3 HSI Sensor - ForensicXP-4010 Imaging Spectrometer

The HSI inspection device used throughout this thesis for hyperspectral image acquisition and examination purposes is the ForensicXP-4010 Spectrometer. The ForensicXP-4010 is an imaging spectrometer designed specifically for document examination. The instrument provides a non-destructive tool for authenticity determination of documents and handwriting. The spectrograph performs spectral imaging of absorption, reflectance, transmittance and fluorescence of questioned documents across a wide portion of the electromagnetic spectrum (specifically 400nm to 1000nm), yielding spectral information ranging from the ultraviolet range, through the visible, and into the infrared region of the light spectrum for each pixel in an image. However, the pixel value in any spectral image taken at a specific wavelength does not only depend on the spectral reflectance value at the corresponding object location, but very strongly also on the exposure time and other

settings and characteristics of the camera and light source used in the instrument. In order to be able to derive from the recorded raw spectral images reliable and accurate spectral reflectance data of the object itself, special care has to be taken in the setup of the instrument and in the calibration of each measurement [PSK+08]. The hyperspectrum analysis and processing tools, combined with powerful optical elements, provides the examiner with the ability to detect minute differences between microscopic features of questioned documents that are undetectable by traditional imaging methods. In operation the ForensicXP document examination system consists of a high resolution camera, adjustable continuous interference filter to allow specific wavelengths of light into the camera at any one time and tunable light sources to illuminate the target object with specific incident light 3.2. The ForensicXP system offers several types of illuminating sources to be used in various applications.

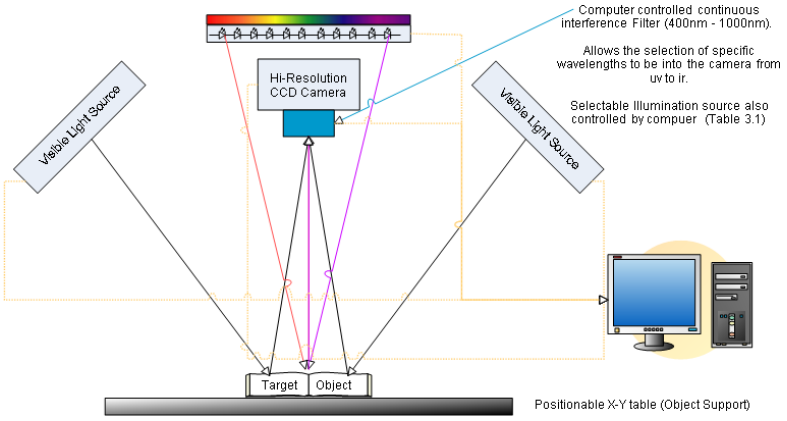


Figure 3.3: Hyperspectral Imaging Sensor basic schematic

3.4 Digital, Multispectral and Hyperspectral Image Acquisition

Images have the dual property of both being large data sets and visually interpretable entities[GG07]. In digital imaging, a pixel (or picture element) is the smallest item of information in an image. A digital image, normally arranged in a 2-dimensional grid, is an array of I rows and J columns of pixels. Images of this form are known as raster images, because they consist of a discrete grid of samples and are a common form of representing scientific and medical data. Each pixel in a digital image is a digitised grey scale value representing intensity measured at a specific wavelength or colour. Colour digital images contain numerous channels of information to describe each pixels colour. For instance, an image from a standard digital camera will have a red, green and blue channel yielding a RGB image. RGB image arrays are made up of width, height, and three channels of colour information. One channel represents the amount of red in the image (the red channel), one channel represents the amount of green in the image (the green channel), and one channel represents the amount of blue in the image (the blue channel). A channel in this context is the grey scale image of the same size as a colour image, made of just one of these primary colours. Multispectral and Hyperspectral imaging extends the number of channels acquired during the capturing of the digital image, incorporating channels of information outside the visible spectrum. In the hyperspectral context, each channel corresponds to a specific wavelength region and contains intensity values for spectroscopic

(a) RGB Image - 3 channels of colour information

(b) Red Channel

(c) Green Channel

(d) Blue Channel

Figure 3.4: RGB Image broken down into Grey Scale Image Channels

information. Multispectral images are produced by sensors measuring light intensity within several specific channels of the electromagnetic spectrum (commonly between 4 and 10 different channels). Hyperspectral sensors, however, measure energy in narrower and more numerous bands than multispectral sensors. Hyperspectral Image systems typically image a scene in hundreds of indexed bands, utilising information from a wide region of the electromagnetic spectrum. Modern hyperspectral image sensors can be used to capture the attributes of light emitted by materials, and its variation in energy with wavelength, at a series of narrow and contiguous wavelength bands. As mentioned, the Forensic XP-4010 forensic sensor measures spectral information in the ultraviolet, infrared and the visible region, specifically the 400nm-1000nm region. The data obtained is in the form of a data-cube

(Figure 3.5) which represents the image information as a data set in three dimensions; two of the data cube's axes represents the spatial data, while its third axis represents the spectral information. This image cube consists of hundreds of spatially recorded images, acquired contiguously over the wavelength region. Choosing an (x,y) point in the data cube will give series of values of intensity of light at a particular wavelength reflected from that particular point on the surface of the object for that point through all the images called a "hyperspectral signature". Similar features in the hyperspectral image will have similar signatures across the wavelength region and a differentiation of these signatures can result in a detailed segmentation of a hyperspectral image. The 3-D data cube can be considered as a stack of images, and at the same time as a two-dimensional array of many low-resolution spectra, one for each pixel. The spectra give information on the reflectance properties of the sample at each point in the image. Different reflectance properties are usually the result of different chemical compositions. For art and document investigations, the composition of pigments is of considerable interest, and distinguishing among them can provide evidence of alterations [Att05].

3.5 Hyperspectral Reflectance Imaging for Ink Analysis and Segmentation

When performing much of the HSI investigations, the fundamental property that we wish to attain is reflectance. The spectral reflectance is the

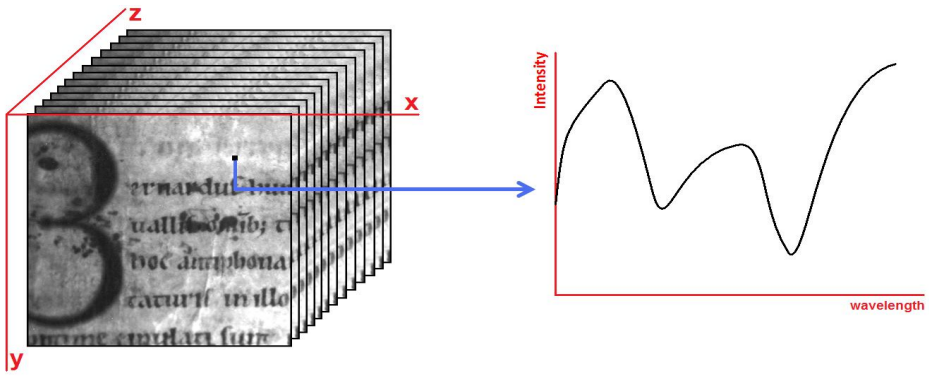


Figure 3.5: Hyperspectral DataCube

fractional amount of incident energy that is reflected from a surface with respect to wavelength [KAP+08]. The amount of reflected light varies with wavelength for the vast majority of materials as incident energy at particular wavelengths is absorbed or scattered to different directions. Materials which have a similar reflectance under natural light may have vastly different reflectance under light of a specific wavelength from other regions of the light spectrum. The importance of this, in the context of analysis of documents, is that this difference in spectral reflectance allows for classification of different features of the document which under natural light visually unassisted look identical. The general objectives of cursive text segmentation include tasks such as word spotting, text/image alignment, authentication and extraction of specific fields [LZT06]. An important step associated with all of these tasks is the segmentation of the document logically into units, for instance text lines, words or letters. In general, this is difficult due to the low quality and the complexity of these documents, and automatic text segmentation of such kind is an open research field. Sophisticated image pro-

cessing of single-image documents is hence the norm so far [LZT06]. Here we describe our recent approach towards segmentation of a different kind, which we refer to as hyperspectral segmentation; the technique is based on the separation and segmentation of different inks by recording and analysing their reflectance properties. This technique is particularly useful for the segmentation of texts that have been edited by various authors over a long time period. Thus, it helps answering basic palaeographic questions and allows dating text by comparing the segments with known dates, or using repositories containing hyperspectral properties of different materials (e.g. inks, paper, etc.). Figure 3.5 shows a simulation of the textual evolution of *Kundige Bok* [Rehng]. It is a sentence from this medieval town law, though re-written on modern paper with modern inks for demonstration purposes. As can be

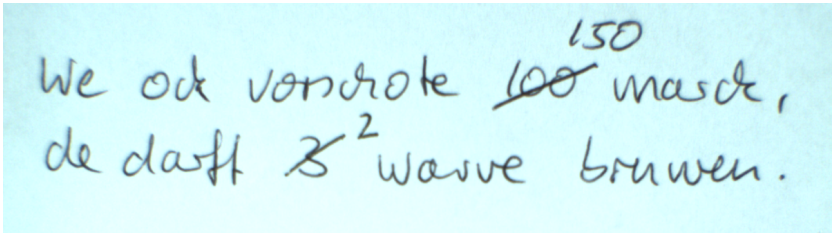


Figure 3.6: Simulation of Kundige Bok - *an example of layered text in historical documents*

seen in the sample (Figure 3.5), and in the originally imaged document, it is hardly possible for the human eye (if at all) to detect whether the changes of the text were made with a different ink or not, thus giving an indication whether the two changes within the sentence originate from a different point in time of the writing process or not. In this recently conducted experiment, the reflectance spectroscopy images recorded revealed the different inks quite

easily. Images of the spectral reflectance were recorded sequentially at 20nm intervals over the 400nm - 1000nm wavelength region using standard illumination. Examination of the inks performance over this wider region of the electromagnetic spectrum reveals both subtle and not so subtle differences in the reflectance measured at specific wavelengths. Figure 3.5 shows a plot of the spectral response of each of the inks in the simulation, measuring the reflectance on three different dots on the manuscripts, marked green, blue and red. A significant difference in the reflectance of all three inks is seen,

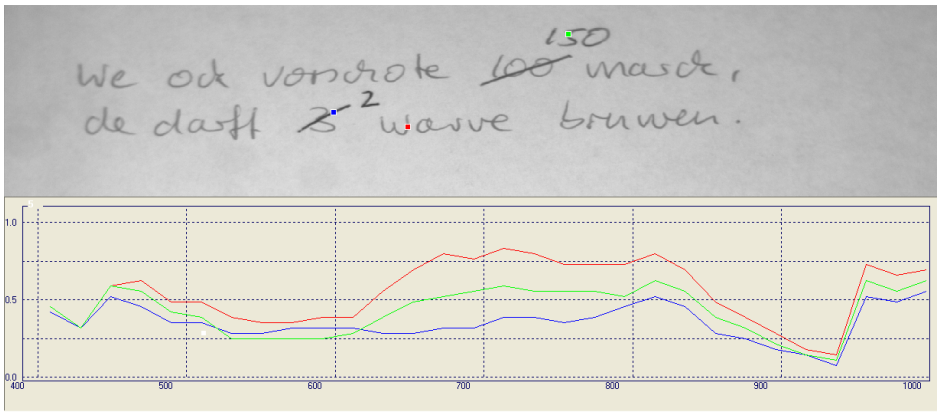
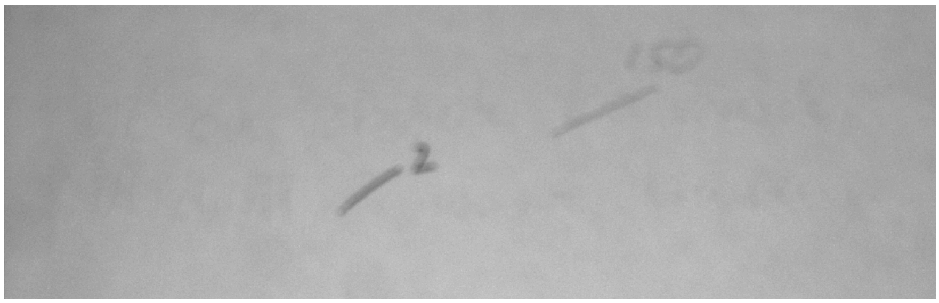
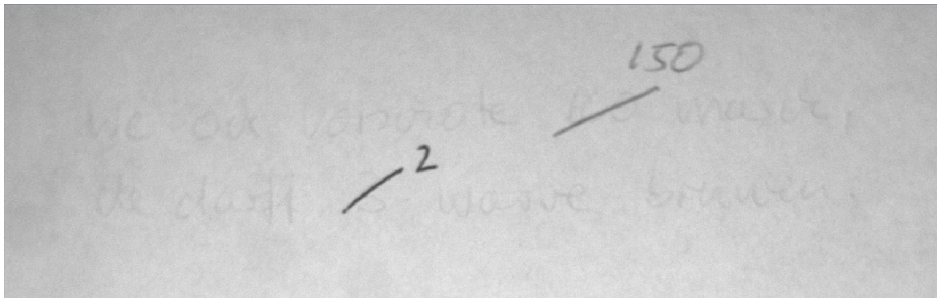


