Paper-based Watermark Extraction with Image Processing

by

Hazem Ali Abd Al Faleh Al Hiary

Submitted in accordance with the requirements for the degree of Doctor of Philosophy.

UNIVERSITY OF LEEDS

The University of Leeds
School of Computing

July 2008

The candidate confirms that the work submitted is his own and that the appropriate credit has been given where reference has been made to the work of others.

Abstract

This thesis presents frameworks for the digitisation, localisation, extraction and graphical representation of paper-based watermark designs embedded in paper texture. There is a growing need for this among librarians and antiquarians to aid with identification, wider accessibility, and providing a further level of document imaging for preservation. The proposed approaches are designed to handle manuscripts with interference such as recto and verso writing, and defects such as non-uniform paper structure, physical damage, etc.

A back-lighting scanning technique is used for capturing images of paper, followed by a selection of intelligent image processing operations, rather than alternatives such as radioactive techniques. This technique requires low cost equipment, and produces a fast and safe solution to capturing all details on paper, including watermarks, and laid and chain lines patterns.

Two approaches are presented: the first takes a bottom-up approach and deploys image processing operations to enhance, filter, and extract the watermark, and convert it into a graphical representation. These operations determine a suitable configuration of parameters to allow optimal content processing, in addition to the detection and extraction of chain lines. The second approach uses a model of the back-lighting effect to locate a watermark in pages of archaic documents. It removes recto information, and highlights remaining 'hidden' data, and then presents a statistical approach to locate watermarks from a known lexicon.

Work is further presented on reconstructing features of the paper mould by aggregating the success of the foregoing steps: this permits an analysis of 'twin' watermarks.

Results are presented from comprehensively scanned eighteenth and nineteenth century manuscripts, including two unusual copies of the Qur'ān, an Islamic Prayer, and various historical manuscripts.

Acknowledgements

First of all, I would like to begin by thanking God, for providing me with the health, strength and patience, guiding me through all the difficulties to start this work and complete it.

My deep thanks go to my parents, my brother Asem and sisters Sireen and Sawsan, they kept encouraging and supporting me for the whole study period. This thesis is my gift to my father, which I hope will make you happy and proud of your son.

Samah, my wife, I owe you my life, thank you for helping me finishing this research, you provided me with all what you can to make me happy, you took care of me in good and bad times, you never complained, my deepest appreciation for you. My warmest gratitude goes also to my baby son Aws. The most beautiful gift I ever got.

Most of all I would like to thank is my great supervisors, Professor Roger Boyle and Dr Kia Ng, you supported me until the end of my PhD, and provided all the help and extensive support I needed. It was an honour working with them.

I would also like to acknowledge the University of Jordan for their support. Additionally I would like to thank the Special Collections of the University of Leeds Brotherton Library for their cooperation in providing the manuscripts used in this research project.

For all my beloved people, thank you very much.

Declarations

Some parts of the work presented in this thesis have been published in the following articles:

Hazem Hiary and Kia Ng, "Optical imaging for watermark: digitisation, segmentation, and vectorisation", In *Proceedings of International Conference on Multimedia, Image Processing, and Computer Vision (IADAT-micv2005)*, pages 178-182, Madrid, Spain, March 30–April 1 2005. International Association for the Development of Advances in Technology (IADAT).

Hazem Hiary and Kia Ng, "Segmentation approach for paper-based watermark extraction", *IADAT Journal of Advanced Technology on Imaging and Graphics (IJATig)*, 1(2):62-65, December 2005.

Hazem Hiary and Kia Ng, "Watermark: From paper texture to digital media", In *Proceedings of 1st International Conference on Automating Production of Cross Media Content for Multi-channel Distribution conference (AXMEDIS 05)*, pages 261-264, Florence, Italy, November 30–December 2 2005. IEEE Computer Society Press.

Hazem Hiary and Kia Ng, "Automated paper-based watermark extraction and processing", In *Proceedings of 2nd International Conference on Automating Production of Cross Media Content for Multi-channel Distribution conference (AXMEDIS 06)*, pages 291-298, Leeds, UK, 13–15 December 2006. IEEE Computer Society Press.

Hazem Hiary and Kia Ng, "A system for segmenting and extracting paper-based watermark designs", *International Journal on Digital Libraries (IJDL)*, 6(4):351-361, July 2007.

Kia Ng and Hazem Hiary, "Digital acquisition and extraction of paper-based watermark designs with image processing", In *Translated Studies of Arabic manuscripts papers – Selections, under supervision of Anne Regourd*, Sana'a, Yemen, 2008. French Institute of Archaeology and Social Sciences, and German Institute of Archaeology.

Roger D Boyle and Hazem Hiary, "Watermark location via back-lighting and recto removal", Submitted to the *International Journal of Document Analysis and Recognition (IJDAR)*, July 2008.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

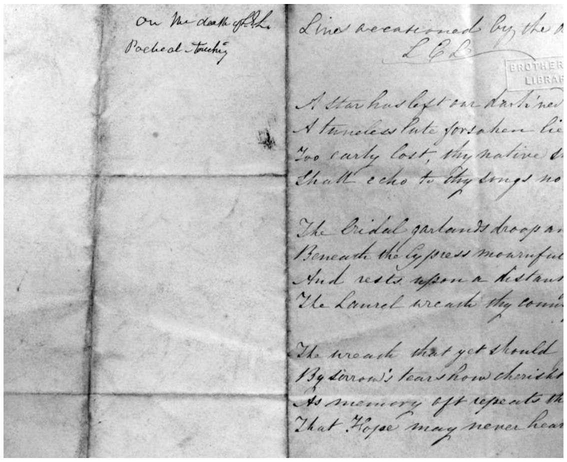
Watermarks in paper are enigmatic because they are hidden. They can also be beautiful, and informative. Seeking, identifying and cataloguing them has long been a human interest [21, 78, 97, 129].

The first known watermark was produced in 1282, originating in Fabriano [78]. These designs were mainly used as trademarks of the paper-makers, and later to trademark paper, a proof of the manufacture date, and an indication of paper size. Use has developed over the centuries and nowadays paper watermarks are used to identify paper owners and are also used for authentication to protect important documents such as bank notes, passports, and tickets from forgery and theft.

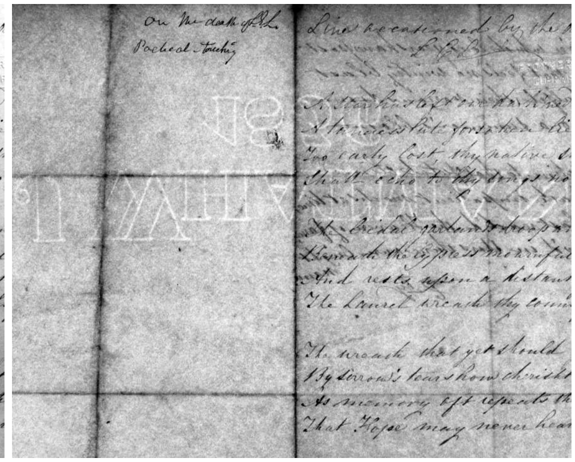
1.1 Research motivation

The motivation behind the study of watermarks is to assist in the tracing of old documents and artefacts to provide plausible historical relationships and background information, such as date and origin. However, there exist some complications for this study:

- Paper watermarks are, by design, hidden and may only be seen when the document is faced against light, for example.
- Many documents of interest are delicate or in private collections: it can be difficult for researchers to have access to watermark collections without permission.



(a)



(b)

Figure 1.1: Historical paper captured using back-lighting, (a) Reflected, (b) Transmitted. This document is taken from the works of Henry Litolff [14]. Digitised with permission from the Special Collections of the University of Leeds Brotherton Library [123].

- Watermarks are usually embedded on paper with writing on front (recto) and back (verso). In addition, there are often paper defects such as folding marks, paper texture, etc. These introduce interference that obstructs watermarks and make studying them difficult.

Many reproduction techniques have been developed to assist in these studies. These include manual tracing, radiographic techniques, and the use of cameras with back-light. This thesis uses back-lighting as it is simple, fast, and requires relatively low cost equipment to deliver fully digital output. Digital images can be compared, processed, stored and retrieved easily. Furthermore, this technique allows further image processing approaches to be applied easily on images. Captured images are of a high resolution, which allows the observer to see very small details of the image.

However, relying on reproduction techniques is not enough in most cases, because of noise and interference left on paper which obstructs the watermark design. To demonstrate this problem, Figure 1.1 illustrates captured images of a sheet, using the back-lighting acquisition technique. Figure 1.1(a) shows the sheet image with normal light (reflected), and Figure 1.1(b) shows the image using back-lighting (transmitted). The watermark ‘J WHATMAN 1836’ (flipped) is visible in the transmitted image. As is clear, recto and verso features, in addition to other paper defects, are all visible in the transmitted image and obstruct the watermark design.

Another example is shown in Figure 1.2, which illustrates a sample from a more difficult dataset, where the watermark design (lower part of a double-headed eagle) can be

(a) (b)

Figure 1.2: Historical paper captured using back-lighting, (a) Reflected, (b) Transmitted. This document is taken from the 'Mahdiyya' copy of the Qur'ān. Digitised with permission from the Special Collections of the University of Leeds Brotherton Library [123].

seen faintly at the right edge of Figure 1.2(b). The paper sheets of this dataset are thick, as is the writing stroke.

1.2 Thesis objectives

This thesis attempts a solution for the preceding complications. Paper watermarks are located and extracted using two different approaches: these were developed to cover a wide range of manuscripts of various characteristics, including paper thickness, watermark visibility, noise distribution (paper structure, background illumination, etc.), recto and verso inscription of varying thickness. This research project aims to:

- Prototype wider accessibility and distribution of artefacts of interest by establishing web-archives of manuscripts [76, 77], especially the 'hard-to-reach' data sources such as the library special collections.

- Digitise these artefacts to provide long term preservation and to combat paper decay issues. The digitisation process enables a further level of document imaging for a more complete preservation since many digitisation efforts have ignored these invaluable contents embedded in the paper. Storage space costs have been reduced to a level that permits large manuscripts to be digitised and stored without difficulty.

- Minimise, as much as possible, the interference that obstructs the watermark designs. This is an important feature since this project is targeted at processing manuscripts that have been written on. Most existing related work suffered from this interference that prevented capture of clear designs.

- Develop algorithms that permit effective approaches to automate parameter selection. Most other work lacks adaptive selection.

- Provide measures of chain lines (caused by the wires attached on the mould during paper production). Providing such information is helpful in studying and dating documents [127, 146].

- Enhance detail features of watermarks by computing the mean shape from a collection of watermarked documents that hold the same design. This is helpful in combining partial similar watermarks from different documents back to a complete design.

- Distinguish ‘identical’ from ‘twin’ watermarks. Watermarks are often twins because paper was often made with two pairs of moulds with similar but not necessarily identical watermark designs. This was to accelerate the process of paper-making. This distinction can be important for studying documents [126, 128].

- Provide scholars (especially those who do not have experience in using computer systems) with tools that can deal with patterns interactively to offer a simple and easy environment.

1.3 Thesis overview

The previous sections have given an introduction to the problems associated with studying paper watermarks, and highlighted the thrust of the work presented in this thesis: this is organised as follows:

Chapter 2: Literature review presents a coverage of background and literature surveys relevant to the research. It covers paper watermarks and their history, an introduction to the history of paper making and the stages of paper and watermark creation, including hand-made and machine-made paper-making. It also discusses the motivation behind the study of watermarks, and existing related work and trends in these studies. Finally it discusses the motivations for our research, and highlight its advantages compared to others work.

Chapter 3: Source material and Digitisation procedures provides a description of material used for prototyping. These data are principally manuscripts of the eighteenth and nineteenth centuries, held by the Special Collections at the Brotherton Library of the University of Leeds. We also present the digitisation setup used for image acquisition; this is equipped with hardware to permit the back-lighting technique. We then present a description of the characteristics and quality of paper and watermarks found in our datasets.

Chapter 4: A bottom-up approach demonstrates a framework for the extraction of paper watermarks with the back-lighting technique. It describes the use of digital image processing techniques to remove foreground and background interference, detect and extract chain lines, and extract watermark patterns. Results from various system stages are used to illustrate and explain the framework design and processing. This approach deals with data of the kind presented in Sections 3.1.1 and 3.1.2.

Chapter 5: Modelling back-lighting introduces an approach to removal of recto features, followed by highlighting of watermark patterns, and goes on to present a statistical approach to location of watermarks from a known lexicon. Adaptive parameter selection is also introduced. Results are presented from a comprehensively scanned eighteenth and nineteenth century editions of the Qu'rān and an Islamic Prayer. These data are presented in Sections 3.1.3, 3.1.4, and 3.1.5. This approach aggregates similar watermarks together to provide their accurate details. It also distinguishes 'twin' from 'identical' watermarks.

Chapter 6: Post processing presents further post-processing to the bottom-up approach. This includes vectorising bitmapped output images, and presenting applications of interactive image and vector editing functionalities to allow manual removal of defects and unavoidable noise on the paper. Further, this chapter introduces evaluation criteria for the extracted patterns.

Finally, Chapter 7 discusses the conclusions and contributions we reached in this research, and discusses the capabilities and possible improvements of the approaches we presented. It suggests future directions regarding this area of research. This Chapter is followed by Appendices of sample test data and output.

Chapter 2

Literature review

This Chapter presents background and literature surveys relevant to this research. It covers the beginning of paper watermarks, a brief history of paper-making and the stages of paper and watermark creation, a discussion on the motivation behind the study of watermarks, and existing related work and trends of this research area. Finally, this chapter also discusses the motivation for the research, and highlights its advantages compared to others' work.

2.1 Paper watermarks and their history

Paper watermarks are changes in paper thickness, and they are normally viewed by holding the paper against light. They are the designs that have been embedded in the paper during manufacture. A paper mould is a rectangle-shape frame made from wood, covered with a laid or wove wire surface, and used for making a sheet of paper [18]. The watermark is usually made by twisting wires into shapes that are sewn onto the mould [124]. The watermark area is always thinner than any other areas in paper.

The production of paper watermarks was initiated over 700 years ago by paper-makers in paper mills in Italy. The oldest known watermarked paper was produced in 1282, originating in Fabriano. It was discovered by Briquet, and first recorded in 1900 [22], and later in his 'Les filigranes' [21], no. 5410 (cf. II 316) (also featured in [82], p52). It is a Greek Cross with circles at the cross-point and cross-ends, as illustrated in Figure 2.1.

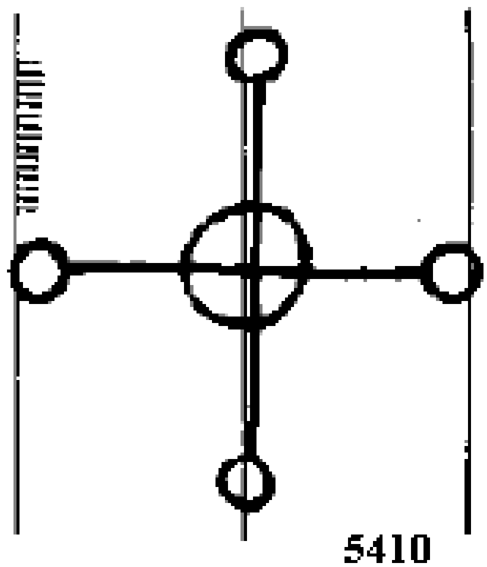


Figure 2.1: The earliest known watermark [82]

Hunter [78] discussed the theories for the usage of watermarks in the early days. These include using watermarks as trademarks of the paper-makers, or as an identification mark for sizes of moulds used for forming paper, or as symbols of religious groups called 'Albigenses' who used watermarks to identify the members of their group. Another theory suggested that these watermarks came from the imagination of paper-makers, just to show their artistic skills. A further theory for making watermarks was to help workmen who could not read to help them to identify the moulds to use.