Figure 3.7: Spectral Plot of simulation inks (400nm - 1000nm)

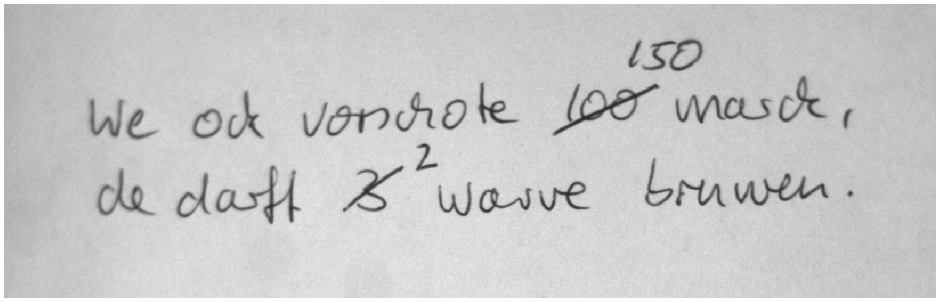
most notably between the 700nm to 800nm range. Further analysis can be carried out by the examination of the inks at the particular wavelengths of interest recorded in the hyperspectral image. The reflectance measurement that is of paramount importance is 700nm as this is where all three inks provide the greatest differentiation in spectral profile. To illustrate the change in reflectance at different regions of the electromagnetic region measurement s taken at 400nm and 900nm are included in figure. Three different inks can clearly be distinguished by the difference in spectral profiles across the wave-



(a) Reflectance at 400nm



(b) Reflectance at 700nm



(c) Reflectance at 900nm

Figure 3.8: Spectral Reflectance of Simulation Inks at 400nm, 700nm, and 900nm

length region. The underlying original text, represented by a red dot and spectral profile and two different alterations, denoted by green and blue dots and corresponding spectral profiles, made to the original text. The alteration, made at the position indicated by the blue dot ('3' substituted by '2')

was likely made with a different ink than the change indicated by the blue dot ('100' substituted by '150'), and further to this, both amendments were made with different ink with which the original text was written (red dot). Taking into account that medieval scribes produced their ink individually [Wat75, Hoh98], using traditional recipes, it can be concluded - if the simulation was based on a real medieval manuscript - that the sample was original text (red) was revised on two separate occasions using two dissimilar ink production methods which is possibly an indication of different scribes also. In the *Kundige Bok* case study we refer to here, this would be an important step forward towards a complete revelation of the textual evolution and with it the development of medieval town law in late 15th century-information that was not known before. Identifying segments of the text written with the same or different ink is only the first step, however, and does not solve the dating issues by itself. It must be accompanied by the expert's view on the manuscript. Dating requires one more piece of information. Consider, for instance, an entry that is undated but from which hyperspectral analysis reveals the same ink signature as a dated entry elsewhere in the book, or even a different source. It can then be dated by inference. However, building up a database of historical ink signatures in a certain (chronological and/or local) context, could establish the basis for an (semi-) automatically created facsimile edition of a manuscript (be it medieval or modern) by visualisation of the different stages (Figure 3.5) of the text and could also lead to automated markup preparation-catering, for example, as a tool for the creation of genetic editions.

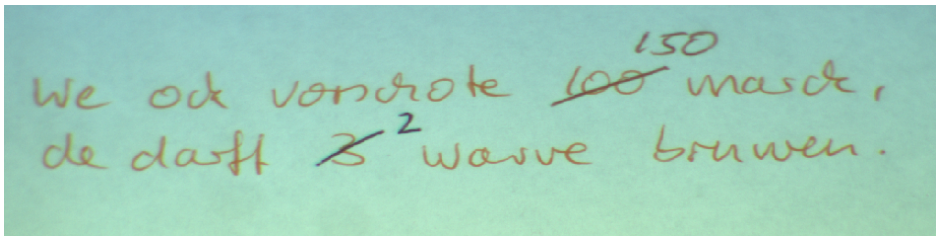


Figure 3.9: False coloured image identifying three different inks

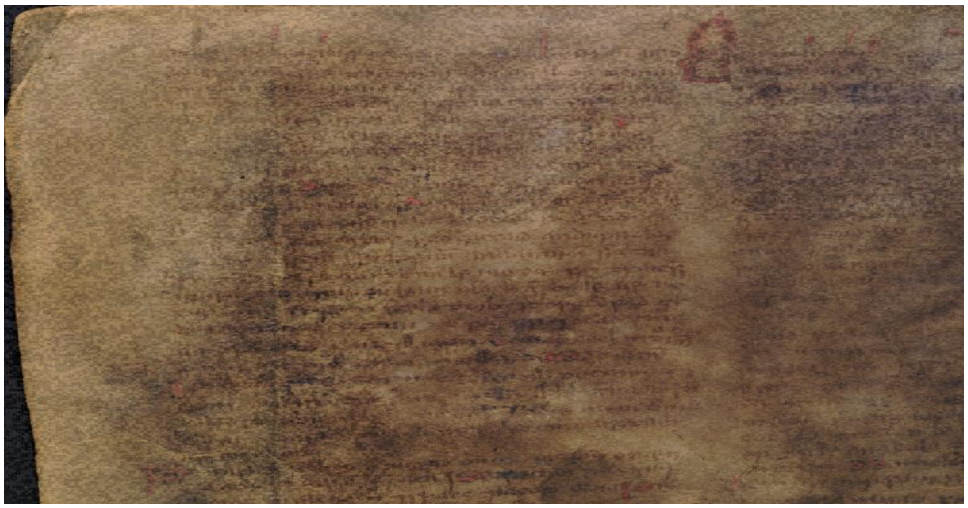
3.6 Hyperspectral Imaging for Faded Text Enhancement and Recovery

The conversion of large collections of historical documents into digital libraries and archives is a task currently been undertaken by document conservationists and archivists. This task encounters significant challenges that standard imaging systems cannot address[AC06]. For instance, figure 1.1(a) is a extract from the *Liber Flavus Fergusiorum* [uaryb], a 15^{th} Century manuscript written on vellum by numerous scribes containing mostly religious content. The manuscript is in good condition overall but with some pages faded over time. As the ink is extremely faded on certain pages of the document due to physical wear over a long period and the formation of mold, the process of interpreting the underlying text becomes difficult. The approach taken to enhance the degraded text was to hyperspectrally analyse the manuscript under a visible illuminating source, recording sequential images of the manuscripts reflected radiance at 10nm intervals from 400nms to 1000nms. The resulting images illustrate the documents reflectance at each specific wavelength. The contrast of reflected energy between that of the text

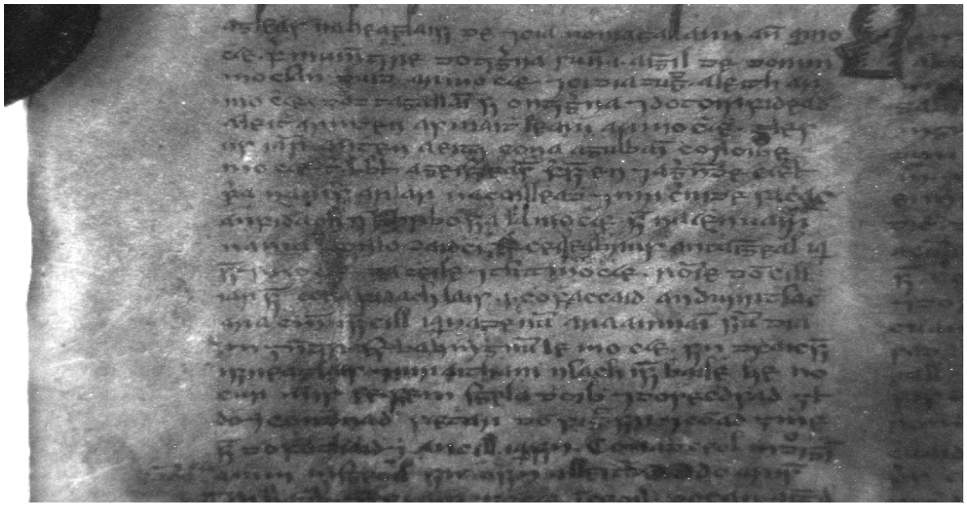
pixels and background/mould was found to be greatest at the lower end of the visible range, specifically 430nm (Figure 3.10(b)). In quantitative terms, these results prove to be a considerable improvement on previous attempt of document historical experts to read the manuscript's text. Initially, only a small fraction of the page was legible to experts due to the extent of the pages degradation. After hyperspectral analysis almost 100% of the text is now legible to experts in the documents history and to the naked eye. This is a significant advancement in recording the document in digital form and in terms of interpreting the manuscripts content.

3.7 Text Recovery using Fluorescence Spectroscopy

Luminescence is a phenomenon where an object emits light due to chemical reaction, electrical energy, subatomic motions, or stress. A particular type of luminescence, called *fluorescence*, is of particular interest when performing hyperspectral analysis of documents. Fluorescence occurs when an object emits a high wavelength (low energy light) following illumination by a shorter wavelength (higher energy light) due to molecular absorption of part of the incident light [Lak06]. The object absorbs part of the incident light which causes excitement of the molecules within the object and then emits energy in the form of lower energy light usually within the visible range of the spectrum. This example shown in section 3.11(a) demonstrates how the application fluorescence spectroscopy can be useful in recovering unreadable text from



(a) Faded, Removed Text - *Liber Flavus Fergusiorum (b), Royal Irish Academy, Dublin.*



(b) Enhanced Text, Reflectance captured at 430nms

Figure 3.10: Reflectance Spectroscopy for Faded Text Recovery

historical documents in our sample 16th Century book cover. In this thesis, we are especially interested in the use of hyperspectral analysis to support the recovery of hidden text, and in particular, the kind of recovery that

requires substantial conservation efforts, or the disassembly of manuscripts and their bindings. We believe that it is possible to employ non-invasive and non-destructive investigative techniques to search for hidden text prior to embarking on physical analyses and treatment of historical texts. In order

(a) Text Recovery - *16th Century Book Cover*

(b) Front Cover

Figure 3.11: Original 16^{th} Century Book Cover

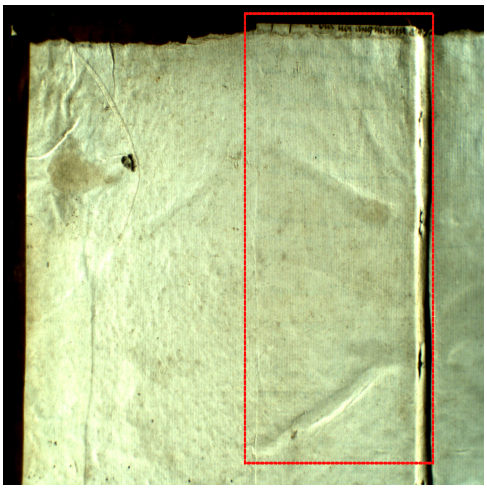
to investigate non-invasive hyperspectral techniques for fluorescence based text recovery, we obtained an exterior 16th Century book cover (located in the Russell Library at National University of Ireland Maynooth) that has become degraded with time and contains mold in places (Figure 3.11). The exterior cover, which is intact for the most part, is not of interest here. The interior cover's structure, shown in figure 3.11(a) consists of an underlying text which has been pasted over with a clean blank faced sheet of paper. Using fluorescence spectroscopy, i.e, light induced fluorescence in the pages it is possible to reveal the underlying text as shown in figure 3.12(b) and 3.12(c). The underlying text is assumed to be degraded with time but is unreadable

due to the presence of the overlaying, pasted down sheet. When the cover was illuminated with high energy light (specifically 505nm) it was found that this light produced fluorescence in the pasted down and underlying pages. Both pages absorb part of this incident light, and in turn emit a low energy (red) light with wavelength 720nm. Using an appropriate filter, our camera records a grey scale image of the resultant fluorescence intensities. As a result of the ink having very low fluorescence relative to the paper, any portion of the interior cover that contains the ink will not fluoresce at 720nm, and is represented by black pixels in our grey scale image, thereby revealing the underlying text. Following initial light spectroscopy images, we apply image processing techniques, such as thresholding, to further enhance the contrast between page and text and increase the legibility of the recovered text.

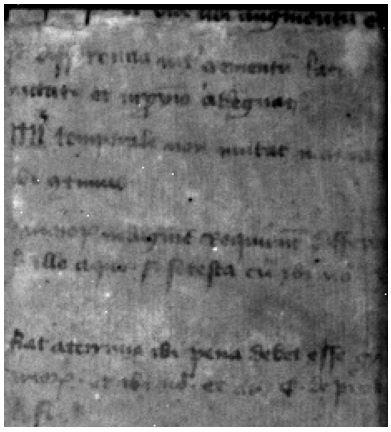
Almost all of the previously unread text contained in interior cover is now legible to librarians who have a particular interest in this manuscript. The results from this experiment show that fluorescence spectroscopy can prove to be a vital tool in the investigation of unreadable or hidden text for scholars and professionals in the field of historical text analysis and conservation.

3.8 Conclusion

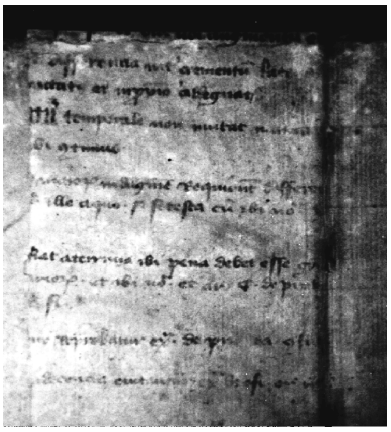
In this chapter we have demonstrated our successful application of spectral imaging methods as a practical tool for the examination of historical documents and, in particular, its effectiveness for enhancing the legibility of text that is obscured or faded over time. The method was also shown to be valuable in differentiating alterations of edits to text when seemingly



(a) Historical Book Cover



(b) Recovered Text using Fluorescence Spectroscopy



(c) Thresholded, Histogram Equalised Image

Figure 3.12: Fluorescence Spectroscopy for Text Recovery

similar inks have different optical properties. Vast improvements have been demonstrated in improving the visibility and legibility of historical text using both reflectance and fluorescence spectroscopy that has been unreadable or unexaminable with other methods.

Chapter 4

Processing of Hyperspectral Images for Feature Identification

4.1 Introduction to Dimensionality Reduction

Advances in hyperspectral sensor technology present the possibility to image a scene in hundreds of spectral bands, where each band covers a specific range of wavelengths. This provides a wealth of detailed spectral information of data that can be used in a wide varied of applications. Adding an extra dimension to the data causes an exponential increase in the volume of the data. While this provides more information to work with it also makes the task of processing the entire data significantly greater. This problem, also known as the curse of dimensionality[JR07], requires the application of dimension-

ality reduction algorithms to hyperspectral data. Dimensionality reduction algorithms are usually employed to serve two fundamental purposes[HC94]:

1. to reduce data volume and dimensionality for data processing purposes,

2. and to detect and classify objects for feature extraction.

Dimension reduction algorithms typically achieve these goals by projecting the high dimensional data to a lower dimensional space, without loss of significant discrimination information. This reduced dimensionality is known as the intrinsic dimensionality. Image transforms play a key role in almost all image processing and analysis applications [GW92, Jai89]. Typically, image chapter4s are used to convert an image from its initial form in a new more meaningful one. This more meaningful form can be a more interpretable image, an image with certain features emphasised or diminished, or an image of reduced dimensions. In this section, we shall describe techniques which can be implemented to reduce the data volume and dimensionality without loss of critical data in the hyperspectral scene. We shall also discuss how feature extraction can be achieved utilising the vast amount of spectral information in hyperspectral images.

4.2 Principle Component Analysis

Principle Component Analysis (PCA) is a linear transformation that has found application in many fields such as image compression and facial recognition due to its simplistic, non-parametric method of extracting relevant

information from high dimensional data[Shl09]. The fundamental purpose of PCA is to reduce dimensionality of a data set containing a large number of inter-related variables, while retaining as much of the variance present in the data set as possible.This is in order to reduce the dimensionality of the data set and to identify new meaningful underlying variables. PCA achieves this by identifying representative elements of the data, called Principal Components (PCs), where the variance is maximised[Jol02]. This reduction is done by transforming high dimensional data onto a new simplified set of variables, which are ordered by the amount of variation retained.

To perform PCA on a hyperspectral data cube, each of the stacked images, or image bands, is represented in a one dimensional vector, $\mathbf{x}$, where each row of image pixels is written sequentially in a vector of the following form:

$$x = (x_1, ..., x_n)^T$$

The mean of that vector is then denoted by

$$\mu_x = \frac{\sum_{i=1}^{n} x_i}{n} \tag{4.1}$$

Calculating the covariance matrix between two of such image vectors (x,y) can be done using the following formula:

$$Cov(x, y) = \frac{\sum_{i=1}^{n} (x_i - \mu_x)(y_i - \mu_y)}{n - 1} \tag{4.2}$$

The components of $Cov(x, y)$, denoted by Cov_{ij}, represent the covariances between the variable components x_i and y_i. The covariance is a measurement of how much the two images vary from the mean with respect to each other. A covariance matrix is calculated for the entire datacube; calculating the covariance between each image band, such that the covariance matrix for the entire hypercube is in the form:

$$C = \begin{pmatrix} Cov(b_1)(b_1) & Cov(b_1)(b_2) & \dots & Cov(b_1)(b_{n-1}) & Cov(b_1)(b_n) \\ Cov(b_2)(b_1) & Cov(b_2)(b_2) & \dots & \dots & Cov(b_2)(b_n) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ Cov(b_{n-1})(b_1) & \dots & \dots & Cov(b_{n-1})(b_{n-1}) & Cov(b_{n-1})(b_n) \\ Cov(b_n)(b_1) & \dots & \dots & Cov(b_n)(b_{n-1}) & Cov(b_n)(b_n) \end{pmatrix}$$

As $Cov(a, b) = Cov(b, a)$ this covariance matrix is, by definition, symmetric. Along the main diagonal the covariance value is calculated between on dimension and itself, resulting in the variance. In mathematical terms, to extract the most variant PCs the eigenvalues and eigen vectors muct be found, which can be derived from the covariance matrix. Although many eigen vectors may exist for a given covariance matrix very few of these shall be PCs. We are only interested in the most variant eigen vectors as these account for a considerable majority of disparity between the image bands. In general, a matrix acts on a vector by changing both its magnitude and its direction. However, a matrix may act on certain vectors by changing only their magnitude, and leaving their direction unchanged (or possibly reversing it). These vectors, which only change magnitude, are the eigenvectors of the matrix.

A matrix acts on an eigenvector by multiplying its magnitude by a factor, which is positive if its direction is unchanged and negative if its direction is reversed (figure 4.1). This factor is the eigenvalue associated with that eigenvector.

Let A be a square matrix. A non-zero vector b is called an eigenvector of A if and only if there exists a number (real or complex), λ, that satisfies the following:

$$Ax = \lambda x$$

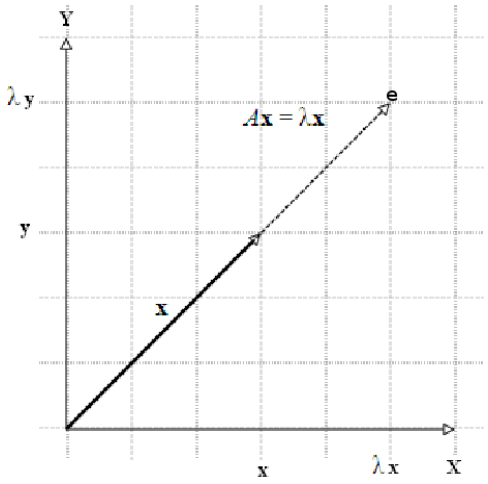


Figure 4.1: Eigen Vectors and Eigen Values

If such a number λ exists, it is called an eigenvalue of A. The non-zero vector, e, is called eigenvector associated to the eigenvalue λ. The eigen vectors e_i and the corresponding eigenvalues λ_i are the solutions of the equation:

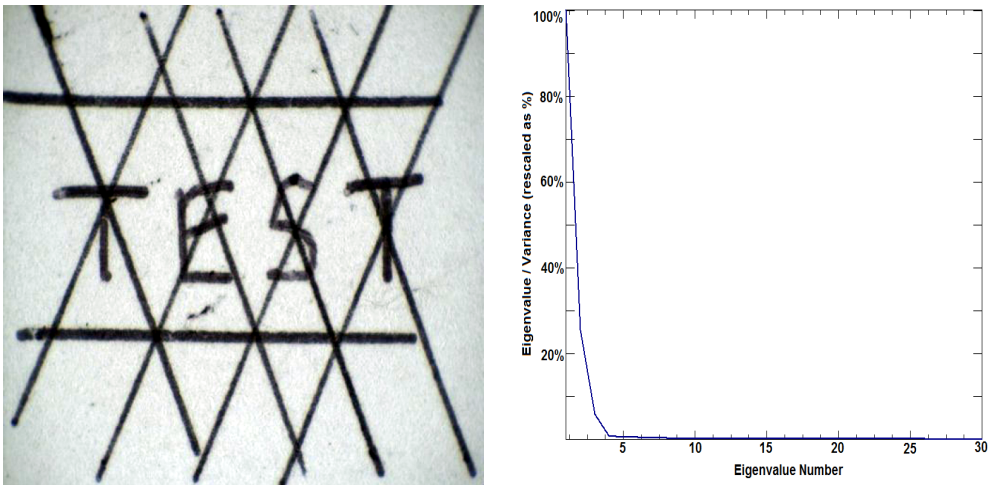
$$C_x e_i = \lambda_i e_i, i = 1, ..., n \tag{4.3}$$

This matrix of eigen vectors is know as a Feature vector. By ordering the eigen vectors in the feature vector in the order of descending eigenvalues (largest first), one can create an ordered orthogonal basis with the first eigen vector having the direction of largest variance of the data. In this way, we can find directions in which the data set has the most significant amounts of energy. Reconstructing the eigen vectors, from one dimensional vectors in the form $x = (x_1, ..., x_n)^T$ to two dimensional images, will result in our PC images, ordered in a manor so that the PC with the most variant component comes to being the first element, the second greatest variance is the second image, and so on.