Watermarks quickly spread through Italy and then over Europe, and the Arabic world, including the Maghreb in the 14th century [15]. Most paper was watermarked by the 15th century [124], but the term 'watermark' did not appear until the 18th century [78]. They are known as 'Wasserzeichen' in German, 'filigrane' in French, and 'papiermerken' in Dutch. By the 18th century, the usage of watermarks in Europe and America was to trademark paper, a proof of the manufacture date, and an indication of paper size. It was also used as a mark against counterfeiting on money and other formal documents [78].

Hunter [78] discussed the classification of watermarks from early days until the 18th century in four classes, based on their shapes. The first class includes the early watermarks, which have the forms of crosses, ovals, circles, knots, triangles, etc. The second class consists of shapes of the human figure, including a whole body, and human parts, such as head, feet, and hands. The third class consists of flowers, trees, leaves, vegetables, grain, plants, and fruits. Finally, the fourth class includes wild and legendary animals, such as unicorns and dragons, as well as snakes, fish, snails, turtles, crabs, scorpions, and varieties of insects. This class also includes bulls' heads, dogs, camels, elephants, leopards, goats, lambs, cats, horses, deer, and a large variety of birds. Examples of animal watermarks, with type, date used and description are in [16].

Hunter also mentioned the use of watermarks in bank notes. The first use of watermarks in Bank of England notes was in 1725. However, this did not prevent forgeries. The first case of forgery of watermarked bank notes of the Bank of England was recorded in 1758, followed by many other cases. Some cases were difficult to discover due to the accuracy of counterfeiters, which led to the invention of triple paper (coloured watermarks) in 1818 by Sir William Congreve, by forming and couching three sheets of paper as one sheet. However, this was rejected due to its production difficulty.

Another attempt to avoid forgeries in bank notes was the invention of light and shade watermarks, invented by William Henry Smith in 1848. This technique has the advantage of introducing any degree of density or lightness into paper watermarks. The first appearance of watermarks in stamps was in England in 1840 [78].

There are three main different types of paper watermarks:

1. Line (typically known as wire) watermark.

2. Shadow (light and shade) watermark.

3. Combined watermark, a combination of line and shadow watermarks in one paper sheet.

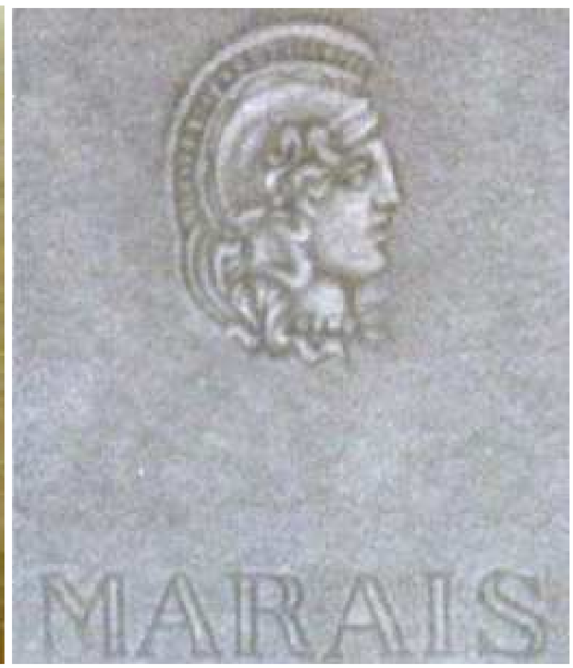
Further types of watermarks are given in [80, 92]. Figure 2.2 illustrates some examples of paper watermarks. Figure 2.2(a) illustrates an example of a wire watermark, Figures 2.2(b) and 2.2(c) show examples of light and shade watermarks, and Figure 2.2(d) illustrates a combined watermark.

Wire watermarks are made using lines to form various patterns, such as letters, numbers, portraits, or other designs. They appear lighter than surrounding paper areas. Light and shade watermarks have patterns resulted from relief sculptures on the mould, alternative names for this type are: chiaroscuro, tonal, shaded, shade-craft, and shadow watermarks [120]. These designs give the watermark further variations to support more features. They appear as dark and light areas when holding the paper against light. The advantage of using light and shade watermarks is to create more detail compared to wire watermarks. However, these watermarks are more expensive, depending on the size and the quality of the mould model [78]. Figure 2.3 illustrates an example of a shade and light watermark, and the mould used to produce it.

(a)

(b)

(c)



(d)

Figure 2.2: Examples of paper watermarks, (a) A European printing paper, (b) A Spanish Official Sealed paper, (c) Part of a bank note, (d) A European printing paper. With permission from Gabriel García [61]

(a) (b)

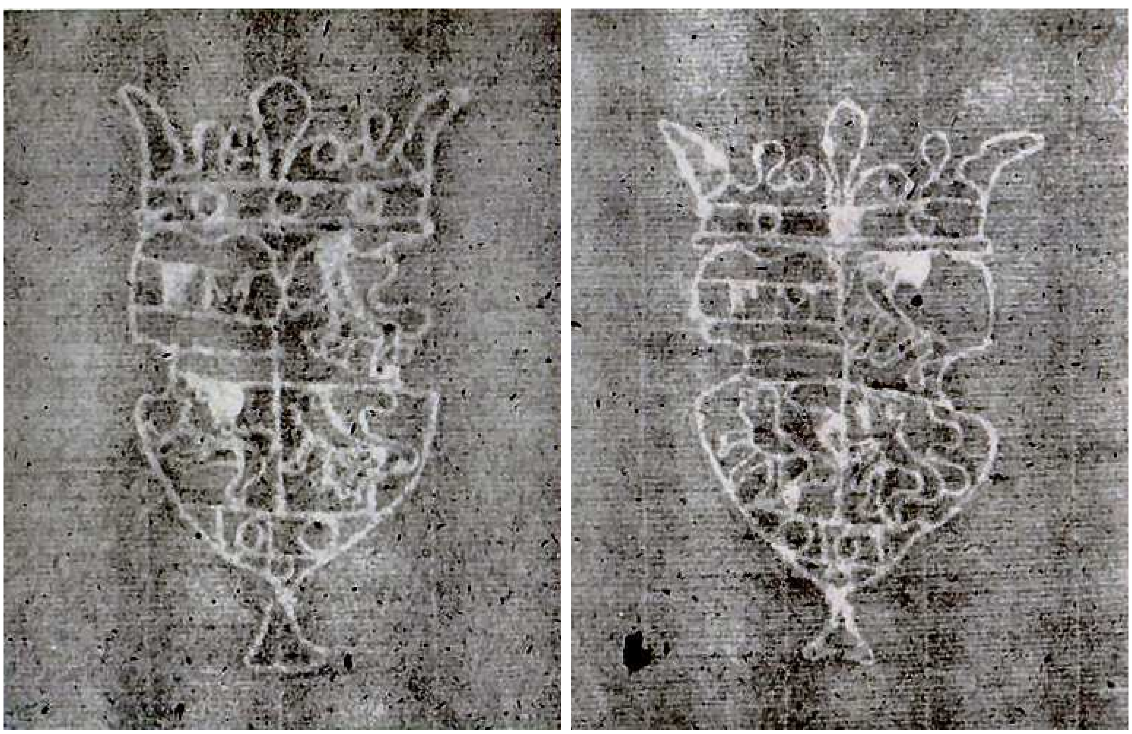
Figure 2.3: (a) Light and shade watermark, (b) Mould used to produce this watermark. With permission from Cindy Bowden [104]

Some paper-makers used to take popular watermarks from their original owners. This led to the introduction of the 'countermark' – an initial or symbol indicating the paper-maker's name, appearing opposite the main watermark on the other half of the mould and usually smaller than the watermark. This can be used to determine the paper-maker [92], and they are common after about 1650 [126]. Figure 2.4 shows an example of a counter-mark 'C L' which is found in a manuscript described in Section 3.1.5.

In many mills, paper was often made with two pairs of moulds with two very similar but not necessarily identical watermark designs. This was to accelerate the process of paper-making. Moulds were made in pairs from the early 17th century, which is why watermarks are generally twins. Also, double moulds, or divided moulds, appeared in the 18th century. They are used to make two sheets at once, and also result in twin watermarks [124, 126]. An example of twin watermarks can be found in Figure 2.5. One of the obvious changes in this example is the date: the year 1610 is written correctly in Figure 2.5(a), while the date is reversed in Figure 2.5(b). Paper and watermark production is detailed in Section 2.2.

Using watermarks as an anti-counterfeiting measure in bank notes and stamps was an inspiration for the use of watermarks in digital media, which also need to be secured

Figure 2.4: Example of a countermark ‘C L’



(a) (b)

Figure 2.5: Twin watermarks: Shield FM and Three Lions. With permission from David L. Vander Meulen [133]

from theft and forgeries. The term ‘digital watermark’ was first used by Komatsu and Tominaga in 1988 [33]. Early publications that focused on watermarking digital images include Tanaka *et al.* in 1990 and Tirkel *et al.* in 1993 [87]. Since then, the concept of watermarking has continued to evolve to identify, authenticate, and protect current digital materials such as digital images, audio, and video recordings [84]. This thesis considers only paper-based watermarks. Further reading on digital watermarking can be found in [5, 33, 75, 83, 105, 148].

Nowadays, paper watermarks are typically used to identify paper owners and for authentication to protect important documents such as bank notes, passports, and tickets from forgery and theft. Watermarks have also been used as a safeguard against espionage in many manufacturing plants, being embedded in identification cards for employees [78]. A discussion of the importance of watermarks and their study nowadays can be found in Section 2.3.

2.2 Paper and watermark making

Paper-making was invented in about A.D. 105 in China by T’sai Lun. The Arabs learnt the technique in 751 from Chinese prisoners in Samarkand after the battle of Talas: since then, paper-making moved from East through Shiraz in 790, Baghdad in 793, and Cairo in 900 to the West, in Fez in Maghreb in the 12th century [82]. The first appearance of paper-making in Europe was in Xativa (south of Valencia), Spain in 1151, and then in Fabriano, Italy in 1276. Paper-making first appeared in England two centuries later in 1495, and in Pennsylvania, America in 1690 [78, 124]. The following sections explain the procedures of hand- and machine- made paper, and indicates at which stage the watermark is embedded.

There are two principle types of paper, laid and wove.

- In laid paper, laid wires are placed horizontally along the mould, as mentioned in Section 2.1. The mould is a rectangle-shaped wooden frame, covered with a laid or woven wire surface, with a small spacing between wires, which are used to let water drain during paper formation. *Chain lines* are placed vertically along the mould. These wires are thicker, and the spacing between them is larger than between the laid wires – they are used to hold laid lines [18]. Figure 2.6 shows an example of a laid mould, and also shows a watermark ‘Fleur de Lys’ (lily), on a shield, crowned, and a ‘G J’ monogram: it also has a countermark ‘G JONES / 1809’ [17].

- The other type is wove paper, which first appeared in 1755. This paper is made

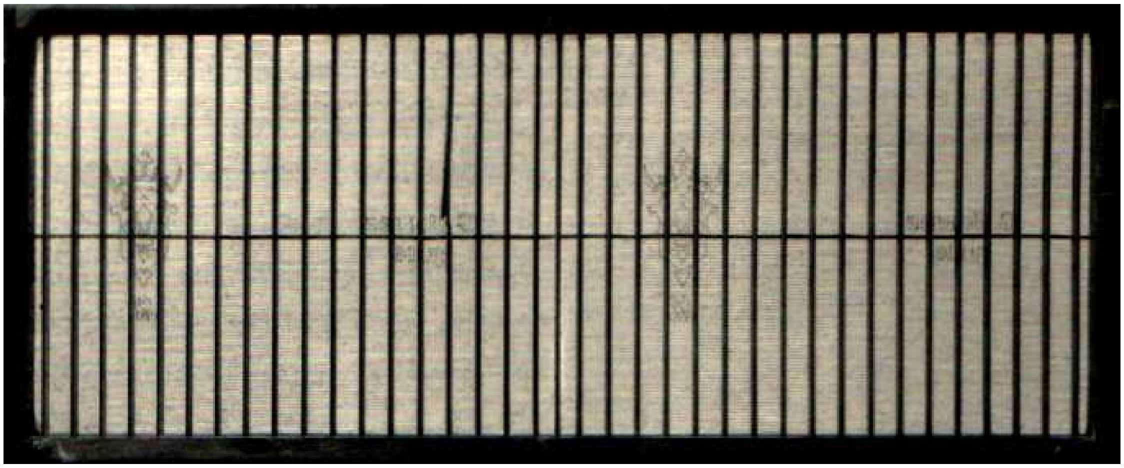


Figure 2.6: Laid mould [17]

using a mould with a finely woven wire mesh [78].

Both types have watermarks inserted as wires twisted into shapes and sewn on. Examples of wove and laid paper are shown in Appendix B in Pages 154 and 156 respectively.

2.2.1 Hand-made paper-making

Hand-made paper-making in paper mills has changed little from its early days until today. The stages of paper-making include preparing raw materials, beating, formation, drying, sizing, finishing, and quality control [17].

The raw material of paper is cellulose fibre derived from plants, or from old materials, such as old rags, ropes, sailcloth. Rags were sorted and checked if suitable, then cut into small squares, then boiled under pressure to soften them.

The next stage is beating. A Hollander beater with a heavy roll is used for beating rags. The quality, durability and characteristics of paper depend on the quality of rags and the way they were beaten. Large rag fibres are then broken using the ‘breaker’, which is a form of Hollander. Beating is used to separate individual fibres.

The next stage is paper formation, which is done in a vat room. Watermarks are embedded into paper in this stage. Experience is necessary to produce proper sheets, the ‘vatman’ forms the paper using a thread- (sieve-) like mould and deckle. A deckle is a removable frame around the mould. Moulds are used to make thin flat sheets.

Fibres are held in water (pulp), and the mould is dipped, shaken and pulled out – shaking will increase the sheet strength. The water then starts to drain through wires, and the paper pulp is left on the mould surface. The sheet thickness depends on the consistency of pulp in the vat, the deckle depth, and the vatman’s skill. The vatman then removes the deckle, and places it on the second (twin) mould, and starts to form another sheet. The first mould is then taken by the 'coucher', who takes the sheet off the mould, and puts it on a 'felt' (a wet woven blanket), and returns the mould to the vatman, who then puts another felt on top of the sheet, and takes the second mould, and so on, creating a stack of felts and sheets, called a 'post'.

The post is then pressed. This will make sure that more water will be removed, and will strengthen the paper sheets. Some mistakes may occur in the formation and couching processes, such as folded corners and edges, inconsistent thickness in sheets, etc. After post pressing, the 'layer' then separates paper from felts, and builds a 'pack' of wet paper. The paper is then taken for drying if a rough paper texture is required. If the paper texture required is smooth, then sheets are pressed.

The next stage is drying. Paper is hung on ropes to dry. This process can be lengthy, depending on the drying environment, such as temperature and humidity, and sheets' weight and size. The sheets are then placed in a cool place, so that air passes over the surface.

Sheets are then sorted. Bad sheets are returned for re-pulping, and the remaining sheets enter the sizing stage. Sheets are cut to a specific sheet size. After sizing, they are pressed to provide a good flat surface.

Finally, sheets are inspected for quality control. Actually, this stage was rarely done except by paper mills who cared about their name, and were famed for making fine and quality paper.

We see thus that watermarks are embedded into paper during the formation stage; also, we can see how paper types and qualities vary, and how faults may occur during paper-making [17].

2.2.2 Machine-made paper-making

The paper machine was first invented by Nicholas-Louis Robert in 1798, in Essons, near Paris [17]. He did this in order to make paper-making simple and cheap, and also because he did not like the restrictive practises and services of the paper-makers. Due to disagreements regarding money and rights between Robert and his paper mill boss, Leger Didot, development of the machine was prevented until John Gamble, a brother-in-law of Didot, moved the model to England, and took a patent in England in 1801. Henry and Sealy Fourdrinier bought a share in the new machine's right, and developed it. It soon became known as the Fourdrinier machine; the first working machine was produced in 1804 by Bryan Donkins, and since then, the paper machine continued to improve.

The Fourdrinier paper machine used to produce a continuous web of paper, until the invention of the cylinder mould machine in 1809 by John Dickinson, which changed the machine to produce single sheets rather than the continuous web. In order to simplify the drying process, drying cylinders were patented in 1820 by T.B. Crompton [18]. In its early stages, the paper machine was making paper without watermarks, until the invention of the dandy-roll. This is a roll covered with wire mesh, which has the watermark design as wires attached. It was invented by John Marshall (but not patented by him because there were no specifications recorded [78]) and patented by John and Christopher Phipps in 1825. This dandy-roll gave the look of laid and wove paper, and allowed the addition of both types of watermarks – wired and light and shade – to machine paper.

A brief description of a Fourdrinier paper machine (built from after 1820) is as follows: "it consists of a stuff chest containing pulp. The pulp is transferred to a vat before passing through a slice onto forming wire. The width of the sheet is controlled by the deckle straps. The wet sheet is transferred to an endless felt passing under a first press and a second press roll. The continuous wet sheet then passes round three heated drying cylinders before being reeled up dry on the reel" [18].

Watermarks are embedded after formation. Dandy-rolls are placed on the forming table, and press the formed paper sheets that pass under it. This gives a flexibility when the watermark position needs to be changed. A description of a paper machine and its functions is in [31].

This machine was an invention to cover the increasing demand for paper. The process is fast, simple and cheap. However, watermarks produced by paper machines lack the good contrast and shading found in hand-made processes [78].

2.3 Motivation for the study of paper watermarks: palaeographic issues

Watermarks in paper have attracted a wide range of interest from researchers for centuries. The motivation behind the study of watermarks is to trace old documents and artefacts to provide plausible historical relationships and background information. However, watermark designs are available not only in several different forms, but also dynamically change over time. This has introduced some complications that have hindered more systematic study of the artefacts. Sometimes, using watermarks to date or find the place of origin of documents is not accurate.

Not all watermarks hold dates (the oldest watermark that holds a date was in 1545 [82]), and we may not know for how long the same mould was used – maybe years. Further, there may not be any record of the time lag between paper production and its use. An example can be found when looking into the 'J WHATMAN' watermark. Its origin was from the Whatman mill, established by James Whatman in England, in 1731. Paper-makers took that watermark and used it for their own paper for many years [78]. A history and variation of this watermark is in [17].

On the other hand, watermarks can be used to correct errors in dating documents, especially if an identical watermark is found in definitely dated paper [66]. There are many examples for using watermarks as paper evidence. One example was the Shakespearean quartos published by Thomas Pavier: a false date of 1619 was given for all of them, but Sir Walter Greg proved in 1908 that those quartos were actually published at three different dates, 1600, 1608, and 1619 [124]. He determined that the watermarks in the quartos appeared in only these years, a discovery confirmed by Allan Stevenson [125]. Another example was the dating of the Missale speciale, which had an incorrect printing date. Stevenson found out that the Missale speciale was printed in 1473 by studying the watermarks in the Missale, and compared it with other identical watermarks from different books [124, 128, 129].

The size and orientation of the watermark can sometimes reveal some information about the size and quality of the original paper [66]. Knowing the original paper size can be helpful in determining paper usage, because paper of a specific size was used for specific uses [17].

Sometimes when studying watermarks, some slight differences can be observed between marks that are supposed to be the same. There are several possible reasons for this. Firstly, the watermarks may be twins, as discussed in Section 2.1. Two moulds may have been used in the same mill in order to accelerate the paper-making process, and it would be very difficult to make them identical. Secondly, it is possible that some watermark wires become detached, and imperfect repair may result in a different watermark design [78].

Twin watermarks are very helpful in dating documents. An interesting challenge for scholars nowadays is how to distinguish 'identical' from 'twin' watermarks. Stevenson proposed 10 differences, such as difference in sewing dots positions, chain line positions, and spacing regarding the watermark, countermark detail and position, etc [126]. He presented many examples of twin watermarks, and also highlighted the importance of sewing dots in the identification of twin watermarks, even if they are unclear [103, 126, 128]. Detailed criteria affecting identity when comparing identical watermarks can be found in [92]. Chapter 5 of this thesis considers possible approaches to locating these very subtle differences from images.

The study of chain and laid (also called 'wire') lines is also used to study and date paper, especially if there is no presence of watermark on paper [127, 146]. These lines are caused by the wires attached on the mould. Chain lines and their sewing marks can identify paper based on its variations and spacing (indentation) – these lines can be useful, with the presence of a watermark, to tell if it is identical or twin. However, these lines' positions may change gradually during the mould life [146]. Also, the spacing between them may change due to paper shrinking during drying process in paper-making [17].

The position of the watermark in various parts of the paper can also be related to its date, these position relations are detailed in [92].

The study and the investigation of the date and shape detail of paper watermarks was extended to detect forgeries of documents, wills, patents, bills, etc. Many examples of detecting forgeries in paper can be found in [78].

Due to the importance of paper watermarks, and in order to classify different paper materials, the International Association of Paper Historians [80] created a taxonomy of terms for describing the components of paper, including the watermark. Each watermark is assigned a code (e.g. E8 for snake), and these codes are arranged in tree structures, (e.g. Birds → Eagle → double-headed). The First International Conference On The History, Function And Study Of Watermarks discussed the importance of watermarks and their study, and was published in [97].

There are several published catalogues of watermarks, including 'Les filigranes' by Briquet, which contains over 16000 traced watermarks. He visited hundreds of paper mills in order to amass this collection [21]. Other collections can be found in [29, 66, 71, 106, 118], and a list of books of reproduced watermarks by tracing is in [81], together with a number of traced watermarks in each book.

Paper decays over time because of natural processes. To combat this, digitisation has been widely applied as one of the preservation approaches to keep a visual record of the artefacts, by creating a digital copy of the paper materials. Digitisation guidelines and best practises are available from many recent and current projects and institutes, such as Pulman [107], Minerva [95], AHDS [3], and MUSICNETWORK Imaging [99], more are in [30, 36, 38, 39, 138, 139]. However, most of these projects are only concerned with the paper surface, not watermarks or other paper 'internals', meaning that many watermarks may be lost forever when the sources decay.

Scholars require easy access to study different watermark collections. This requirement has led to the establishment of a number of web-based archives of watermarks to assist wider accessibility. Examples include [4, 42, 48, 52, 69, 88, 98, 153, 156] (these databases are mentioned in Section 2.4.1). A list of web databases is compiled in [13]. These archives can also help in preserving the watermarks from paper decay.

Gants [58] studied historical manuscripts written in the early seventeenth century, including the *Workes* of Beniamin Jonson, and built a digital catalogue of watermarks used by William Stansby in the printing of the *Workes* of Beniamin Jonson (London, 1616) [52].

He was also involved in several other digitisation projects, such as “The Cambridge edition of the works of Beniamin Jonson” [25, 56]. The aim of this project was to provide all the works of Beniamin Jonson in electronic form. Another project was “The early English booktrade database” [53, 60], this project aimed to provide a quantitative analysis of English materials printed and published in the period 1475-1640. In these projects, he studied textual materials, watermarks, and chain line spacing. More description of his approaches is in Section 2.4.2.

LIMA (Literary Manuscript Analysis) [70] is a website for the study of manuscripts, including handwriting, paper and watermarks. Another website is at the American Museum of Paper-making, of the Institute of Paper Science and Technology (IPST) [104], which provides information about watermarks and lessons on how to make them. Another website which provides rich information about paper watermarks and their history can be found in [120].

2.4 Watermark reproduction techniques and existing related works

As discussed in Section 2.3, scholars study watermarks in paper, together with countermarks, sewing dots, laid and chain lines, to pinpoint date and origin. However, paper watermarks can only be seen when faced against light, and also most watermarks are usually obstructed by writing ink and other noise in paper. To solve these problems, many approaches have been developed in order to reproduce watermarks. These include hand-drawn tracing, rubbing, photosensitive paper (Dylux), Ilkley, phosphorescence watermark imaging, transmitted light photographs (back-lighting), beta-, electron- and soft X-radiography, and thermography. Back-lighting is more of an acquisition (capturing) rather than reproduction technique. The following section gives a description of each technique with examples, followed by existing related works.

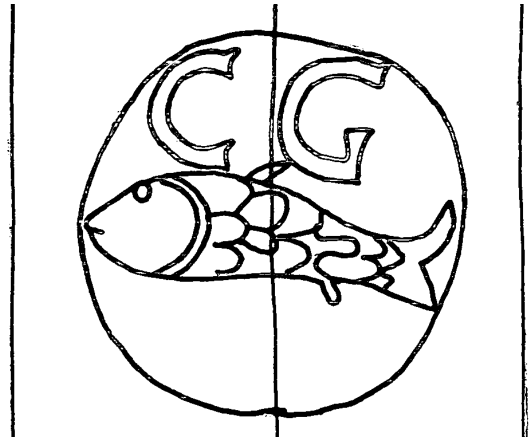


Figure 2.7: Watermark: Fish inside a circle, no. 44342. With permission from 'Vorlage: Hauptstaatsarchiv Stuttgart, J 340' [69]

2.4.1 Techniques of watermark reproduction

Manual tracing: Hand-drawn manual tracing of the watermark pattern requires a light table (back-light), blank paper and a pencil. This technique is simple and easy. However, it is a time consuming and highly subjective task. It is hard to trace watermarks obstructed by interference [6, 7] and thick paper [46], also, tracing may cause some damage to the paper [66]. Well-known catalogues of traced watermarks include [21, 29, 71]. Web archives of traced watermark images can be found in [69, 98]. Figure 2.7 shows an example of a traced watermark: a fish in a circle, with 'C G' letters.

Rubbing: The rubbing technique works by placing a clean sheet over the watermarked paper and diagonal strokes with a pencil are made with its unsharpened end from the paper upper left to lower right [80]. Rubbing is quick, easy and does not require special equipment, but it does not produce good results, and may damage the paper [6]. Many examples of watermark reproduction by rubbing can be found in [68], and web-based archives of watermarks reproduced by this method can be found in [88, 153]. Figure 2.8 shows an example of an anchor watermark reproduced by rubbing.

Dylux: The photosensitive paper 'Dylux' method was developed by Thomas Garvell [65]. It requires DuPont Dylux 503-1B yellow coated paper [41], a visible (fluorescent) light, an ultraviolet light, and a frame of two glass plates. The frame is used to make sure no shifting occurs during the process between the Dylux and original watermarked papers. Dylux 503-1B paper is used because it behaves in two different

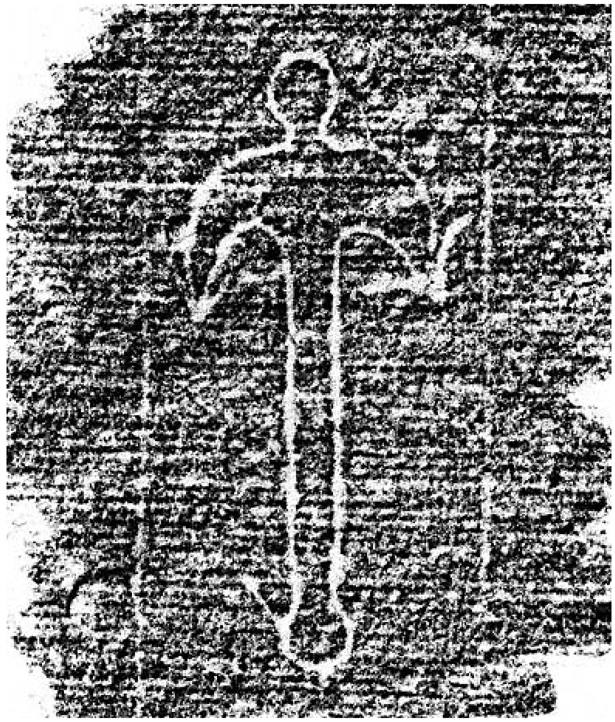


Figure 2.8: Watermark: Anchor, no. WM I 52712. With permission from Marieke van Delft [88]

ways to visible (400-500 nm range) and ultraviolet (200-400 nm range) light [54]. Since the watermark area in paper is thinner than other areas, the visible light will colour the whole paper in white, while the ultraviolet light will colour paper areas other than the watermark area in blue. This is helpful in separating the watermark from background.

This method works by placing the Dylux paper in the frame with the original watermarked paper laid over it, and the frame is then closed. The frame goes under the visible light source, three to four inches from the paper, and the yellow coated paper then becomes white. The second step is imaging or printing: the Dylux paper is taken from the frame, and held under the ultraviolet light source, at a distance of one foot, until the blue colour is formed. The result image consists of a blue background with white watermark [35, 64].

The advantages of this method include the relatively low cost equipment, time-saving, and production of watermarks without dark room conditions. However, this method also captures any design that interferes with the watermark, and its effectiveness depends on the paper thickness, ink opacity and light source types [117]. Also, exposure to both visible and ultraviolet light is time-limited. Any delay or

Figure 2.9: Watermark: Flower, no. FLR.005.1. With permission from Daniel W. Mosser [98]

move too soon will result in low contrast between blue and white colours, and this will affect the result.

The use of this method is not permitted in many libraries and museums because of the use of ultraviolet light [64]. The DuPont corporation [41] stated in their MSDS (Material Safety Data Sheet) no. DU002873 that the chemicals used in Dylux proofing papers release gases, so users should be cautious and use a well ventilated environment. A catalogue of watermarks reproduced by the Dylux method can be found in [66]. Web archives of watermark images reproduced by this technique can be found in [4, 52, 98]. Figure 2.9 shows an example of using this technique.