4.3 Hyperspectral Ink Segmentation and Dimensionality Reduction using PCA

This example serves to illustrate how inks that appear identical under natural light can be distinguished. In this example we have written the word "TEST" in one black pen and crosshatched lines in a different black pen. We have illuminated the target with a constant white spot light. As the eye perceives the image the underwritten text, "TEST", and the overwritten crosshatched lines appear to be from the same pen (figure 4.2(a)). A segmentation can be achieved using PCA based upon differences in the inks respective reflectance.

The covariance matrix of the hyperspectral image data was calculated. Based on the formulation of the covariance matrix, the eigenvectors and eigenvalues were then resolved. Detailed in figure 4.2(b) is the eigenvalues

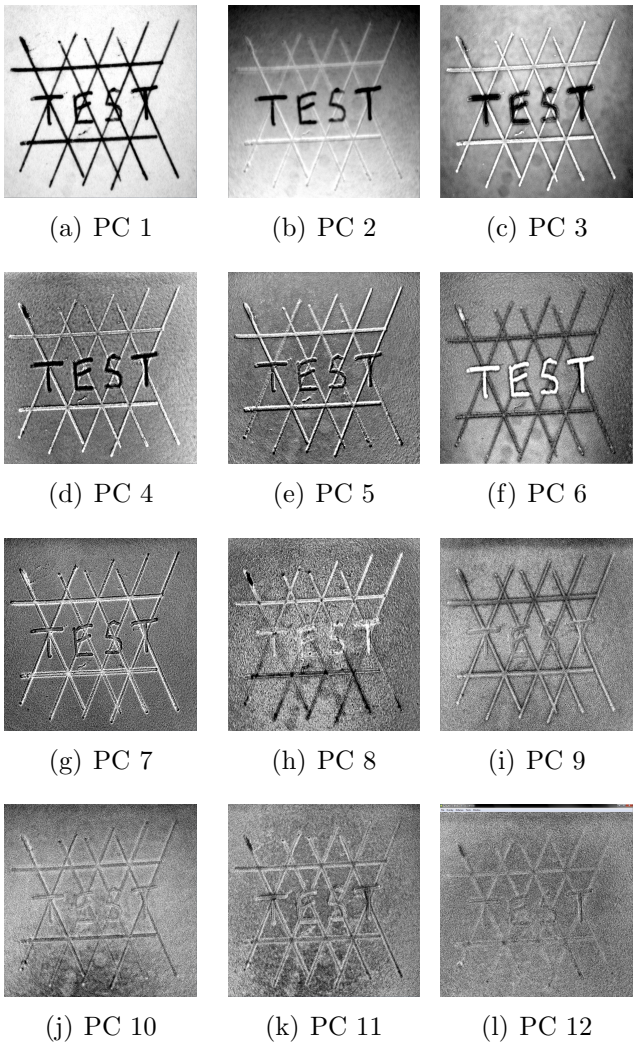


(a) RGB Image of example text (b) Eigenvalues and % variance

Figure 4.2: Hyperspectral segmentation using PC Analysis

plotted for each eigenvector. We can see from this that the vast majority of the images variance can be described by the first few PCs. Specifically, about 85% of the hypercubes variance can be described in the first 3 principal components and almost 98% of the entire images variance is contained in the first 6 PCs. For dimensionality reduction purposes choosing to ignore some of the less significant components we can construct an image with is a very accurate representation of the initial hypercube data, containing much of the variance, but in significantly less dimensions by just retaining these first 5-6 PCs. PCA performs a coordinate rotation that aligns the transformed axes with the directions of maximum variance. When performing PCA for forensic document analysis often the observed data has a high signal-to-noise ratio, and as such the principal components with larger variance usually correspond to features or objects of interest and lower ones correspond to noise. Features with similar variance will fall on the same principal component. Visualising

these principal components as two dimensional images will yield an image emphasising certain features in the image with a similar variance from the overall mean (figure 4.3). Besides the functionality of pca for dimension



(a) PC 1 (b) PC 2 (c) PC 3

(d) PC 4 (e) PC 5 (f) PC 6

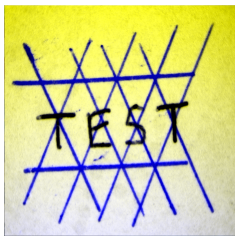
(g) PC 7 (h) PC 8 (i) PC 9

(j) PC 10 (k) PC 11 (l) PC 12

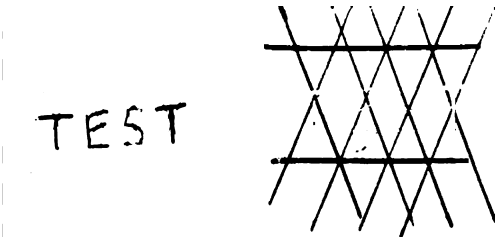
Figure 4.3: First 12 PCs of "Test" Example

reduction, it can also be used for feature extraction by grouping pixels with similar variance together. We can extract specific features of interest from

the hyperspectral images by selectively choosing principal components with specific features, and combining these into an RGB composite image, using PCs as the red, green and blue image channels (figure 4.4(a)). With the application of a false colouring algorithm (which colours similar intensities in an image in the same colour) and a process of thresholding to the image we can isolate specific features of interest and as a result extract these features from our image. This allows us to recover the underwritten text(figure 4.4(b)) or conversely only display the crosshatched lines(figure 4.4(c)).



(a) RGB Composite Image - Ink Segmentation (PC1, PC2, PC6)



(b) Underlying 'Test' Text

(c) Overwritten Crosshatching

Figure 4.4: Principal Component Analysis for Feature Extraction

4.4 Independant Component Analysis

Hyperspectral image pixels are often a combination of numerous different components of various substances. A spectrally mixed pixel consists of a number of spectral components, called end-members, mixed linearly together. These end-members represent a pure signature for a specific component or class. The majority of data analysis techniques attempting to separate mixed multiple signal sources are based on a differentiation of signal spectral differences [PSS+00]. This is termed as "spectral or pixel un-mixing", and it is used to determine the contribution of each material in a mixed pixel. Unmixing hyperspectral image data can be seen as an unsupervised method for blind source separation where the objective is to determine the contribution of each component in the mixed signal without prior knowledge of the sub components [BGC04]. Independent component analysis (ICA) is a method for unsupervised classification of hyperspectral imagery which provides a possible way to unmix the different components of a mixed pixel. It assumes that the mixed signals are close to statically independent. ICA can be applied to hyperspectral images where the data consists of linearly mixed signals, using all the statistical information not just the first order (mean) or second order (covariance) statistics, as used in principal component analysis (section 4.2). The ICA algorithm reduces the multi-dimensional hyperspectral data into independent parts (or basis functions) and identifies each independent component from looking at the given mixed observations with little or no prior information about the mixing parameters. The application of ICA to historical documents for restoration has yielded fruitful

results in the past, such as in [TSMB04b], where the recovery of underwriting has been achieved from synthetic examples of palimpsest text. The effects of bleed-through and show-through can also be minimised, and often removed, by utilising the ICA technique for documents. When the independence assumption is correct, blind ICA separation of a mixed signal gives very good results [TBS04]. To formulate the problem of unmixing hyperspectral data as that of a blind source separation problem we make the following assumptions about the hyperspectral data:

- We assume that each pixel (of index t in a total of T) in a hyperspectral scene has a vector value $\boldsymbol{x}(t)$ of N components, and that we have M superimposed sources (or layers in our image) at each pixel t, represented by the vector $\boldsymbol{s}(t)$ and each of these superimposed layers or sources is mutually statistically independent.

- As we are dealing with hyperspectral images of documents we also assume that each of these layers will behave almost uniformly across the wavelength region in terms of their reflectance. We also capture the hyperspectral data (described in section 3.3) in a controlled manor so that noise and blur can be neglected.

From these assumptions we now have a random N-vector $\boldsymbol{x}$, which is generated by linearly mixing the components of a M-vector $\boldsymbol{s}$ through an $N \times M$ matrix, $\boldsymbol{A}$, where the layer functions $\boldsymbol{s}_i$(t), i = 1,2, ... , M denote the amount of contribution of the M patterns to the intensity at point t. Estimating $\boldsymbol{s}(t)$

and $\boldsymbol{A}$ from $\boldsymbol{x}(t)$ is then a problem of blind source separation.

$$\boldsymbol{x}(t) = \boldsymbol{A}\boldsymbol{s}(t) \tag{4.4}$$

where t = 1,2, ... , T. If the prior distribution of each source was known, the joint prior distribution for $\boldsymbol{s}$ is given by

$$P(s) = \prod_{i=1}^{N} P_i(s_i) \tag{4.5}$$

where s_i = (s_i(1), s_i(2), ..., s_i(T)). The separation problem can then be formalised as the maximisation of equation 4.5, subject to the constraint in equation 4.4. Letting the unknown matrix A^{-1} be W, the problem simplifies to a search for a W = (w_1, w_2, ..., w_N)′, such that when this W is applied to the data x = (x_1, x_2, ..., x_N), produces the set of vectors, $\hat{s}_i = w'_i x$ that are as nearly statistically independent as possible, and whose distributions are given by P_i. By taking the logarithm of equation 4.5, the problem solved by the ICA algorithm[TBS04] is given by

$$\hat{W} = \arg\max_{W} \sum_{t} \sum_{i} \log P_i(w_i' x(t)) + T \log |\det(W)| \tag{4.6}$$

4.5 Blind Source Seperation of Hyperspectral Data using ICA

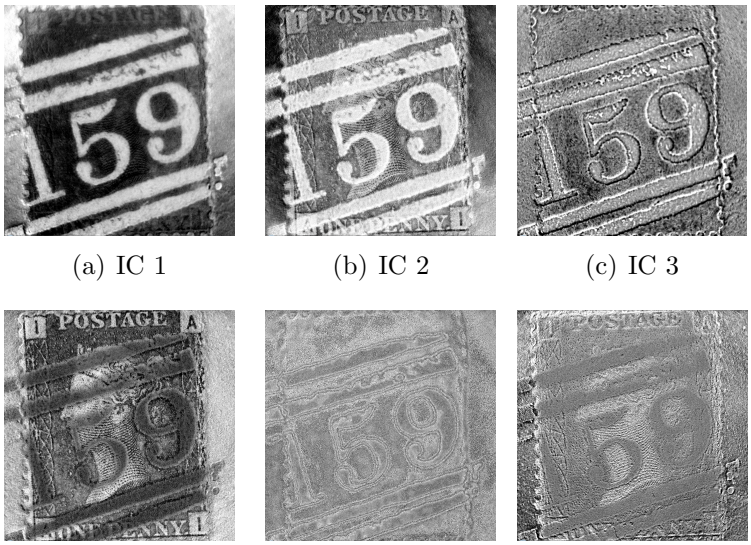
The main idea of ICA when applied to images is that each sub image (resulting independent component) may be perceived as linear addition of features

$a_i(x, y)$ weighted by the coefficients s_i. Features are represented by columns of mixing matrix A and s_i are elements of appropriate sources [VJ06]. For the purpose of investigating the potential of blind ICA for the enhancement of historical documents we have chosen an example which incorporates an obliteration of a specific form. In many historical manuscripts and text, unwanted text was not removed, but merely obliterated by covering the section of unwanted text with a block of ink. In our example we have a postage stamp which has been ink stamped for official purposes but serves to illustrate this scenario where interesting underlying features have been obscured by an overlying layer of ink. The purpose of performing the ICA algorithm

Figure 4.5: Obliterated, Obscured features - RGB digital image of Postage stamp

on the hyperspectral image obtained of the stamp is not necessarily to reduce the dimensionality of the hyperspectral image but to reduce the image to a smaller subset of images (its independent components) which contain statistically independent features respectively. Thus, achieving a blind source separation of hyperspectral data. With respect to images, and specifically

hyperspectral images, the aim of dissolving an image into statistically independent components (ICs) is to automatically segment and extract an image into features which contain similarities. The following is the results of the ICA algorithm for the first 6 independent components. We can