Ilkley: Ilkley is another method for watermark reproduction. It was developed by Robin Alston in 1976. It requires two glass plates, a light source (desk lamp) with photographic timer connected to it, and a Kodak Precision Line film LPD4. It works by placing the film over the glass plate, the watermarked paper is laid over the film, and the other glass plate is placed above. After that, it is exposed to light for 5 seconds (using the timer), the film is then removed and processed manually to reveal the watermark design. This method is simple and quick, and the film produced can be duplicated quickly and easily [117]. However, this method requires dark room conditions for exposure, and will capture any details in the paper in addition to the watermark. Hence, it is only useful for reproducing watermarks in clean paper without interference. Figure 2.10 illustrates a watermark image captured using this technique.

Phosphoresence: The phosphorescence watermark imaging reproduction technique re-

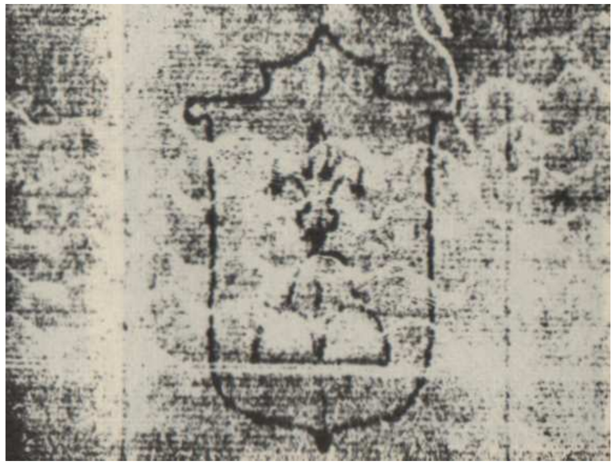


Figure 2.10: Watermark: Fleur de Lys on a shield. With permission from David Schoonover [117]

quires an ultraviolet and infrared light, a phosphorescent pigment plate, a glass plate, and a photographic film (e.g. Agfa HTP-3 blue-sensitive line film). These lights are used because the infrared waves go through the whole watermarked paper, causing the phosphorescent pigment plate to be dark, while the ultraviolet waves cause the plate to glow only in the locations of the watermark and laid and chain lines (thin areas in paper).

The plate is first excited by an ultraviolet light for 10 seconds at a distance of 10 cm, which makes the plate glow. Then, the watermarked paper is placed above the pigment plate, and the glass plate is laid over it. It is then exposed to the infrared and ultraviolet lights simultaneously for 20 seconds at a distance of 30 cm. The lights are then turned off, and the pigment plate is removed and placed immediately beneath the photographic film to make an image of the watermark [119]. This method is quick. However, the image quality depends on the distance between pigment plate and light sources, and also on the paper thickness and ink opacity. This method also captures image interference in addition to the watermark design. Figure 2.11 illustrates an example of a reproduced watermark (Fleur de Lys in a circle) using this technique.

Back-lighting: This acquisition method requires a high resolution digital CCD (Charge Coupled Device) camera and a light source (a thin foil of light with even homogeneous illumination behind the paper, used to visualise the watermark pattern) or light box. This technique uses the camera to capture reflected (with normal light) and transmitted (with back-light from slim light or light box) images of the water-

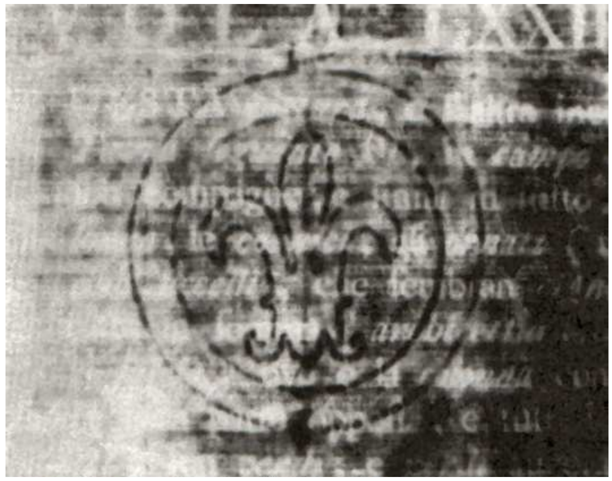


Figure 2.11: Watermark: Fleur de Lys in a circle. With permission from Carol Ann Small [119]

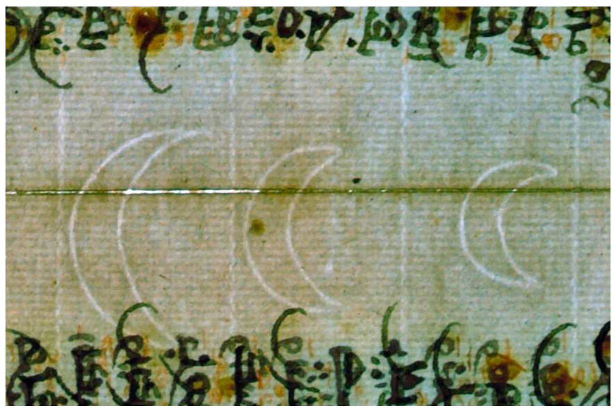


Figure 2.12: Watermark: Tre lune (three crescents or moons)

marked paper [6, 13, 27, 130, 145, 149].

This method is quick and produces good image quality, it requires relatively low cost equipment, and it does not require darkroom conditions. It differs from the earlier techniques in that it is digital. This is very helpful when further processing to images is required. This method made the study and investigation of paper watermarks easier for individual scholars [145]. However, it captures all the details of paper, including the watermark and any other designs that may interfere with it. Web archives of watermark images reproduced by back-lighting are in [42, 48]. Figure 2.12 shows a tre lune (three crescents or moons) watermark image obtained using this technique, taken from data described in Section 3.1.5: further examples are in Appendix B.

Thermography: Thermography, or thermal photography, is a reproduction technique de-

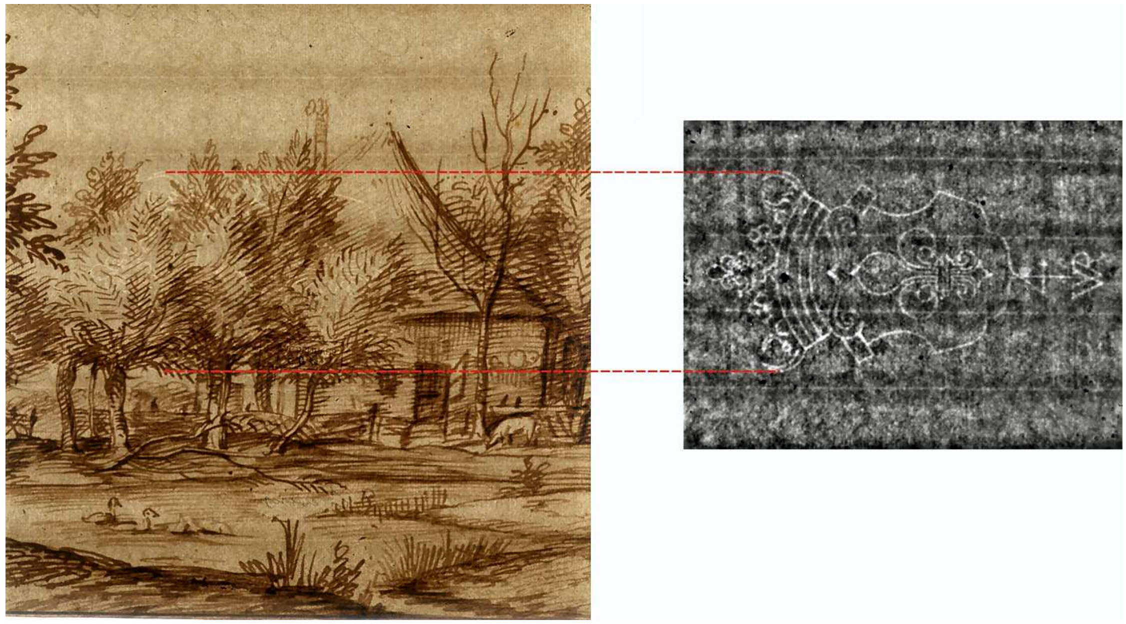


Figure 2.13: Watermark: Fleur de Lys on a shield, crowned [93]. With permission from Peter Meinlschmidt [50]

veloped at the Fraunhofer Institute by Neuheuser *et al.* in 2005 [93, 100]. They benefited from the fact that writing ink on paper is transparent (not absorbed) under thermal radiation (infrared light). This technique works by placing a thermal source (warm plate) at a temperature of 35 to 40 ^{o}C behind the watermarked paper, and using an infrared camera in front of it. The camera is sensitive to thermal radiation; it records the changes of the watermark density in paper, and generates a digital watermark image. This method is fast, and produces good watermark images. The limitation is concerned with the safety of the watermarked paper: it is safe as long as it is at a distance (of 1 cm) from the warm plate, and the exposed time is only one second [93]. A result of using this technique is illustrated in Figure 2.13, the original Rembrandt drawings are from the Herzog Anton Ulrich-Museum [74], and thermographic images from Fraunhofer-Institute for Wood Research – Wilhelm-Klauditz-Institut (WKI) [50].

Radiographic techniques: There are three radiographic techniques for watermark reproduction: Beta-, soft X- (low voltage) and electron-radiography. Their advantage comes because of the ability to display changes of paper thickness, no matter what is printed on it [145]. The reason behind using X-rays in recording watermarks was because they are not absorbed by writing ink (usually Carbon) on paper [140].

1. The Beta-radiography method was developed in the late 1950s by D P Erastov,

Figure 2.14: Watermark: Fleur de Lys, no. AT5000-553_257. With permission from Alois Haidinger [156]

from the Academy of Sciences at Leningrad. It uses beta-isotopes (Carbon-14) to record variations in paper thickness (watermark, countermark, chain and laid lines, and sewing dots) on an X-ray film [117]. The watermarked paper is placed between the beta-isotope plate and the X-ray film. Beta rays are radiated from the plate, go through the paper and expose the film. A detailed description of this method can be found in [6, 117].

Beta-radiography gives an accurate image of the watermark with minimum interference, and films produced can be duplicated easily, but unfortunately is time consuming (two to twenty four hours per page [119, 137]) and expensive (approximately $2500 per plate [119]). For this reason, only large institutes and museums use it [145], and it requires darkroom conditions [117].

There are also some concerns regarding radiation safety [119]. Results of watermark images of radiographic techniques may be blurred depending on the paper thickness [112], and the imperfect contact of the watermarked paper, the beta-isotope plate and the X-ray film [34]. A web archive of watermark images reproduced by this technique can be found in [156]. Figure 2.14 shows a reproduced Fleur de Lys watermark using this technique.

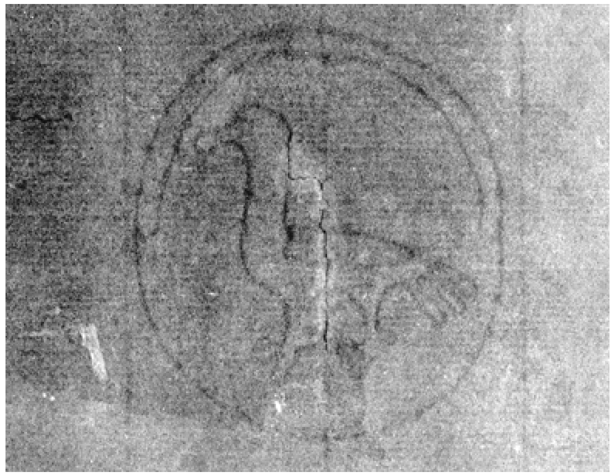


Figure 2.15: Watermark: Bird in a circle, no. IT-CBF-46 A. With permission from Georg Dietz [42]

2. Soft (or low voltage) X-radiography was described by Bridgman [19], and further developed and improved by dentists Van Hugten [89, 142] and Van Aken [140]. A low voltage energy (5keV-10keV: kilo electron volts) is radiated from the X-ray source through the paper to a phosphor plate – exposure takes 2 minutes. The phosphor plate is then read by a laser reader (originally used for dentistry), and the watermark image takes 4 minutes to be generated digitally [145]. The reason for using low voltage radiation, which produces very long wavelengths, is because it gives high contrast (sharp) images.

   This method gives very good watermark images. Moreover, it is cheaper, faster (requiring 5-30 minutes [137]) and relatively safer (as long as 10 keV voltage is not exceeded) than beta-radiography. Van Hugten used modified dental X-ray equipment in order to make the setup portable for mobile use, and Van Aken improved the contrast in results, and allowed non-darkroom conditions, but this technique is still expensive. A detailed description can be found in [137, 140]. A web archive of watermark images reproduced by soft X-radiography is in [42]. A watermark image reproduced using this technique is in Figure 2.15.

3. Electron-radiography was described by Bridgman [19, 20], and further developed by Schnitger *et al.* at Deutsche Staatsbibliothek and Technische Universität in Berlin [115, 116, 158]. With this method, X-rays of high energy are pointed to a lead sheet to emit electrons, and these electrons go through the watermarked paper to a photographic film, as in beta-radiography. The film will hold an image of the watermark with minimum interference.

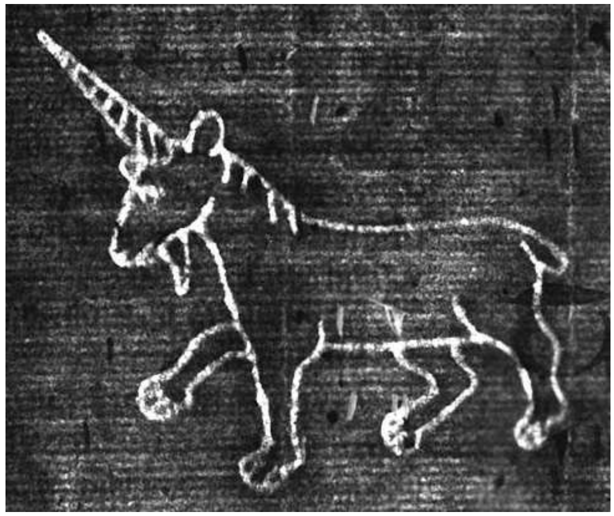


Figure 2.16: Watermark: Unicorn, horizontal to left, no. WM I 00063. With permission from Marieke van Delft [88]

> This technique produces very good watermark images and is faster than other radiographic techniques (requiring 1 second [137]), and does not require dark-room conditions. It has the advantage over other radiographic methods that in the case the writing ink was metallic, X-rays will be absorbed by this ink and will appear in the final image, while electrons will not [19]. However, it is very expensive, and requires safe (radiation shield) conditions. A web archive of watermark images reproduced by this technique is in [88]. Figure 2.16 shows a result of a reproduced unicorn watermark, using electron-radiography.

Among these techniques, radiographic techniques give the best result of watermark images, as these results do not suffer from interference caused by writing ink and other obstacles: beta- and electron- radiography need to be scanned for digital processing and archival, soft X-radiography gives the highest resolution, and produces sharper images compared to other radiographic techniques. It also records the entire paper sheet in a single exposure [7], and needs short exposure time. Electron-radiography is the fastest method among radiographic techniques (not faster than transmitted light). Back-lighting method is considered the best among non-radiographic methods, advantages of using back-lighting is discussed in Section 2.5. A full comparison of radiographic and back-lighting techniques, together with requirements and description is in [137]. However, these radiographic techniques are still expensive, especially for individual scholars, needing specialised equipment, and limited to small formats of paper, depending on the size of the X-ray films and plates [12, 145]. It is also unsafe due to radiation hazards.