(a) IC 1 (b) IC 2 (c) IC 3

(d) IC 4 (e) IC 5 (f) IC 6

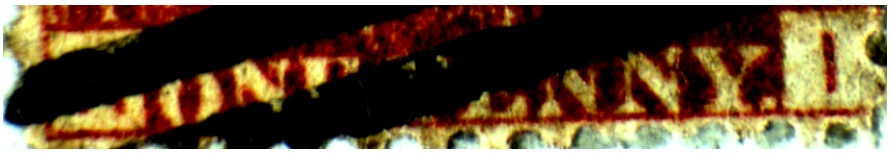
Figure 4.6: First 6 ICs of "stamp" Example

see that particular ICs have enhances certain features in the hyperspectral scene. Specifically, detailed examination of the first and fourth IC produces an unmixing of the overlapping elements or features in the image. The first IC has augmented the over-pattern (i.e. the postal numerical stamp 159, figures 4.7(c) and 4.7(b)) and conversely the fourth independent component has augmented the under-pattern (i.e. the queens head and associated markings, figures 4.7(f) and 4.7(e)). Visual inspection of the results of the ICA algorithm show a vast improvement in the ability to make out sections of

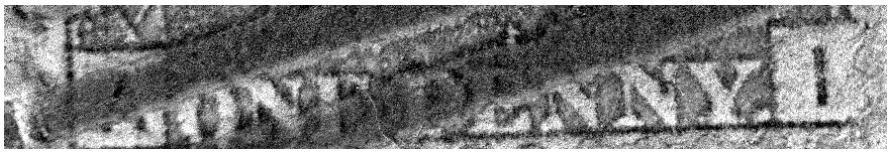
(a) Original RBG Image (Section)

(b) Under-Pattern augmented

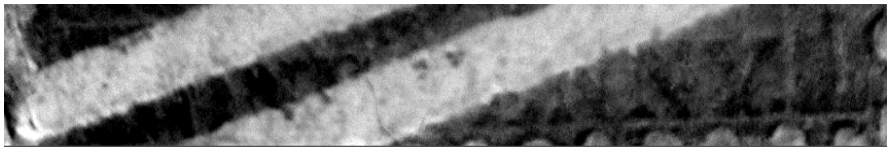
(c) Over-Pattern augmented



(d) Original RGB Image (Section)



(e) Detailed section of Under-Pattern



(f) Detailed section of Over-Pattern

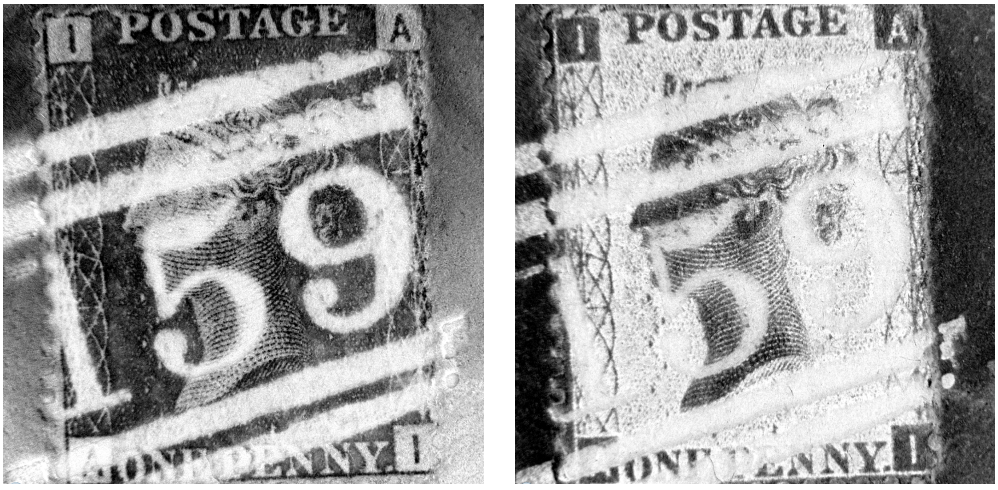
Figure 4.7: Practical Application of ICA for historical documents

the underlying pattern which have been obscured or obliterated by the interfering over-pattern. This constitutes a vital tool in the analysis of historical manuscripts and documents that may contain obliterations such as these and the ability to uncover underwritten and previously unreadable text.

4.6 Transformation Results and Discussion

The intention of this chapter is to demonstrate different applications of the above mentioned methods in hyperspectral image processing. Both Principal Component Analysis and Independent Component analysis (ICA) are transformations that rely on statistics of the given data set. PCA is based on the second order statistics whereas ICA exploits higher order statistics in the data. This technique has been employed for a variety of applications such as blind source separation, speech enhancement, and data mining. An important point to note about PCA and other image transforms is that while it computes the axes of maximum overall variance, there is no guarantee that it will actually increase the separation between a particular pair of classes or features in an image[SGS+97]. It only guarantees to maximize the overall variance givenIndependence component analysis, in the context of blind source separation, aims to recover components, that are as statistically independent as possible. As ICA is based on higher order statistics it can be assumed that the resulting components extracted by ICA will be more significant and expressive than that of PCA. From section 4.4, we have seen that ICA can be an effective feature extraction tool for hyperspectral data separating class or source information into different bands. However, the

number of bands contained in hyperspectral imagery makes the ICA algorithm a time expensive task. Although not applicable in this experiment, noise may also have a large factor on the effectiveness of the ICA algorithm. It has been shown[VA04] that ICA works better on data that has been preprocessed with PCA, where an improved hybrid application of both the PCA and ICA algorithms has been proposed. Following from this we have initially performed PCA on the same sample data set as in section 4.5 and then preceded to perform the ICA algorithm on the resulting principal components. The following results were obtained: As no interfering sources such as noise



(a) PCA-ICA Component 1 (b) PCA-ICA Component 2

Figure 4.8: Hybrid PCA and ICA Automatic Hyperspectral Segmentation

or atmospheric conditions were recorded in our sample data set, the PCA-ICA algorithm offer only marginal benefits in only very subtle differences to ICA alone in terms of a separation of the unmixed classes. In fact, for the purposes of segmenting the layers of our image into an underwritten-pattern and an overlying one, the blind ICA algorithm provided quicker and almost

as legible segmentation for reading the underlying pattern. As a result of these experiments we have shown that both PCA and ICA prove to be attractive techniques to providing an automatic segmentation of hyperspectral data.

Chapter 5

Classification Techniques for Hyperspectral Document Segmentation

5.1 Introduction

The general objectives of cursive text segmentation include tasks such as word spotting, text/image alignment, authentication and extraction of specific fields. An important step associated with all of these tasks is the segmentation of the document into logically units, for instance text lines, words or letters. In general, this is difficult due of the low quality and the complexity of these documents, and automatic text segmentation of such kind is an open research field [BMC09]. Sophisticated image processing of single-image documents is hence the norm so far. Here we describe our recent approach

towards segmentation of a different kind, which we refer to as hyperspectral segmentation; the technique is based on the separation and segmentation of different inks by recording and analysing their reflectance properties. When dealing with hyperspectral images, image classification is a means to convert the entire spectral raster data set into a finite set of classes that represent the features seen in the imagery respectively. As discussed in the previous chapter, an important issue associated with the classification of hyperspectral images is the large size of data produced by current hyperspectral imaging systems. For this reason, pre-processing steps such as dimensionality reduction are often taken prior to performing a classification of hyperspectral imagery. The vast majority of the most variant features in hyperspectral imagery can be represented and summarised in just a few images, which is often achieved by the application of principal component analysis to reduce the dimensionality. Often, one wishes to go one step further than reducing hyperspectral images to their principal components; one desires to reduce the entire data cube to a single segmented image in which one of a small number of digital grey levels is assigned to each pixel [SRC02]. Thus, classifying all pixels in the image into a specific number of key classes that can be used to identify or distinguish individual features occurring in an image. Hyperspectral image classification has been employed to great effect in the area of remote sensing and can be used to identify vegetation, mineral sources, and to plot geographical coast lines accurately [Cha03]. Additionally, classification can provide a means to geometrically allocate areas or features in the imagery for the calculation of spatial attributes such as area, length and perimeter

of specific objects in a remotely sensed image. Similar techniques have been applied to the field of document analysis to distinguish ink differences and enhance extrinsic element of a document such as watermarks and frame patterns [KAP+08]. Hyperspectral classification and indeed most image classification algorithms can be categorised three main groups - unsupervised classification, supervised classification and a hybrid approach combining the merits of both methods. Typically, supervised classification methods require prior manual identification and classification of a small subset of known features within the image data. This manual classification of a small subset of the image is then used as a training set for a learning algorithm. A statistical package is used to determine an average representative spectral signature for the training set's identified class, which is then used to classify the remainder of the image. An unsupervised classification scheme uses spatial statistics to classify the image into a predetermined number of categories or classes. These classes are statistically significant within the imagery, but may not represent actual surface features of interest. Hybrid classification combines both techniques to improve the accuracy and efficiency of the classification process. This chapter presents both automatic and semi-supervised methods of classification and segmentation of hyperspectral imagery of historical documents and manuscripts. The historical document hyperspectral images have been generated by imaging a section of a the historically significant *Lebor na hUidre* or *The Book of the Dun Cow* courtesy of the Royal Irish Academy [uarya]. The manuscript is the earliest surviving manuscript written in Irish and contains the oldest version of many historical, biblical, and

literary materials. The surviving manuscript is in poor condition with some of the pages containing almost completely faded text. The extract chosen contains legible text, faded text, mould and other noise patterns and as such is a good exemplar for examination. The historical document hyperspectral images have been generated by imaging a section of a the historically significant *Lebor na hUidre* or *The Book of the Dun Cow* courtesy of the Royal Irish Academy [uarya]. The manuscript is the earliest surviving manuscript written in Irish and contains the oldest version of many historical, biblical, and literary materials. The surviving manuscript is in poor condition with some of the pages containing almost completely faded text. The extract chosen contains legible text, faded text, mould and other noise patterns and as such is a good exemplar for examination providing a reasonable variety of legibility problems on which to test our hyperspectral segmentation techniques. The optical instrumentation used for examination is identical to that used throughout this thesis and described in detail in section 3.3. The hyperspectral image cube acquired from the manuscript consisted of 30 images measuring the reflectance of the manuscript taken at wavelength intervals of 20 nanometres from 400 to 1000nm.