2.4.2 Existing related work

There is much literature on the location and extraction of watermark designs after being reproduced. Most of these works were to build watermark databases. Depending on reproduction techniques is not enough to study watermarks in most cases, because of noise and interference left on paper which obstructs the watermark design, and because radiographic technique are only in the hands of large institutes, not individual scholars.

The advantage of using digital, rather than non-digital, techniques is because they can observe information in images at scales that may be too small or too large for the human eye. Digital images can be compared, processed, stored and retrieved easily [54]. This Section discusses related work, together with its advantages and disadvantages.

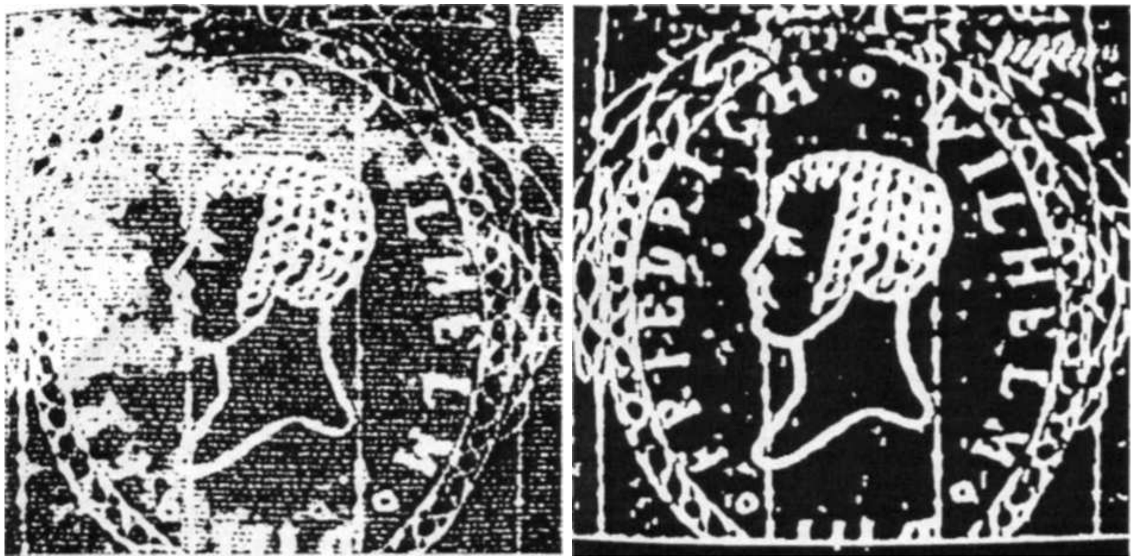
Combining back-lighting digitisation with various image processing operations offers an effective and simple to use technique for extracting the watermark design from paper. The motivation for using such operations is to isolate and remove noise and other interference, including writing ink, uneven background illumination, and the existing damage on paper [157].

Digital image processing is the science of manipulating digital images. These processes include noise reduction, contrast enhancement, image sharpening, filtering, segmentation, objects recognition, morphological operations, edge detection, image analysis, etc. The purpose of using such processes includes improving the image visual appearance to human eye, such as noise reduction, and preparing images for non-interactive processing such as feature analysis and measurement, such as edge detection [113].

The most commonly used processes in this review of related work is mathematical morphology. This is a combination of an image and a *structuring element* using a set operator (e.g., union, intersection, difference, etc). The structuring element is a shape that may be square, disc, line, diamond, etc. In all morphological operations (e.g. dilation, erosion, opening, closing, reconstruction, etc), image data are processed and modified depending on the structuring element. These operations simplify the image features, preserve its shape characteristics, and can remove irrelevancies [63, 67, 91, 122]. The morphological top-hat transform is also widely used in this research area to remove non-uniform image background, defined as $TopHat = A - (A \circ S)$, where $\circ$ is morphological opening, and A and S are the image and the structuring element respectively [63].

Edge detection is an operation for feature detection and extraction in images that identifies image edges: places in an image that correspond to features boundaries. Edge detection methods include Sobel, Prewitt, Roberts, Laplacian of Gaussian, and Canny [63,122].

Other operations include enhancing images using histograms [63]. Adjusting image contrast and brightness is an example of using histograms in image enhancement. Image



(a) (b)

Figure 2.17: (a) Input watermarked image, (b) Output binary image. With permission from Volker Märgner [132]

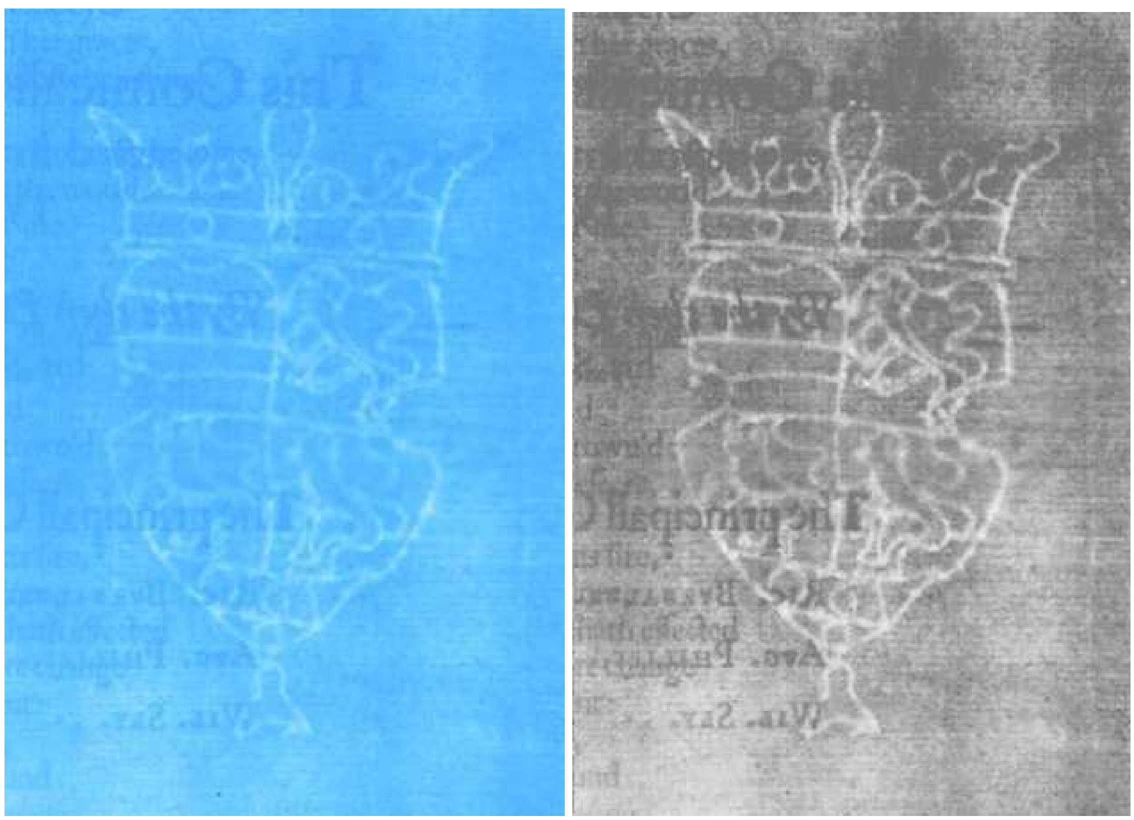
subtraction is also considered in this Section, defined as the difference between two images A and B, denoted as $D(x,y) = A(x,y) - B(x,y)$, where x and y are the coordinates of pixels pairs in images A and B.

Zamperoni [157] proposed a watermark database system in which it is possible to perform watermark image retrieval. He used back-lighting and image processing in order to extract watermarks, using only the transmitted (backlit) image. First, he removed chain lines using morphological closing or frequency filtering to give an image A. Then he used the top-hat transform to approximate the background, and subtracted it from image A, followed by contrast enhancement, to give B. Then, he separated the process into two steps: the first one takes image B and cleans it (removal of noise, which also results in removal of part of the watermark), then dilation is applied to smooth the resulting binary image, to give B_1. The other step enhances B, in which the watermark signal becomes stronger, but interfered with noise, to give B_2. B_1 and B_2 are then grouped together by the AND operator. The result is finally filtered by a median filter.

The resulting watermark is binary; this is an advantage because data size is reduced, and so searching a database for watermarks will be easier and faster. In this case, the watermark pattern can be converted to a contour easily, in other words, the watermark patterns can be presented by a sequence of numbers (contour coding [63]). This coding will provide further data size reduction. However, results of this system suffered from interference. Figure 2.17 shows an input transmitted image and output binary result.

Gants [54] studied watermarks found in the *Workes* of Beniamin Jonson (London, 1616). He applied image processing techniques to enhance and reduce interference in images reproduced using the Dylux and beta-radiography techniques [57]. He first scanned these reproductions, and converted images to grey-scale, and then shifted the contrast and brightness to make the watermark, together with laid and chain lines, look clearer. Then, he analysed the histogram to select narrow bands of grey shades areas, and shifted pixel values of these areas to the values of surrounding areas, so it fades into the background.

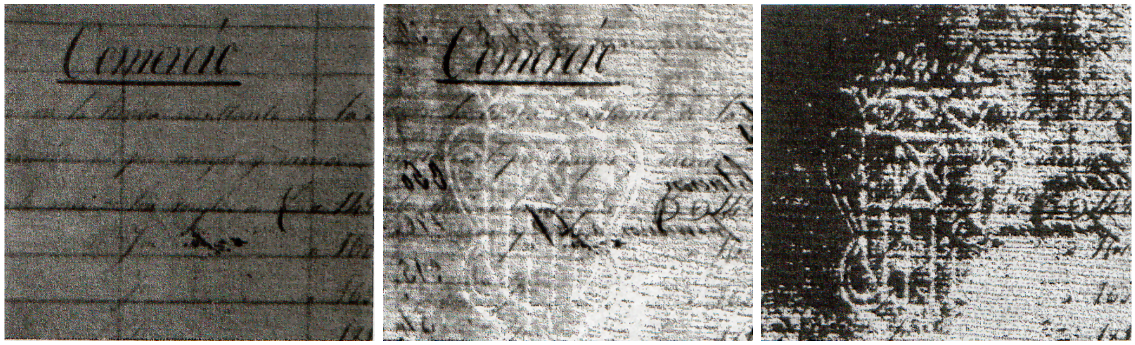
He also used the above enhancements to study watermarks reproduced using back-lighting [55], and studied and identified papers by measuring the spacing between chain lines [59]. However, results after enhancements still suffer from interference. Figure 2.18 shows an example of Gants' work.



(a) (b)

Figure 2.18: (a) Watermarked image (from Beniamin Jonson's *Workes* of 1616), reproduced with Dylux method, (b) Output result after enhancement. With permission from David Gants [54]

Stewart *et al.* [130] also used back-lighting with image processing; they presented two techniques, image segmentation, and modelling ink and paper optics. They discussed the use of histogram thresholding in extracting watermarks. A trial and error process was



(a) (b) (c)

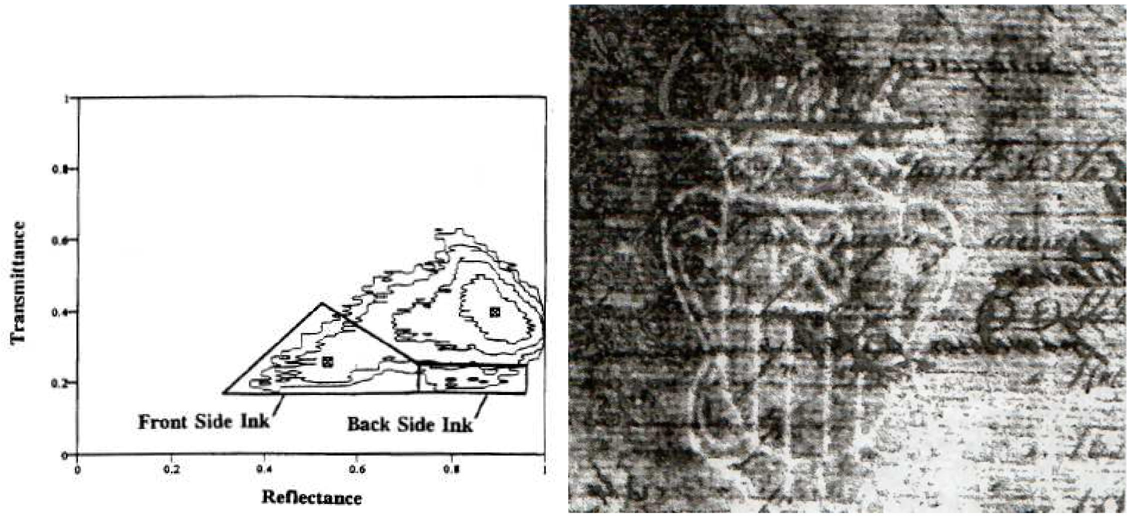
Figure 2.19: (a) Input reflected image, (b) Input transmitted image, (c) Output result after thresholding. With permission from Jonathan S. Arney [130]

used in order to pick a threshold to separate ink from watermark in grey-scale images, and values of image pixels less than the threshold are changed to the value of boundary of these pixels, however this technique was not good since it resulted in losing part of the watermark. See Figure 2.19 for input images (reflected and transmitted), together with a result of histogram thresholding of the reflected image.

To solve this problem, they used both histograms of reflected and transmitted images, and built a 2-D histogram, and again used trial and error to perform thresholding. They managed to separate recto from verso ink on the paper, and changed the pixel values of these regions to the mean of the whole image. However, the result suffered from interference caused by ink, which was not removed completely. Figure 2.20 illustrates the 2-D histogram (in low resolution due to source) and the output result.

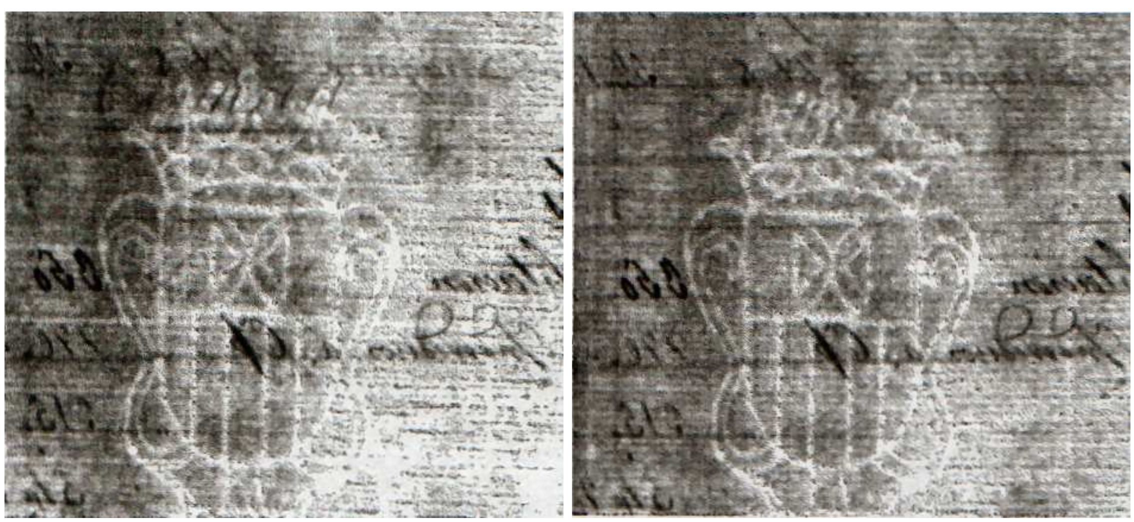
The next method aimed to separate the transmittance of the watermark from the optical density of ink, using the Beer-Lambert and Kubelka-Munk models of light absorption. These models can approximate the behaviour of ink on paper. However, they ignored the verso writing ink, and these models did not remove the recto ink completely, which resulted in interference in the output image. Results of using these models are in Figure 2.21.