5.2 Unsupervised Segmentation using PCA and K-Means Clustering

For classification purposes the data can be analysed in two different ways. The first method is to generate classes based on a supplied training set of

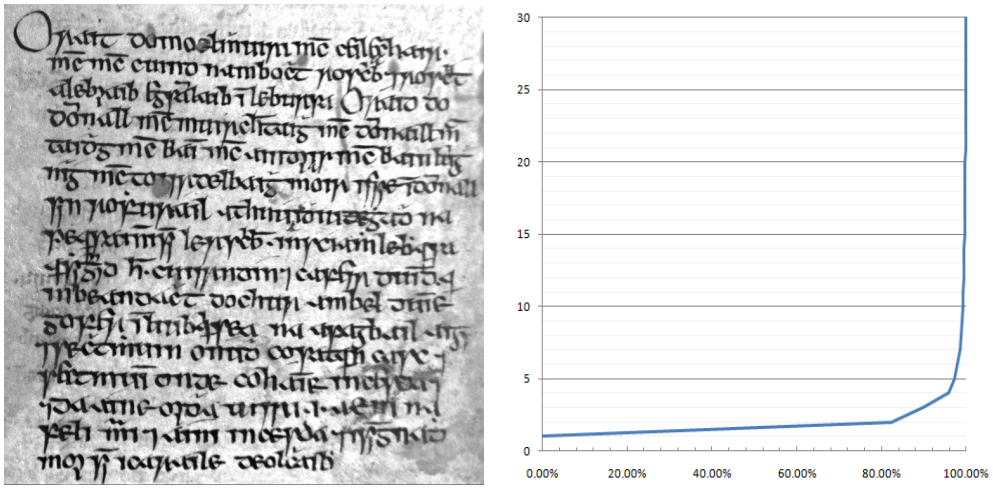
data. The second method is an unsupervised classification of the data set. Unsupervised classification clusters a dataset based on statistics only and classes are formed and the image is segmented based only on a fixed number of chosen classes. Here, we present an unsupervised method of hyperspectral image segmentation for historical documents based on a clustering of principal component data. PCA has successfully been used as a preprocessing step to classification algorithms to reduce the dimensionality of datasets [GSGS+02]. K-Means clustering [KMN+02, DJ88] is an unsupervised classification technique, which initially defines class centroids or means distributed in the dataset, then iteratively clusters the pixels into the nearest class using a minimum distance technique. Each iteration of the algorithm redefines the means of each class and reclassifies each pixel with respect to the new class means. The objective of K-means is to minimise the *sums of squares distances* (errors) between each pixel and its assigned centroid.

$$SS_{distances} = \sum_{\forall x} [x - C(x)]^2$$

Where *C(x)* is the mean of the cluster that pixel x is assigned to.

Each pixel is assigned to the nearest class and this process continues until the number of pixels in each class changes by less than a predefined change threshold or the maximum number of iterations is reached. The ISODATA algorithm is similar to the k-means algorithm with the distinct difference that the ISODATA algorithm allows for different number of clusters while the k-means assumes that the number of clusters is known a priori. K-means clustering is a simple, yet computationally expensive algorithm, particularly

when clustering large datasets such as hyperspectral images [dSFFdA+99]. Principal component analysis was applied as an initial pre-processing step to reduce the dimensionality of the dataset. As we can see from figure 5.1(b),

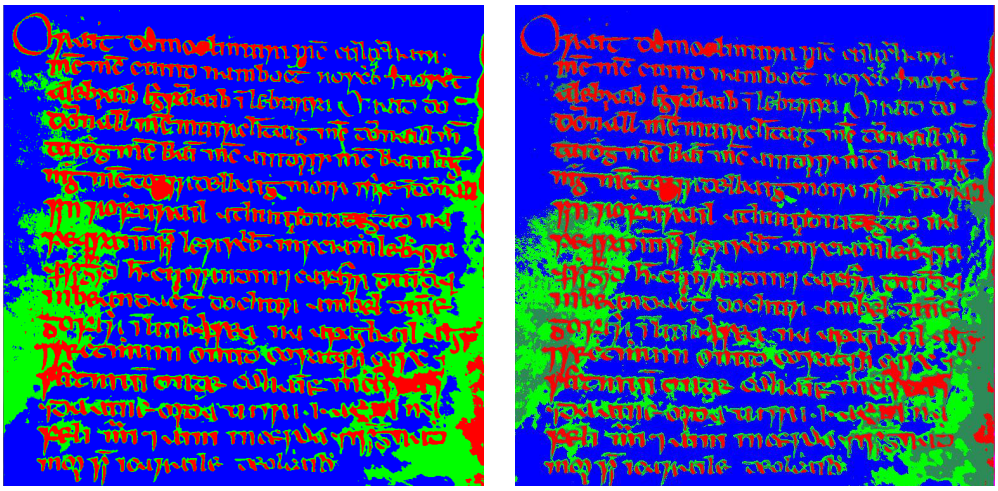


(a) Lebor na hUidre Extract (b) Principal Components and % variance

Figure 5.1: Dimensionality Reduction using PCA

much of the variance of the entire dataset can be described in the first few principal components. Specifically, 98.4% of the overall variance in the dataset is contained in the first 6 principal components. To reduce the dimensionality and amount of redundant data in the dataset we select only the first 6 principal components for our classification process. Following the dimensionality reduction, we apply the k-means clustering algorithm to the dimension reduces data (i.e. the first 6 principal components), varying the number of clusters and distance threshold. The optimal classification was the result of the principal component technique and a K-means clustering with 3 clusters. The ISODATA method of classification did provide good results in clustering mould pixels (green) but the optimal classification of

text pixels was achieved using the K-Means algorithm. Segmentation results obtained using the ISODATA algorithm is presented also 5.2. To improve these initial segmentation results, supervised methods of classification have been employed and examined in the next section at the cost of the process no longer being automatic. The results of clustering or classification of doc-



(a) K-Means Clustered Image (3 clusters) (b) ISODATA Clustered Image

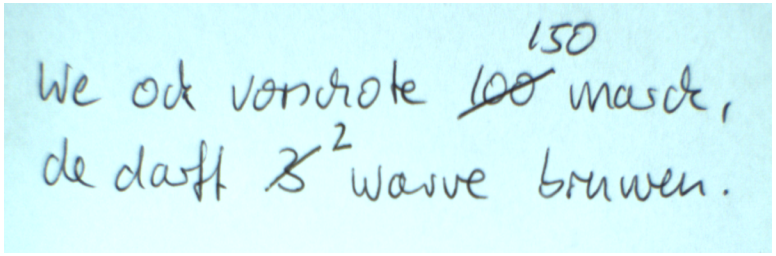
Figure 5.2: Automatic Unsupervised Classification for Segmentation of Hyperspectral Images

ument images are rarely perfect. Numerous factors affect the classification results when working with historical documents, with important ones being the objective of the classification, the spectral and spatial characteristics of the data, the natural variability of document and ink and the digital classification technique employed. Frequently, the classification effort may require preparatory processing prior to classification and the refinement of classes after classification. This method based on principal component analysis and k-means clustering provides a automatic method of segmentation. Further

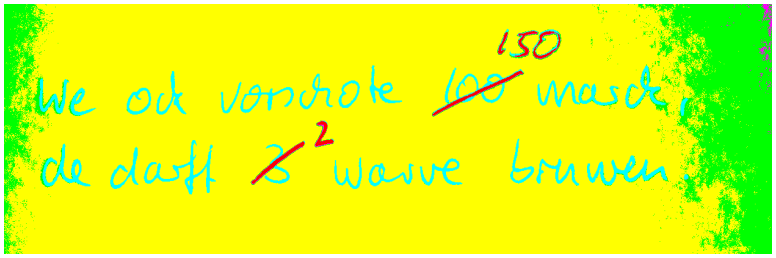
investigations using our simulated pen example from an earlier chapter (figure 3.5) shows that our method of hyperspectral imaging using PCA and clustering algorithms can be used to automatically segment hyperspectral images based fundamentally on a difference in spectral signature. These results provide an automatic segmentation of a hyperspectral scene where individual inks look identical under natural light. As we know from earlier, three different pens were used to create the simulation. Our method correctly segments the underlying text (light blue) from the overlying amendments (red). This provides a more detailed method of investigation and identification of inks beyond visual inspection. The drawback in this example is that although the amendment on the second line (strikethrough and '3' amended to '2') has been classified perfectly the other amendment (strikethrough and '100' amended to '150') has been classified as a mixture of both overlying and underlying text(figure 5.3). This shows that it is different to the underlying text; however, a perfect classification would classify this amendment as its own new class. This can be corrected by altering the amount of clusters used in the clustering stage or using supervised classification algorithms as described in the next section.

5.3 Improving Segmentation with Supervised Classification

Supervised classification of data refers to a classification function that is learned from, or fitted to, a user defined training dataset [RJ06]. Sufficient



(a) Simulation of Kundige Bok - *an example of layered text*



(b) Automatically Segmented Image

Figure 5.3: Segmentation of Simulation of Kundige Bok

training data must be supplied to classify areas of interest using supervised classification methods for hyperspectral data, which forms a basis from which a learning rule is employed to identify a class for the remainder of the input data. Training data for from our input data was constructed by selecting manually identifying and selecting representative or prototype pixels (and their associated spectra) for each desired class in our PCA pre-processed image. This forms the basis from which we apply supervised classification techniques to classify the remainder of the hyperspectral document image. The training set is used to estimate the parameters of a particular algorithm being used. Various comparison methods are then used to determine if a specific pixel qualifies as a member of a class. The performance of four classification methods has been evaluated for the purpose of segment-

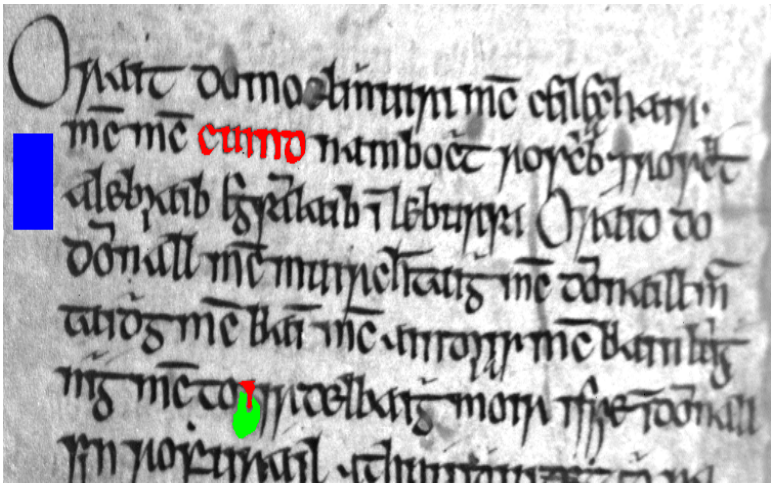


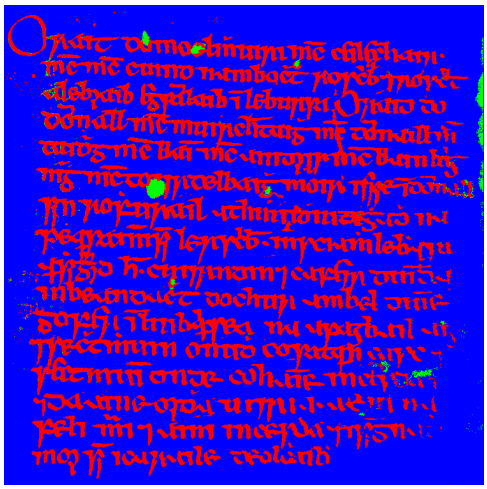
Figure 5.4: Training Dataset: Identifying representative class pixels

ing the data, namely Maximum Likelihood Classification, the Mahalanobis Distance Classifier, Minimum Distance Classification, and the Parallelepided Classifier. The desired classes, defined manually above (figure 5.3), form the common input training dataset supplied to all classification algorithms. The objective of this classification process is to identify and isolate three main classes in our hyperspectral data, the main text (red), the faded ink (green) used to embellish or decorate specific letters in paragraphs, and finally the background parchment (blue). Maximum Likelihood Classification (MLC) is a commonly used supervised classifier for multispectral and hyperspectral imagery [FCTW92]. The classifier assumes the statistics for each class are normally distributed throughout each hyperspectral band and calculates the probability that a pixel belongs to a specific class. The maximum likelihood classifier uses a Gaussian threshold stored in each class signature to determine if a given pixel falls within the class or not. The Mahalanobis Distance

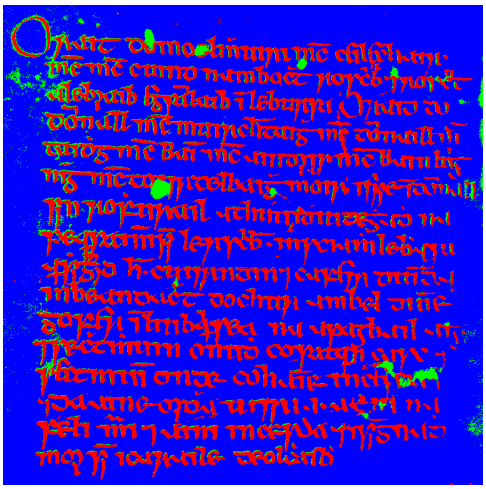
Classifier (MDC) is similar to the MLC but it assumes that the covariance between all classes is equal. The significance of this is that if is computationally faster. The effectiveness of the MLC classifier depends on a good estimation of the mean vector for each class and the covariance matrix [RJ06], which in turn depends on the amount of training data supplied. If the training data is limited it can be beneficial to employ other supervised classification methods such as Minimum Distance Classification (MDC), which does not rely on covariance matrix estimation. The minimum distance classification uses the mean vectors of each specified feature class from the training data and calculated the Euclidean distance from each unknown pixel to the mean vector for each class. Pixels are then classified in to the closest feature class. This technique is computationally less expensive than that of the MLC. Another supervised classification method, Parallelepiped Classification (PPC), uses a simple decision rule to classify multispectral data. This classifier defines class limits or decision boundaries to form an n-dimensional parallelepiped classification to determine if a given pixel falls within the class or not. If the pixel falls inside the parallelepiped, it is assigned to the class. The parallelepiped classifier is typically used when speed is required. The negative aspect is (in many cases) poor accuracy and a large number of pixels unclassified or classified as more than one class.