Rauber *et al.* [109, 110] proposed a system for the management, archival and retrieval of historical papers which contains watermarks in a database that can be accessed via the Internet. To help scholars determine date and origin of unknown paper, it will be efficient if they compare such unknown watermarked paper with known watermarks in the database: this database contained an image and textual description of each watermark. They used back-lighting, followed by specific image processing algorithms [108] such as contrast and contour enhancement to remove laid and chain lines and other spots from papers. They also added scanned images of watermarks traced by hand by Briquet [21] to



(a) (b)

Figure 2.20: (a) 2-D histogram of reflected and transmitted images, (b) Output result after 2-D thresholding. With permission from Jonathan S. Arney [130]



(a) (b)

Figure 2.21: (a) Output result using Beer-Lambert model, (b) Output result using Kubelka-Munk model.With permission from Jonathan S. Arney [130]

their database. They proposed textual and image retrieval classifications of watermarks:

1. The class of the watermark, as presented by Briquet [21].

2. Using the IPH code presented by the International Association of Paper Historians [80].

3. Retrieval by specifying global features, using 12 features (e.g., watermark size, watermark position on paper, spacing between two sequential chain lines, etc).

4. Retrieval by comparing similar images. A similarity task processing algorithm is presented to compare the shape of a given watermark with other watermarks stored in the database: two algorithms were proposed for comparing similarities, Circular histogram and Directional algorithms, details of these algorithms are in [109].

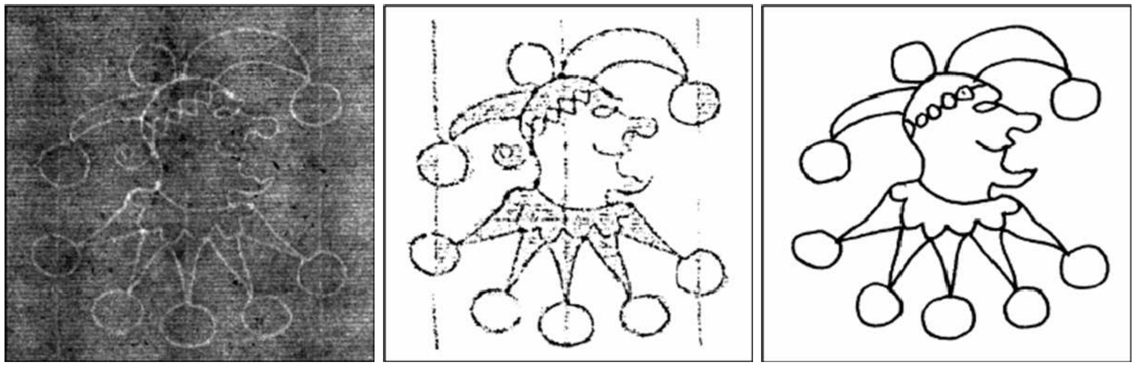
5. Retrieval by drawing an approximate shape: they built a feature which allows historians to draw watermarks manually, in order to be compared using image similarity.

6. Retrieval using small patterns, that is, retrieval using only part of the watermark, where watermarks in the database are indexed into a hash table, and convolution is applied to search for similar watermarks.

Rauber *et al.* also proposed a secure mechanism for copyright protection of material in the database by using digital watermarking. The main drawback of their approach is that they ignored paper with interference and concentrated more on clean paper and the traced scans. The image processing algorithms they used for removing laid and chain lines and other spots are semi-automatic, they did not discuss the selection of parameter values in these processes [108], and it is not clear how they judged retrieval success [111]. An example of their work is shown in Figure 2.22.

Ash *et al.* [7, 8] presented a database project using beta-radiography to reproduce watermarks in Rembrandt's prints – the aim of this project was to help Rembrandt scholars in their research by offering them accessibility and helping them to date his prints. For each watermark, they added information on the watermark description, with laid and chain lines, the date of the document, and a list of other prints which has the same (identical) and possible twin (nearly identical) watermarks.

Moschini [96] used back-lighting and image processing to build a database of watermarks. Some image processing methods were used to enhance and highlight watermarks in images (these processes were not discussed though). Watermarks were entered into the database, together with information of the documents which the watermarks were taken



(a) (b) (c)

Figure 2.22: (a) Input watermarked image, (b) Output image, (c) Output image after applying semi-automatic processing for enhancement. With permission from Thierry Pun [109]

(a) (b) (c)

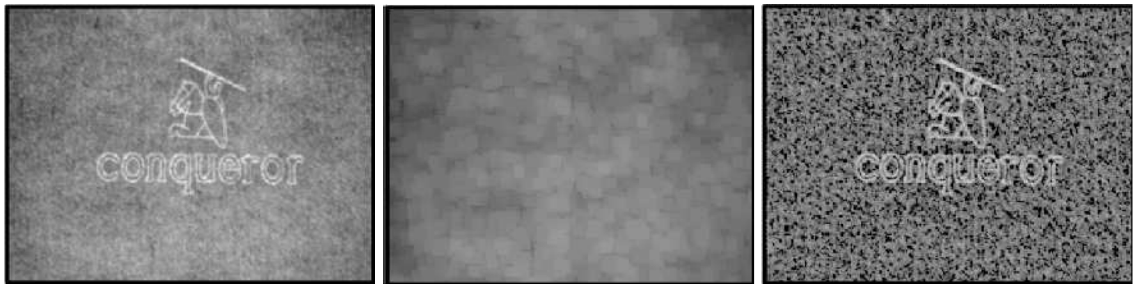
Figure 2.23: (a) Input reflected image, (b) Input transmitted image, (c) Output result [46]

from. This project was used to date and identify Italian artefacts in the National Central Library in Florence, Italy.

Edge [46] also used back-lighting. He used a flatbed scanner (instead of a camera) with a transparency adaptor to capture watermark images in musical manuscripts; he captured both reflected and transmitted images of the watermark. These images were enhanced in order to minimise interference – he used ‘Photoshop’ [1] software to do the enhancement. The reflected image is first inverted, its opacity is changed, and then superimposed with the transmitted image. Figure 2.23 shows input images and result of this approach.

This approach has its limitations. From Figure 2.23(c) we see the existence of interference and furthermore this approach does not work with bound manuscripts, because it uses a flatbed scanner. He also used commercial software for image manipulation, and trial and error for the parameter choice for changing image opacity.

Christie-Miller [27] developed a hardware back-lighting digitisation system. The system, called APIS (Advanced Paper Imaging System), was developed with the cooperation



(a) (b) (c)

Figure 2.24: (a) Input plain watermarked image, (b) Estimated background, (c) Top-hat image result. With permission from Paul F. Whelan [152]

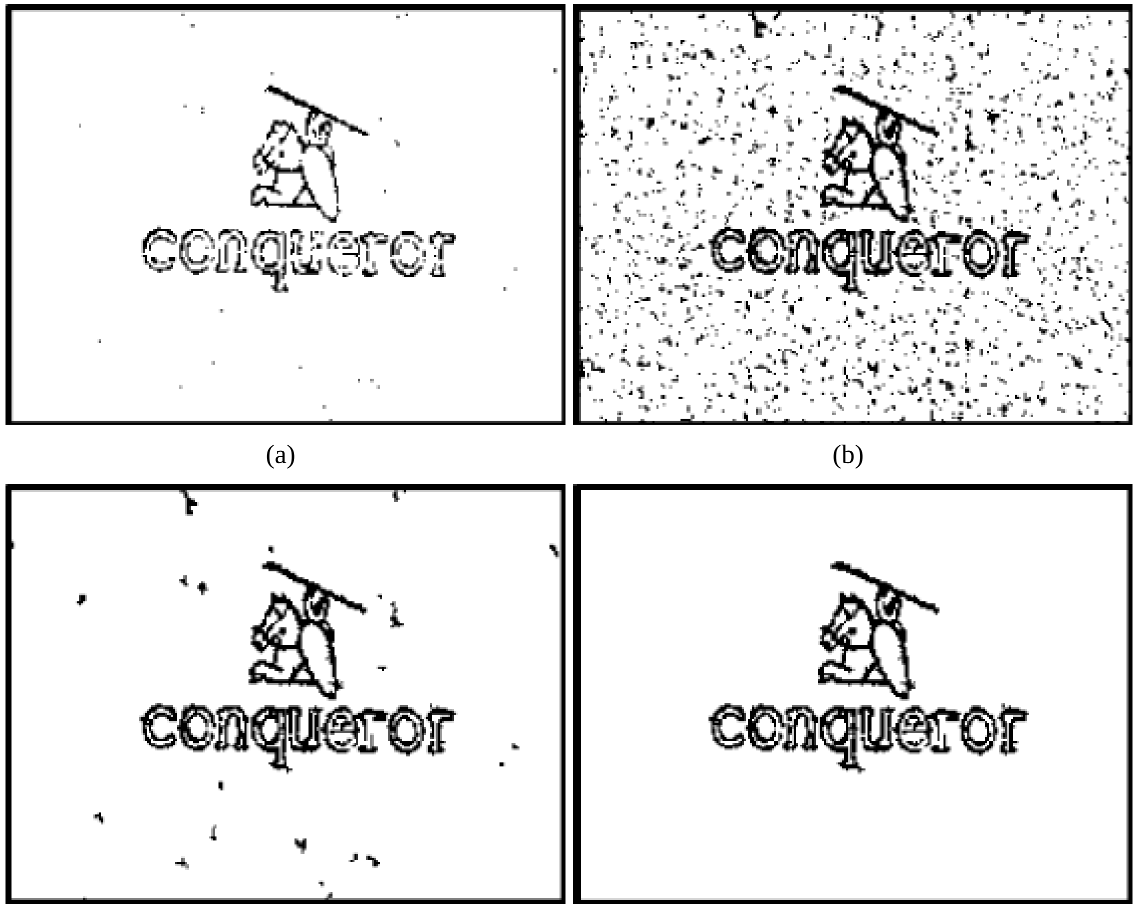
of Solar Imaging Systems Ltd [121]. The purpose of this system is to record the paper structure (including watermark) in order to provide digital fingerprinting [28] which helps in identifying stolen manuscripts. Another purpose was to preserve valuable artefacts and store them digitally, which also assist in studying these artefacts. It allows digitisation of bound manuscripts (opened at 45^o), so the digitisation is safe and does not damage manuscripts.

Whelan *et al.* [152] used back-lighting and image processing in order to extract watermarks from continuous web paper. They work on papers with and without laid and chain lines (laid and wove paper). In the case of wove paper, they started by removing the noisy background by applying the morphological top-hat transform to estimate and remove the image background. However, they did not discuss how they picked the structuring element size for opening operation. The estimated background is then subtracted from the original image (named the top-hat result, A). See Figure 2.24, the input image in Figure 2.24(a) has only a watermark without any interference.

Then morphological reconstruction by dilation is applied to clean any remaining noise; a double threshold operator is used. They first analysed the histogram of image A, and followed assumptions in order to find two thresholds – a detailed description of assumptions and thresholds is in [152]. The first threshold was used for the marker image, the second was used for the mask image, then they reconstructed the mask from marker images (with result B). See Figure 2.25.

The next step is cleaning and filtering. Morphological closing is applied to image B, and small connected features less than a threshold are removed (these features are probably noise): the result is named C. They did not discuss how they picked the structuring element size in the closing operation, or the threshold value. Finally, image B is intersected with C to get the result. Figure 2.25(d) shows a result after extraction.

They also worked on laid papers. They transformed the image using a Discrete Fourier



(a) (b) (c) (d)

Figure 2.25: (a) First threshold of top-hat image: marker image, (b) Second threshold of top-hat image: mask image, (c) Reconstruction of (b) from (a), (d) Output result after filtering. With permission from Paul F. Whelan [152]

Transform in order to remove laid lines – see Figure 2.26(b). Laid lines appear as peaks in the frequency domain due to their high frequency: they applied a selective lowpass filter (a smoothing filter [63]) to these high frequency peaks in order to remove them (as in Figure 2.26(c)), with result *A*. Then they removed chain lines by applying morphological opening – they used subsets of line segments (because of the shape of chain lines) as structuring elements for opening – with result *B*. Then they subtracted *B* from *A*, and applied the previous morphological operations in order to get the result. Figure 2.26 illustrates an example image (which has a watermark, together with chain and laid lines) and the output result.

This method used only the transmitted (backlit) image, and did not benefit from the reflected image. The major drawback of this technique is it did not handle interference caused by writing ink and other features which may obstruct the watermark design. Instead, it concentrated on dealing with watermarked paper without any interference.

Lubbe *et al.* [141] worked on watermarked images of Rembrandt's etchings, reproduced by soft X-radiography. The purpose was to detect and extract patterns of chain lines in order to identify the date of these etchings. Chain lines are first highlighted in images with filtering and morphological operations. These lines are then detected by vertical data projection in images using a selective threshold. However, they assumed that chain lines are always vertical, but watermark images can sometimes be skewed or rotated from the reproduction process. Further, they did not discuss the selection of parameters in the highlighting and extraction of chain lines.

Further improvement to this work was done by Staalduinen *et al.* [144], by finding the orientation of these lines in any direction. Chain lines were located using Fourier and Radon transforms [136] (discussed in Section 4.2.2.1) were applied to find the orientation of these lines in the image. The visualisation of these lines is enhanced using Gaussian filtering. However, the detection is based on the assumption that there is a specific average distance between sequential chain lines, and the number of chain lines in the paper. This is true as long as all lines appear in paper – some may not appear in cases of paper cutting and folding, as appears in Figure 4.15 in Section 4.2.2.1. They also did not discuss the thickness measurement of these lines.

Karnaukhov *et al.* [86] enhanced the blurred watermarked images resulting from the beta-radiography watermark reproduction technique by applying image restoration methods (e.g., Wiener and regularisation filters, which are used for noise reduction in images). An example of a watermarked input image and its output after filtering is illustrated in Figure 2.27.

Wenger *et al.* [150] proposed the INTAS project: A Distributed Database and Processing System for Watermarks [79]. The aim of this project was to build a database for watermarks existing in Russia and West Europe, which can be accessed widely, and will help scholars to study these watermarks and date undated documents. Another aim was to study and improve reproduction techniques, including radiographic, back-lighting and rubbing techniques.

Results of this project appeared in [149]; it included the birth of the first two electronic watermark databases in Russia. This project also resulted in analysing and evaluating reproduction techniques. Reproduced images were enhanced (contrast enhancement), and watermark contours were approximated using semi-automatic processes [151] for identification purposes. These enhanced images are then entered into the database. Emanuel Wenger is the coordinator of the *Bernstein – The memory of papers* [13], an ongoing project for studying watermarks in paper. It aims to create a digital environment for re-

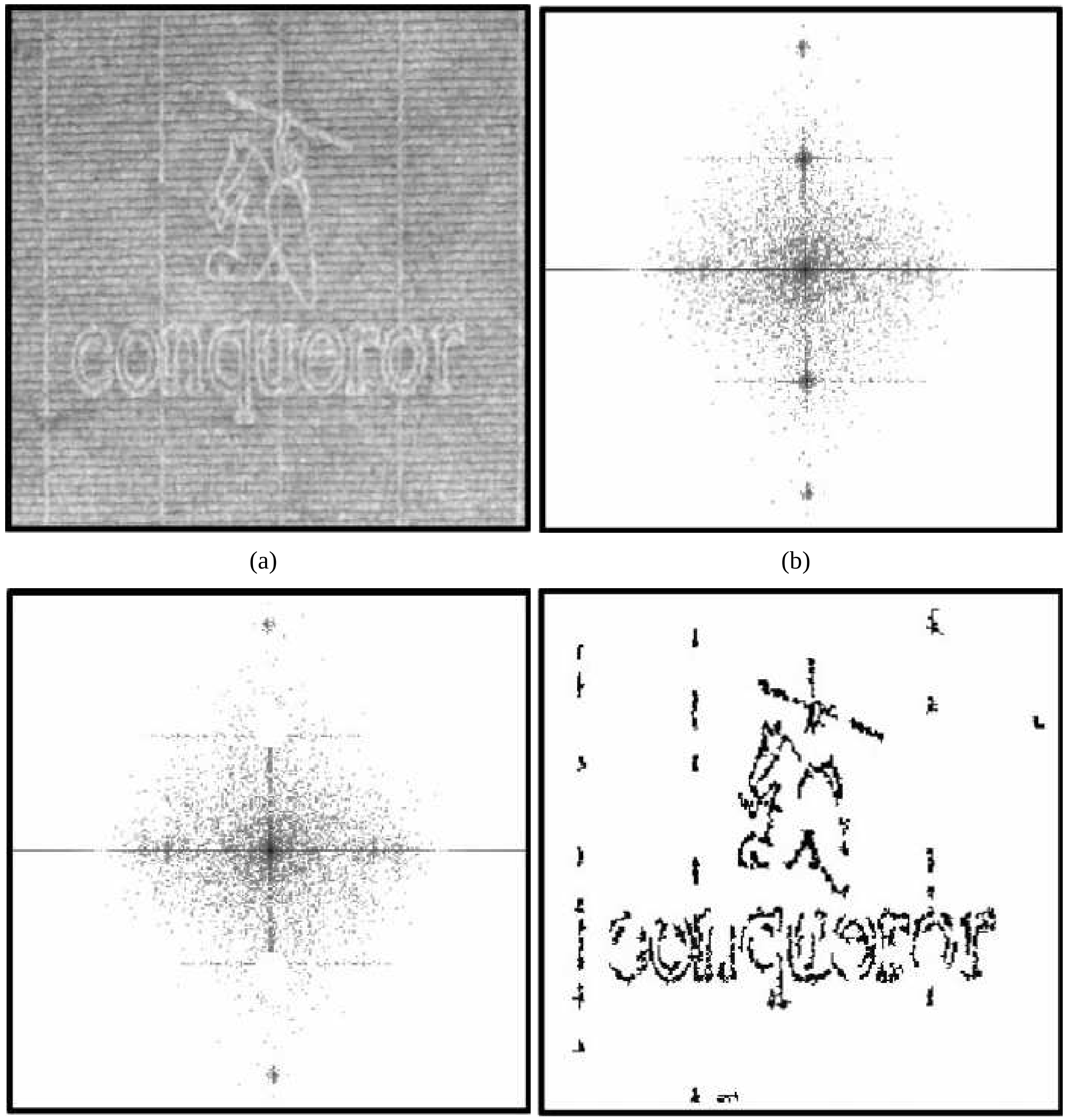


Figure 2.26: (a) Input watermarked image with wire and chain lines, (b) Discrete Fourier Transform frequency spectrum as an intensity function, (c) Selective lowpass filtering of (b), (d) Output result after filtering and double threshold. With permission from Paul F. Whelan [152]

(a) (b)

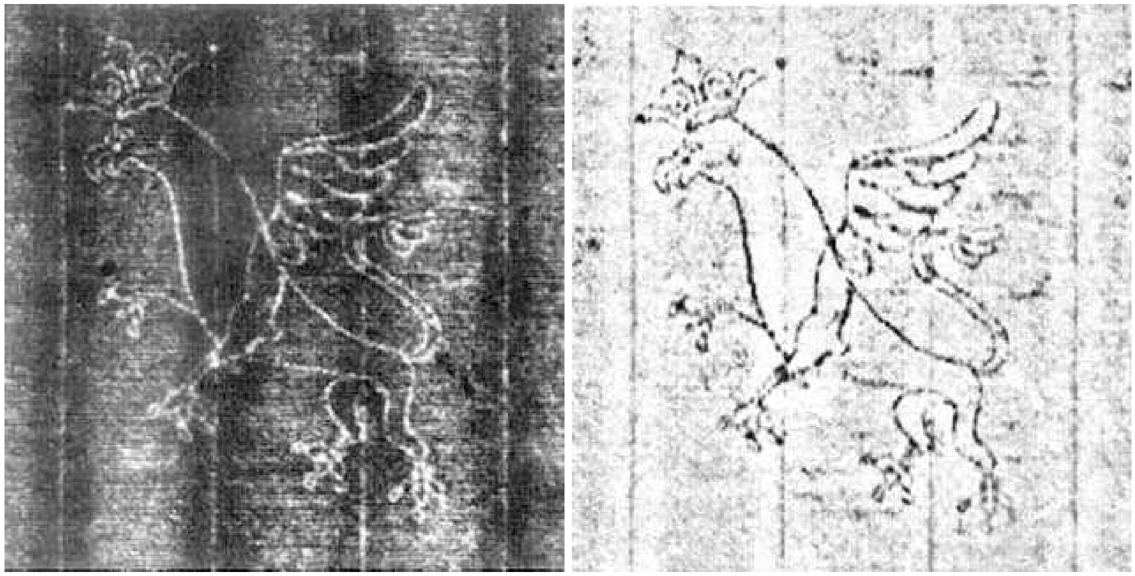
Figure 2.27: (a) Input blurred watermarked image, (b) Output result after applying image filtering. With permission from Alois Haidinger [86]

searchers to study paper: it will link all the European databases of reproduced watermarks together, and provide image processing tools to measure paper features.

Profil is another watermark database project [34] – its aim is to offer scholars the ability to identify watermarked paper. Data was reproduced using beta-radiography in the National French Library; these watermarks were scanned and entered to the database, together with a description of each watermarked document. Then, processes are performed to remove defects in images. The contrast is enhanced by applying lowpass filtering to the image in the Fourier spectrum, the filtered image is then subtracted from the original, then the image is filtered (e.g., median, Gaussian filters, etc) to remove remaining noisy patterns. An example input and its output result after enhancement are in Figure 2.28.

SHREW 'SHape REtrieval of Watermarks' is a database project for image retrieval of historical watermarked papers. SHREW enhances the visualisation of watermarked images and stores them in a database; a given watermark can then be matched with stored watermarks and similar shapes are retrieved [43].

Input data were traced watermarks by Churchill [29], and images reproduced by electron-radiography. Traced watermarked images were processed for feature extraction: images are first converted to binary using a constant threshold, then noise is reduced using filters (e.g., mean, median and Gaussian), images are then enhanced using morphological closing to strengthen thin and broken lines in tracings. These enhancements were combined with shape retrieval techniques in order to get better results [111]. An example of a



(a) (b)

Figure 2.28: Griffon watermark, (a) Input image reproduced using beta-radiography, (b) Watermark image result after enhancement. With permission from Claire Bustarret [34]

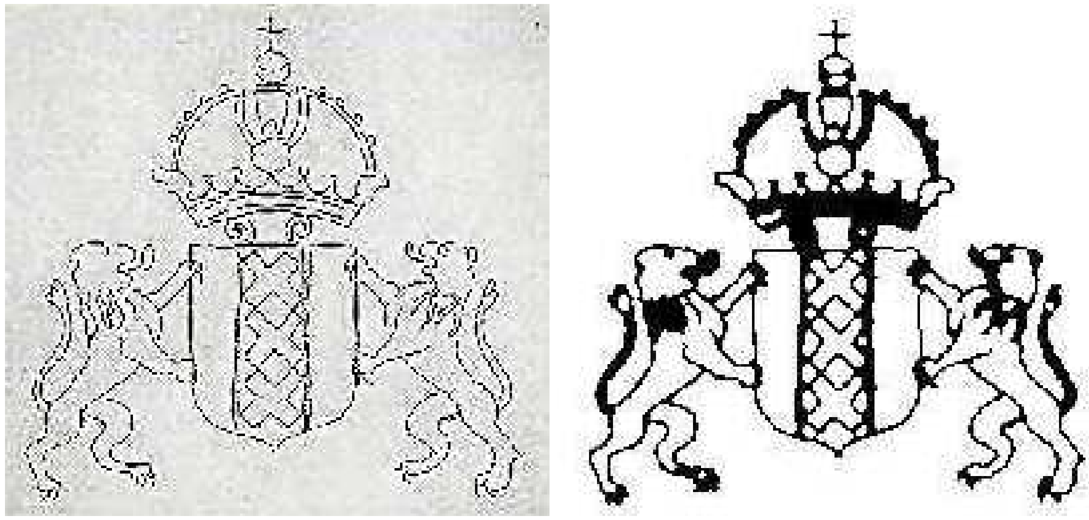
traced image and its output after enhancement is in Figure 2.29.

SHREW was further developed and evaluated in [112]. The other datasets were reproduced using electron-radiography. In addition to their previous enhancements, chain and laid lines were removed by applying lowpass frequency filtering, the background was approximated by applying a median filter *n* times to the watermark image, and then subtracted from the original image. See Figure 2.30 for an input watermark image using electron-radiography, and output after laid line suppression.

The main drawback was the lack of treatment of interference by writing and such; noise reduction by lowpass filtering did not give good results, and results of images reproduced by electron-radiography was not as good as results of traced watermark images.

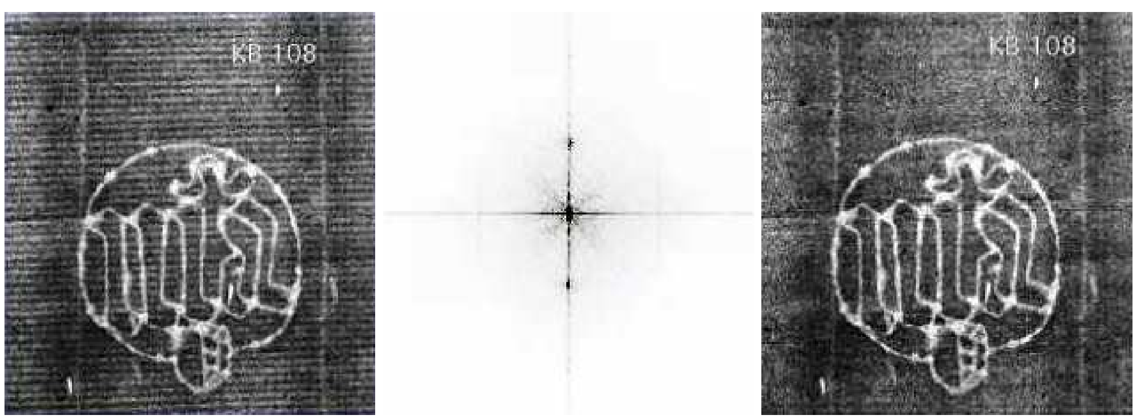
Van Aken [140] improved the contrast in soft X-radiography technique using a hardware solution using Helium gas. His improvement made the exposure time shorter, allowed the non-darkroom conditions, and improved the contrast in results. A result of using this improvement is in Figure 2.31, these images are in low resolution due to the source they are taken from.

Another project that used the combination of back-lighting and image processing was presented by Jin [37] and Ng *et al.* [102]. The approach used the back-lighting system that we also used in our digitisation (described in Section 3.2), followed by image enhancements to extract watermark features. They enhanced the transmitted image contrast, then applied edge detection. Detected features are then converted to vector representation



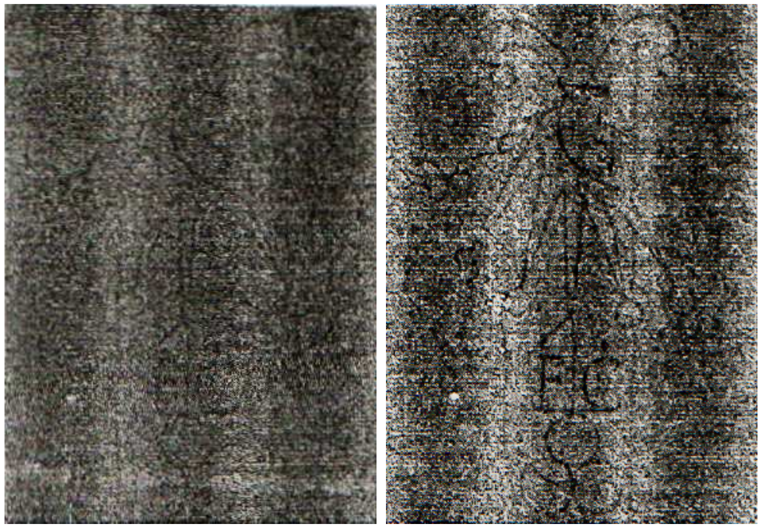
(a) (b)

Figure 2.29: (a) Input traced watermark image, (b) Output image after enhancement. With permission from Jean Brown [43]



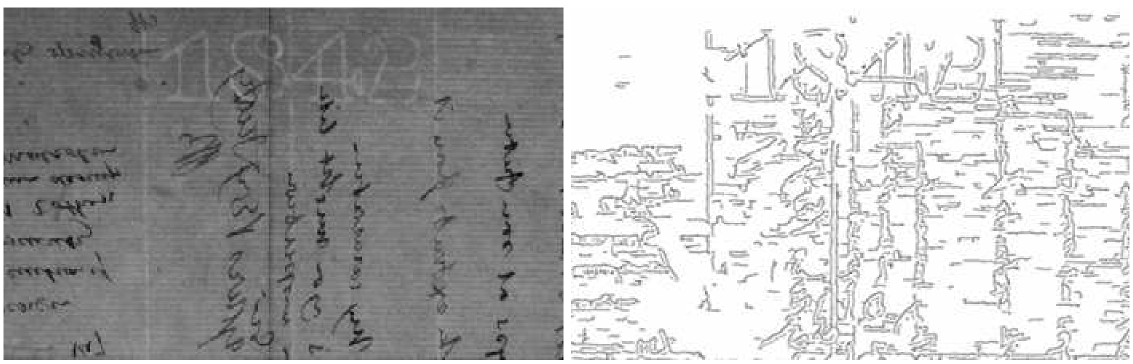
(a) (b) (c)

Figure 2.30: (a) Input watermark image by electron-radiography, (b) Watermark image in frequency domain, (c) Output after laid lines suppression. With permission from Jean Brown [43]



(a) (b)

Figure 2.31: Watermark image using soft X-radiography, (a) without Helium at 10 keV, (b) After improvement, with Helium at 5keV [140]



(a) (b)

Figure 2.32: (a) Watermark image using back-lighting, (b) Output result [37]

in SVG (Scalar Vector Graphics) format [135]. Results of this approach suffer from interference which obstructs the watermark pattern. Example of an input transmitted image and its output is in Figure 2.32. The result of the same watermarked image using our approach is in Figure 4.23 in Section 4.3.