5.4 Classification Results and Discussion

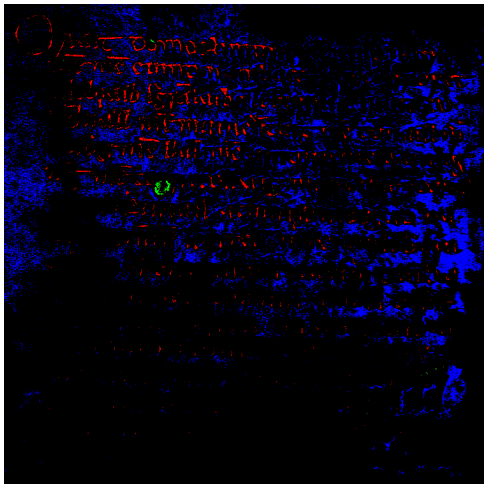
Presented in this chapter is the automatic segmentation of hyperspectral historical document images using hyperspectral imaging, principal component



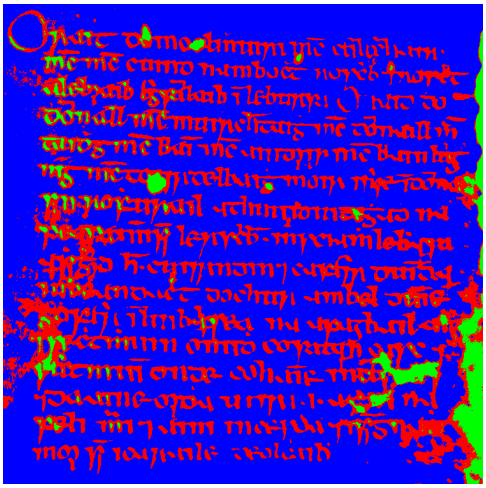
(a) Maximum Likelihood Classification (MLC)



(b) Mahalanobis Classifier



(c) Parallelepiped Classification (PPC)



(d) Minimum Distance Classification (MDC)

Figure 5.5: Supervised Classification and Segmentation

analysis and unsupervised k-means clustering. The results obtained provide a automatically segmented image of the principal features in the extract of the historically significant Irish text - *Lebor na hUidre*. To improve on the

classification result, supervised methods of classification were utilised as a replacement for the unsupervised methods. This generated a method of classification yielding a far superior classification with the drawback that the process is no longer and automatic one and requires some user input. The performances of four supervised methods of classification have been investigated. Maximum likelihood classification provided the best segmentation of the specified classes but was computationally more expensive than that of any other inspected method of classification. The classification of hyperspectral imagery of historical documents faces many obstacles. The vast majority of classification techniques assume that different features of interest that will also differ statistically. As such the application of statistical classification techniques to degraded historical documents in particular is limited. Classification techniques can and have proved useful in the segmentation of hyperspectral imagery for many purposes and have been fruitful in the segmentation and classification of features in our image 5.2. The difficulties associated with the classification of historical documents are that the principle features often become faded to an extent that they also become statistically similar to that of background or insignificant features in the imagery and are classified as such. Post-processing techniques such as principal and independent component analyses, presented in previous, chapters may generate greater results for analysts looking to enhance or isolate faded or degraded features. However, the classification technique presented in this chapter is not without merit. A successful automatic segmentation of historical documents can be achieved, and it has been shown that these results

can be improved with the aid of training data and supervised classification methods.

Chapter 6

Conclusion

In this dissertation, we have examined the application of hyperspectral imaging, and related computational techniques, to the examination of questioned historical documents. The use of hyperspectral imaging for the analysis of historical document has been inspired by its successful application of the classification and feature extraction of remotely sensed data [Cha03, TB92]. We present hyperspectral methods for feature extraction and segmentation, ink distinction for scribe attribution and relative aging of writings, faded text recovery and the diminishing of other deteriorations, all of which are vital for scholars to uncover as much knowledge of historical documents as possible. The research initially investigates and outlines difficulties associated with the examination of historical or fragile documents and how the transformation of these problems into a digital form can often results in their solution. Preservation of fragile historical documents is of the upmost importance to both librarians and archivists looking to conserve a document in its current

state for as long as possible, but also to scholars who utilise historical documents as primary sources. As continued usage of documents is itself a cause of deterioration, often a trade-off exists for librarians between the provision of access of historical documents and their conservation. We identify the digitisation process of hyperspectrally imaging documents such as these as a good resolution of this problem. Much is gained inherently by the process of digitisation of documents such as the ability to make documents available worldwide through the internet and other means. The digitisation process also provides a means for digital conservation and ease of duplication. Hyperspectral Imaging, in particular, is particularly suited to the digitisation of historical and fragile documents due to the non-destructive method of image acquisition and the wealth of data that is obtained. Hyperspectral Imaging is particularly suited as both representation and as a dataset to the digitisation of historical and fragile documents due to the non-destructive method of image acquisition and the wealth of data that is obtained. This research has also shown added benefits of spectroscopy and hyperspectral imaging for further more detailed examination of historical documents. Our experimental investigations demonstrate the advantages of high-spatial reflectance and fluorescence spectroscopy measurement for the non-invasive examination of historical documents. In particular, it can support codicology research by revealing binding structure of a codex or comparing and differentiating ink signatures, and paleographic research by making visible hidden text or by giving support to identify scribes and to solve dating issues. The post processing of hyperspectral digital representation of documents has also been demon-

strated as a vital tool for the forensic analysis of writings and paintings that have unsolved historical questions associated with them. We believe that the inclusion of hyperspectral imaging devices as standard research equipment for usable non-destructive analysis of historic documents is both affordable and attainable and would encourage humanities research institutes, libraries and archives to invest in the technologies, methodologies and proficient personnel to maximise their potential. Furthermore, we believe that equitable humanities computing partnerships are an essential component in hyperspectral imaging projects in order to provide realistic "use cases" for the development of the necessary software tools to support disruptive codicology and paleography research.

Appendix A

Appendices

A.1 Contributing Publications

The following is a list of the contributing peer-reviewed publications to this thesis from the author.

Book Chapters

1. **P. Shiel**, M. Rehbein, J.G. Keating "The Ghost in the Manuscript: Hyperspectral Text Recovery and Segmentation", *Codicology and Palaeography in the Digital Age* , ISBN 978-3-8370-9842-6, Books on Demand, Norderstedt 2009.

International Conferences

1. **P. Shiel**, M. Rehbein, J.G. Keating "Hyperspectral Text Recovery and Segmentation", *Digital Humanities Conference, University of Maryland, Baltimore, MA, USA.* , 2009.

Bibliography

[AC06] A. Antonacopoulos and C. Casado Castilla. Flexible text recovery from degraded typewritten historical documents. In *Proceedings of the 18th International Conference on Pattern Recognition (ICPR2006)*, pages 1062–1065. IEEE-CS Press, August 20-24 2006.

[AF06] C. Arms and C. Fleischhauer. Sustainability of digital formats: Planning for the library of congress collections. Online, July 2006.

[Att05] Michael Attas. Enhancement of document legibility using spectroscopic imaging. *Archivaria*, 57:131–144, 2005.

[Aut07] Israel Antiquities Authority. The dead sea scrolls go digital, 2007.

[BAMCM97] Adel Belouchrani, Karim Abed-Meraim, Jean-Francois Cardoso, and Eric Moulines. A blind source separation technique using second-order statistics. *IEEE TRANSACTIONS ON SIGNAL PROCESSING*, 45(2):434–444, 1997.

[Ban02] Helmut Bansa. Accelerated aging of paper: Some ideas on its practical benefit. *Restaurator*, 23.2:106117, 2002.

[BBD+06] X. Briottet, Y. Boucher, A. Dimmeler, A. Malaplate, A. Cini, M. Diani, H. Bekman, P. Schwering, T. Skauli, I. Kasen, I. Renhorn, L. Klasén, M. Gilmore, and D. Oxford. Military applications of hyperspectral imagery. *Targets and Backgrounds XII: Characterization and Representation*, 6239:62390B, 2006.

[BCLP98] S. Baronti, A. Casini, F. Lotti, and S. Porcinai. Multispectral imaging system for the mapping of pigments in works of art by use of principal-component analysis. *Applied Optics*, 37(8):1299–1309, 1998.

[BGC04] Jessica D. Bayliss, J. Anthony Gualtieri, and Robert F. Cromp. Analyzing hyperspectral data with independent component analysis. In *Proceedings of SPIE*, volume 3240, 2004.

[BM97] Irving J Bigio and Judith R Mourant. Ultraviolet and visible spectroscopies for tissue diagnostics: fluorescence spectroscopy and elastic-scattering spectroscopy. *Physics in Medicine and Biology*, 42:803814, 1997.

[BMC09] Brian D. Bue, Erzsebet Merenyi, and Beata Csatho. Automated labeling of segmented hyperspectral imagery via spectral matching. In *Hyperspectral Image and Signal Processing:*

Evolution in Remote Sensing, 2009. WHISPERS '09. First Workshop on, pages 1–4, 2009.

[BPD07] Vincent Baeten, Juan Antonio Fernandez Pierna, and Peirre Dardenne. *Techniques and Applications of Hyperspectral Image Analysis*, chapter Hyperspectral Imaging Techniques: an Attractive Solution for the Analysis of Biological and Agricultural Materials, pages 289–307. Wiley, 2007.

[BPP+03] Costas Balas, Vassilis Papadakis, Nicolas Papadakis, Antonis Papadakis, Eleftheria Vazgiouraki, and George Themelis. A novel hyper-spectral imaging apparatus for the non-destructive analysis of objects of artistic and historic value. *Journal of Cultural Heritage*, 4:330–337, 2003.

[BS96] G.H. Bearman and S.E. Spiro. Archaeological applications of advanced imaging technologies. *The Biblical Archaeologist*, 59:56–66, 1996.

[Cha03] Chein-I. Chang. *Hyperspectral Imaging: Techniques for Spectral Detection and Classification.* New York: Kluwer Academic, 2003.

[Cha07] Chein-I. Chang. *Hyperspectral Imaging: Signal Processing Algorithm Design and Analysis.* New York: John Wiley and Sons, 2007.

[CLP+99] Andrea Casini, Franco Lotti, Marcello Picollo, Lorenzo Stefani, and Ezio Buzzegoli. Image spectroscopy mapping technique for non-invasive analysis of paintings. *Studies in Conservation*, 44:39–48, 1999.

[Coa09] Digital Preservation Coalition. Definition, 2009.

[DB91] Bobby C. Deaton and William L. Balsam. Visible spectroscopy; a rapid method for determining hematite and goethite concentration in geological materials. *Journal of Sedimentary Research*, 61:628–632, 1991.

[DJ88] R. C. Dubes and A. K. Jain. Algorithms for clustering data. Prentice Hall, 1988.

[dR92] Dianne Van der Reyden. Recent scientific research in paper conservation. *The Journal of the American Institute for Conservation*, 31:117138, 1992.

[dSFFdA+99] Abel Guilhermino da S. Filho, Alejandro C. Frery, Cristiano Colho de Arajo, Haglay Alice, Jorge Cerqueira, Juliana A. Loureiro, Manoel Eusebio de Lima, Maria das Graas S. Oliveira, and Michelle Matos Horta. Hyperspectral images clustering on reconfigurable hardware using the k-means algorithm. In *Proceedings of the 16th symposium on Integrated circuits and systems design*, 99.

[FCTW92] G. M. FOODY, N. A. CAMPBELL, N. M. TRODD, and T. F. WOOD. Derivation and applications of probabilistic measures of class membership from the maximum-likelihood classification. *Photogrammetric engineering and remote sensing*, 58:1335–1341, 1992.

[FK06] Christian Fischer and Ioanna Kakoulli. Multispectral and hyperspectral imaging technologies in conservation: current research and potential applications. *Reviews in Conservation*, 7:3–16, 2006.

[GG07] Hans F. Grahn and Paul Geladi. *Techniques and Applications of Hyperspectral Image Analysis*. Wiley, 2007.

[GSGS+02] C. Gurschler, G. Seerafino, A. Del Bianco G. Spck, M. Kraft, and A. Kulcke. Spectral images for classification of natural and artificial truquoise samples. In *International Conference OPTO*, pages 197–202, 2002.

[GW92] R. Gonzalez and R. Woods. *Digital Image Processing*. Addisson-Wesley, 1992.

[GY95] Laurence G. Grimm and Paul R. Yarnold. *Reading and understanding multivariate statistics*. American Psychological Association, 1995.

[HAS03a] John Havermans, Hadeel Abdul Aziz, and Hans Scholten. Non destructive detection of iron-gall inks by means of multispec-

tral imaging - part 2: Application on origional objects affected with iron-gall corrosion. *Restaurator*, 24:88–94, 2003.

[HAS03b] John Havermans, Hadeel Abdul Aziz, and Hans Scholten. Non destructive detection of iron gall inks by means of multispectral imaging part 1: Development of the detection system. *Restaurator*, 24:55–60, 2003.

[HBB+06] Joseph F. Hair, Bill Black, Barry Babin, Rolph E. Anderson, and Ronald L. Tatham. *Multivariate Data Analysis*, volume 6. Pearson Prentice Hall, 2006.

[HC94] J.C Harsanyi and C. Chang. Hyperspectral image classification and dimensionality reduction: anorthogonal subspace projection approach. *IEEE Transactions on Geoscience and Remote Sensing*, 32(4):779–785, 1994.

[HG08] Tian Han and David G. Goodenough. Noise reduction of hyperspectral remotely sensed imagery: A nonlinear dynamical system approach. In *IEEE International Geoscience & Remote Sensing Symposium*, 2008.

[Hoh98] Peter Hoheisel. *Die Gttinger Stadtschreiber bis zur Reformation: Einfluss, Sozialprofil, Amtsaufgaben*. Vandenhoeck and Ruprecht, 1998, 1998.

[Jai89] Anil K. Jain. *Fundamentals of Digital Image Processing*. Prentice Hall, 1989.

[Jol02] I. T. Jolliffe. *Principal component analysis: Springer series in statistics*, volume 2. Springer, 2002.

[JR07] Xiuping Jia and John A. Richards. *Hyperspectral data exploitation: theory and applications*, chapter Hyperspectral Data Representation, pages 205–225. Wiley-Interscience, 2007.

[KAP+08] Marvin E. Klein, Bernard J. Aalderink, Roberto Padoan, Gerrit de Bruin, and Ted A.G. Steemers. Quantitative hyperspectral reflectance imaging. *Sensors*, 8:5576–5618, 2008.

[KMN+02] Tapas Kanungo, David M. Mount, Nathan S. Netanyahu, Christine D. Piatko, Ruth Silverman, and Angela Y. Wu. An efficient k-means clustering algorithm: Analysis and implementation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7):881 – 892, 2002.

[Lak06] Joseph R. Lakowicz. *Principles of fluorescence spectroscopy*, chapter Introduction to Fluorescence, pages 1–26. Springer, 2006.

[Lin06a] Xiaofan Lin. Quality assurance in high volume document digitization: A survey. Technical report, Digital Printing and Imaging Laboratory HP Laboratories Palo Alto, February 1 2006.

[Lin06b] Brian S. Lindblom. *Scientific Examination of Questioned Documents*. CRC Press, 2 edition, 2006.

[lRB05] J. De la Rosa and F. J. Bautista. Optical properties of paper at 337.1nm. *Revista Mexicana de Fsica*, 51.1:110113, 2005.

[LZT06] Likforman-Sulem Laurence, Aderrazak Zahour, and Bruno Taconet. Text line segmentation of historical documents: a survey. *International Journal on Document Analysis and Recognition*, 9(2):123138, April 2006.

[MF08] Simone Marinai and Hiromichi Fujisawa, editors. *Machine Learning in Document Analysis and Recognition Volume 90 of Studies in Computational Intelligence*, chapter Introduction to Document Analysis and Recognition, pages 1–20. Springer, 2008.

[Pit00] Ioannis Pitas. *Digital image processing algorithms and applications*, chapter Digital Image Processing Fundamentals, pages 1–51. Wiley-IEEE, 2000.

[PSK+08] R. Padoan, Th.A.G. Steemers, M.E. Klein, B.J. Aalderink, and G de Bruin. Quantitative hyperspectral imaging of historical documents: Technique and application. In *ART Proceedings (2008)*, 2008.

[PSS+00] Lucas Parra, Clay Spence, Paul Sajada, Andreas Ziehe, and Klaus-Robery Muller. Unmixing hyperspectral data. *Advances in neural information processing systems*, 2000.

[Qua91] Abigail B. Quandt. *The Documentation and Treatment of a Late 13th Century Copy of Isidore of Sevilles Etymologies*, volume 10. Book and Paper Group (BPG) of the American Institute for Conservation of Historic and Artistic Works, 1991.

[Rehng] Malte Rehbein. Reconstruction the textual evolution of a medieval manuscript. Special Issue. LLC. The Journal of Digital Scholarship in the humanities., (forthcoming).

[Rie08] Oya Y. Rieger. Preservation in the age of large-scale digitization. Technical report, Council on Library and Information Resources No. 141, 2008.

[RJ06] John Alan Richards and Xiuping Jia. *Remote sensing digital image analysis: an introduction*, chapter Supervised Classification Techniques, pages 193–248. Birkhuser, 2006.

[SaAS04] J.H. Scholten and M.E. Klein andTh. A.G. Steemers. Hyperspectral imaging - a novel nondestructive analytical tool in paper and writing durability research. In *Durability of Paper and Writing*, 2004.

[SGS+97] Suresh Subramaniana, Nahum Gata, Michael Sheffield, Jacob Barhenb, and Nikzad Toomarianc. *Algorithms for Multi-*

spectral and Hyperspectral Imagery III, volume Volume 3071 of Proceedings of SPIE–the International Society for Optical Engineering. SPIE, 1997.

[Shl09] Jonathon Shlens. A tutorial on principal component analysis. Technical report, Center for Neural Science, New York University New York City, NY Systems and Neurobiology Laboratory, Salk Insitute for Biological Studies La Jolla, CA, 2009.

[Smi97] K.J Smith. *Evaluation of Whiteness and Yellowness in Color Physics for Industry*. Society of Dyers and Colourists, 1997.

[Smi99] Abby Smith. Why digitize? Technical report, Council on Library and Information Resources Washington, D.C., February 1999.

[SRC02] Jerry Silverman, Stanley R. Rotman, and Charlene E. Caefer. Segmentation of hyperspectral images from the histograms of principle components (proceedings paper). In Sylvia S. Shen, editor, *Imaging Spectrometry VIII (Proceedings Volume)*, 2002.

[SrH01] A.J. Sellen and r. Harper. The myth of the paperless office. MIT press, 2001.

[STB07] Emanuele Salerno, Anna Tonazzini, and Luigi Bedini. Digital image analysis to enhance underwritten text in the archimedes

palimpsest. *International journal on document analysis and recognition*, 9:79–87, 2007.

[TB92] F. Toselli and Johann Bodechtel. *Imaging spectroscopy: fundamentals and prospective applications*, chapter Imaging Spectroscopy for earth remote sensing, pages 1–19. Springer, 1992.

[TBS04] Tonazzini, Luigi Bedini, and Emanuele Salerno. Independent component analysis for document restoration. *International Journal on Document Analysis and Recognition*, 7:17–27, 2004.

[Til00] Richard Tilley. *Colour and optical properties of materials: an exploration of the relationship between light, the optical properties of materials and colour*. John Wiley & Sons, 2000.

[TSMB04a] A. Tonazzini, E. Salerno, M. Mochi, and L. Bedini. Blind source seperation techniques for detecting hidden texts and textures in document images. *International Conference on Image Analysis and Recognition*, Lecture Notes in Computer Science 3212:241–248, 2004.

[TSMB04b] Anna Tonazzini, Emanuele Salerno, Matteo Mochi, and Luigi Bedini. Blind source seperation techniques for detecting hidden texts and textures in document images. *International Conference on Image Analysis and Recognition*, Lecture Notes in Computer Science 3212:241–248, 2004.

[TSMB04c] Anna Tonazzini, Emanuele Salerno, Matteo Mochi, and Luigi Bedini. *Document Analysis Systems VI: Lecture Notes in Computer Science*, chapter Bleed-Through Removal from Degraded Documents Using a Color Decorrelation Method, pages 229–240. Document Analysis Systems VI, 2004.

[Tuc79] C.J. Tucker. Red and photographic infrared linear combinations for monitoring vegetation. *Remote Sensing Enviroment*, 8:127–150, 1979.

[uarya] unknown author. Lebor na huidre / the book of the dun cow. Vellum, 12th Century. R.I.A. MS 24.

[uaryb] unknown author. Liber flavus fergusiorum. Vellum, 15th Century. R.I.A. MS 23 O 48 a-b (2 volumes): Cat. No. 476.

[VA04] Pramod K. Varshney and Manoj K. Arora. *Advanced image processing techniques for remotely sensed hyperspectral data*, chapter Feature extraction from hyperspectral data using ICA, pages 199–215. Springer, 2004.

[Vaa99] J. Vaarasalo. Optical properties of paper in papermaking science and technology. *Pulp and Paper testing*, 17:162181, 1999.

[VJ06] S. Vaseghi and H. Jetelova. principal and independent component analysis in image processing. *Proceedings of the 14th ACM International Conference on Mobile Computing and Networking (MOBICOM '06)*, pages 1–5, 2006.

[VRR+07] Francesca Voltolini, Alessandro Rizzi, Fabio Remondino, Stefano Girardi, , and Lorenzo Gonzo. Integration of non-invasive techniques for documentation and preservation of complex architectures and artworks. In Sabry El-Hakim Fabio Remondino, editor, *Proceedings of the 2nd ISPRS International Workshop*. ETH Zurich, Switzerland, International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences Volume XXXVI-5/W4, July 2007. Proceedings of the 2nd ISPRS International Workshop.

[Wat75] Wilhelm Wattenbach. Das schriftwesen im mittelalter. Zweite, verm Auflage Leipzig: Hirzel, 1875.

[WMW98] G. A. WAGNIERES, W. M.STAR, and B. C. WILSON. In vivo fluorescence spectroscopy and imaging for oncological applications. *Photochemistry and photobiology*, 68:603–632, 1998.

[WN] Curator of Manuscripts William Noel. The archimedes palimpsest project. online.