Van Staalduinen [143] enhanced reproduced watermark images from back-lighting or soft X-radiography techniques by suppression of laid lines, and background variation. The same approach used in [152, 157] was used to detect and suppress laid lines, while background variation was estimated by means of the background mean and variance estimate. Both reproduction techniques were compared qualitatively (from an art expert's point of view) and quantitatively (by image analysis techniques). Results showed that the soft X-radiography technique is better – details of comparison are in [145].

Neuheuser *et al.* [93, 100] used a thermography watermark reproduction technique to distinguish originals from prints, and to identify watermarks in Rembrandt's drawings. Figure 2.33 illustrates the team and the setup they used, with a watermark image result.

Figure 2.33: Thermography setup, with a watermark image result. With permission from Peter Meinlschmidt [50]

Atanasiu [9, 10], working in the Bernstein project [13], developed two applications which helped in studying laid lines. The first is for laid line density measurement, known as 'AD751', which locates the frequency of these lines in Fourier transform [11], and the other is for laid lines suppression and extraction, known as 'BlueNile'. Other useful applications are 'Filigrana', which is another laid lines density measurement tool, 'WatermarkScissors' is an application which segments an image which contains a number of watermarks, into smaller images according to the number of watermarks, and 'WMT' is an application which measures width and height of watermarks interactively [13].

2.5 Discussion

After this introduction of the history of paper watermarks and its making, its importance in early and present days, and after reviewing other approaches for extracting these features, we consider advantages of our approach, and discuss the limitations of other works.

Tracing, back-lighting and radiographic reproduction techniques are the most commonly used approaches by scholars nowadays. The approach presented in this thesis is back-lighting (described in Section 3.2), because it is simple and requires relatively low cost equipment; captured watermark images are generated digitally in a very short time. This makes it easier to preserve and store them in digital archives that can be accessed remotely. Radiographic techniques are more expensive, unsafe, time-consuming and hard to reach for individual researchers. Tracing is simple and cheap, but it is not accurate and needs skill and experience.

Back-lighting allows further image processes approaches to be applied easily in order to highlight watermark patterns and remove interference caused by writing ink (on both sides of paper), together with noisy and uneven background illumination, and other unavoidable existing damage on paper. Captured images are of a high resolution, which allows the observer to see very small details of the image.

Related works reviewed in Section 2.4.2 suffered from interference that prevented a clear watermark design. Other works lacked the adaptive selection of parameter choices in image processing algorithms; our developed approaches managed to output watermark images with minimum interference, and presented effective approaches to automate parameter selection.

This work is divided into two approaches. The first, a bottom-up approach, presented in Chapter 4, was developed to extract watermarks from paper – this approach will help preserving these important artefacts, and will allow wider accessibility for scholars. These data are presented in Section 3.1.2. The system gives effective results with the minimum interference compared to others' work. This approach was further evaluated, and processed to export watermark images to vector forms in Chapter 6. The system was built with an interactive interface in order to aid historians (who do not have experience in using computers) to use it easily.

The second approach attempts to model back-lighting, and is presented in Chapter 5. This approach serves as a watermark image retrieval utility, and was developed to locate watermarks in more difficult data than those in Section 3.1.2. These data are presented in Sections 3.1.3, 3.1.4 and 3.1.5. These data have the importance of being a valuable artefact, since these are complete handwritten collections of the Qur'ān and Prayer; these data are characterised by thick writing strokes on both recto and verso. The paper used in writing this manuscript is thick, and the watermark patterns are not clear, which resulted in high interference, and a weak signal of the watermark shape. This approach aggregates similar watermarks to provide accurate details which may not be clear in individual sheets. It also distinguishes 'twin' from 'identical' watermarks. Results of this approach are promising.

In the context of a complete digitisation, it is not realistic to only extract and preserve paper watermarks that have a clean surface. Most of the manuscripts we are working on for preservation purposes contain important foreground visual information as well as the watermark and hence the proposed methods make use of image pairs (one with normal visible lighting and one with back-lighting) for the digitisation stage. The image capture with normal lighting is used for the digitisation of the surface of the paper while the image pair is used for the framework described.

As a result of our work, on-line web archives of these manuscripts are now available in [76, 77].

Chapter 3

Source material and Digitisation procedures

This Chapter presents a description of material used for prototyping. These data are principally manuscripts from the eighteenth and nineteenth centuries, held by the Special Collections at the Brotherton Library of the University of Leeds [123]. We also present the digitisation setup we used for image acquisition; this is equipped with hardware to permit the back-lighting technique described in Section 2.4, digitising these artefacts will preserve its important historical value and provide better access and distribution. We then present a description of the characteristics and quality of paper and watermarks found in our data.

3.1 Materials used for prototyping

3.1.1 Modern paper

We used our digitisation setup to capture watermarks in modern paper which holds a logo of the University of Leeds as a watermark. This is positioned in the paper centre: an example of such currently used paper is in Figure 3.1 (zoomed and enhanced for display). We used this type of paper as a benchmark for the approach presented in Chapter 4.

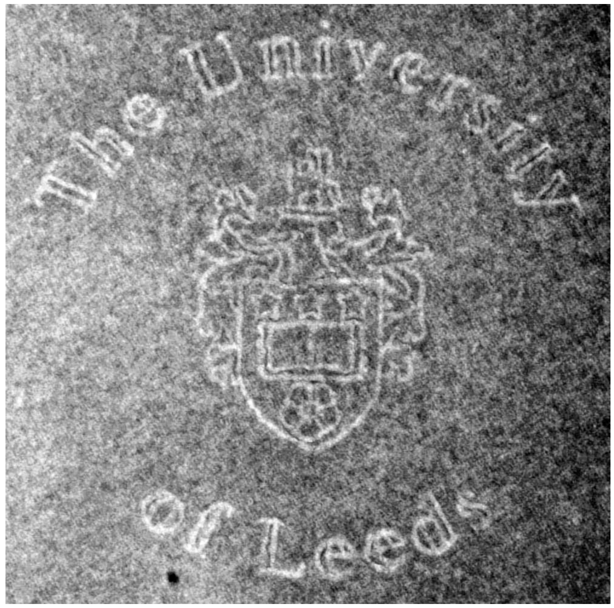


Figure 3.1: Modern transmitted paper (zoomed and enhanced for display)

3.1.2 Individual manuscripts

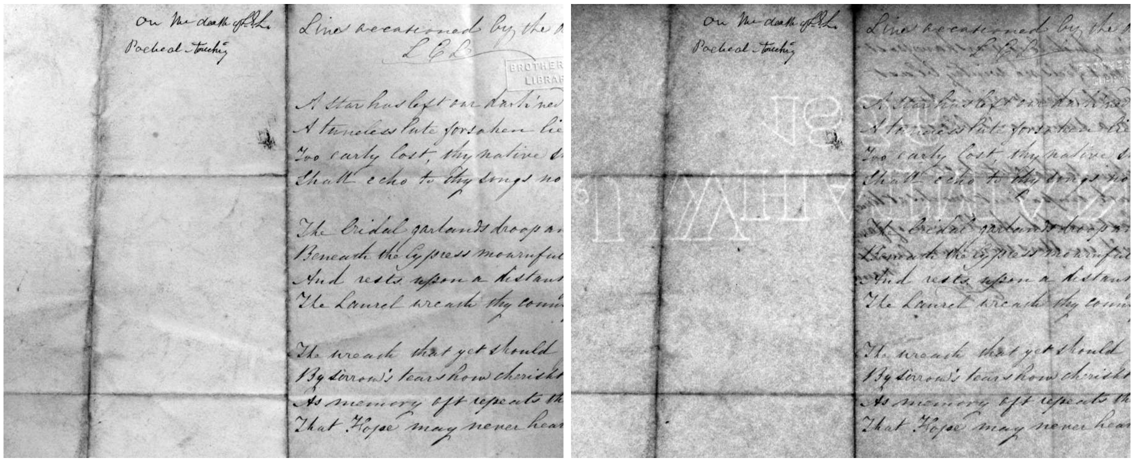
Part of our data was individual musical and handwritten manuscripts. These manuscripts are taken from the works of Henry Litolff [14], digitised with permission from the Special Collections at the Brotherton Library of the University of Leeds [123]. Paper used for these manuscripts is laid (with chain and laid lines) and wove, and has a variety of watermarks. Examples of these manuscripts (zoomed and enhanced for better visualisation) are in Figures 3.2 (for wove paper with the 'J WHATMAN/1836' watermark) and 3.3 (for laid paper with the '1824' watermark). Further full illustrated examples of these manuscripts are in Pages 153 – 156. These manuscripts were used in the approach discussed in Chapter 4.

3.1.3 The 'Mahdiyya' copy of the Qur'ān

This manuscript is held by the Special Collections at the Brotherton Library of the University of Leeds (MS Arabic 619). It is a complete copy of the Qur'ān written in 1881 (1299 Hijri) in Sudan. It was taken 18 years later by Bimbashi T. E. N. Lewis, a British major, in Um Debrekat in Sudan. The Qur'ān was "found in the saddle-bag of an Emir who was killed near the Khalifa (Abdullahi) on the occasion of the latter's death at Um Debrekat (Gedid) on 24th November 1899" [24].

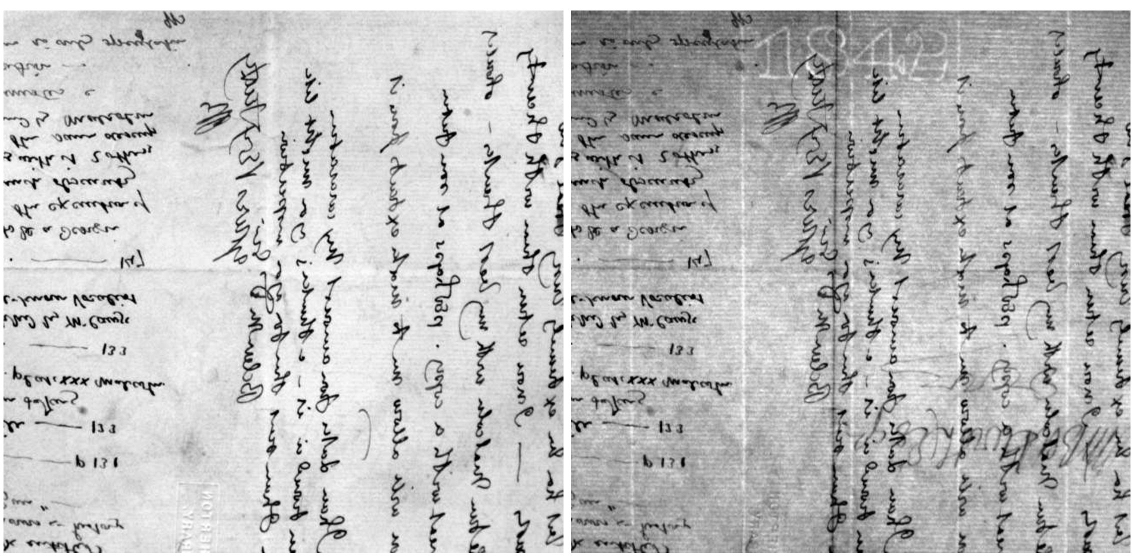
A brief description of the manuscript, taken from Brockett [24]:

> The manuscript is written on laid paper, folios 346 (except pages 247, 341



(a) (b)

Figure 3.2: Historical wove paper (zoomed), (a) Reflected, (b) Transmitted



(a) (b)

Figure 3.3: Historical laid paper (zoomed), (a) Reflected, (b) Transmitted

Figure 3.4: Cover of the ‘Mahdiyya’ copy of the Qur’ān

> and 342, which were taken from a different paper type), paper dimensions are 234–238 × 160–164mm, writing area is 170–175 × 100–102mm, 13 lines of writing per sheet, the manuscript is written in east Sudani naskh. Writing and vocalisation is in black ink, while sūra titles, verse-dividers, recitative notations and marginal notes are in red ink, no decoration exists, and cover is made of leather (as illustrated in Figure 3.4).

Except for three pages, only one paper has been used for this Qur’ān, bearing a watermark and its countermark. The watermark is the two-headed (or double-headed) eagle of the Austro-Hungarian Empire with a sword and sceptre. The countermark ‘Andrea Galvani Pordenone’ with a moonface-within-shield, reveals the name of the fabricant, Andrea Galvani, and the city where the mill was based, Pordenone (situated in the Frioul, in the North-East of Italy). This countermark was first used in Egypt in 1868, and in Sudan from 1870 [147]. Page 247 bears a tre lune (three moons) watermark, with human faces and the arc curved at the top and bottom edges. Pages 341-2 hold another moonface-within-shield design.

The watermark and countermark are divided into two parts in this manuscript. None of the pages contain a complete design of the watermark or countermark, these designs appear on the edge of paper sheets. After using this manuscript in our approach presented in Chapter 5, and after superimposing the similar designs together, we later determined that there is another countermark placed under the double-headed eagle, probably ‘A G’, a

(a) (b)

Figure 3.5: Sample from the ‘Mahdiyya’ copy of the Qur’ān (zoomed), (a) Reflected, (b) Transmitted

well-known countermark that denote Andrea Galvani. Complete and clean designs of watermark and countermark are illustrated in Figures C.7 and C.8 in Appendix C. A sample of this manuscript (zoomed and enhanced) is shown in Figure 3.5. This example shows part of the paper, with lower part of moonface-within-shield, and the ‘Andrea Galvani Pordenone’ countermark. Sample full illustrations can be found in Pages 157 – 160.

3.1.4 Islamic Prayer

This manuscript is an Islamic Prayer and also held by the Special Collections at the Brotherton Library of the University of Leeds (MS Arabic 86). Catalogue notes identify it as:

> Kitāb Durrat ‘iqd al-naḥr fī ‘asrār ḥizb al-baḥr. No date is given but it is believed to be in the 18th century. The commentary (on the Prayers) is by the Ṣūfī ‘Abd al-Raḥmān b. Muḥammad b. ‘Alī b. Aḥmad al-Bisṭāmī (d.858/1454, [23] vol.II, p300). The main Prayer (or Prayers) is by Nur al-Dīn Abu al-Ḥasan ‘Alī b. ‘Alī b. ‘Abd al-Jabbār al-Ḥasanī al-Idrīsī al-Mi’marī al-Shādhilī (d.656/1258, [23] vol.I, p583). The work comprises the Muqaddimah, Prayers by al-Shādhilī, Aḥmad al-Malawī, a Risāla by Abu al-Ḥasan al-Hindī, and a ḥizb by Ibrāhīm al-Dasūqī.
>
> The manuscript comprises 32 folios, 8.5 × 6in, written in single columns of 17 lines to page, within a border of two red lines, 5.75 × 3.25in. It is on good, waxed, vertically-laid paper (horizontal layer to the inch), in clear Naskh, with a few vowel- and orthographic signs. Rubrics and original text are in red, with no annotations. The folios are loose within stained, brown

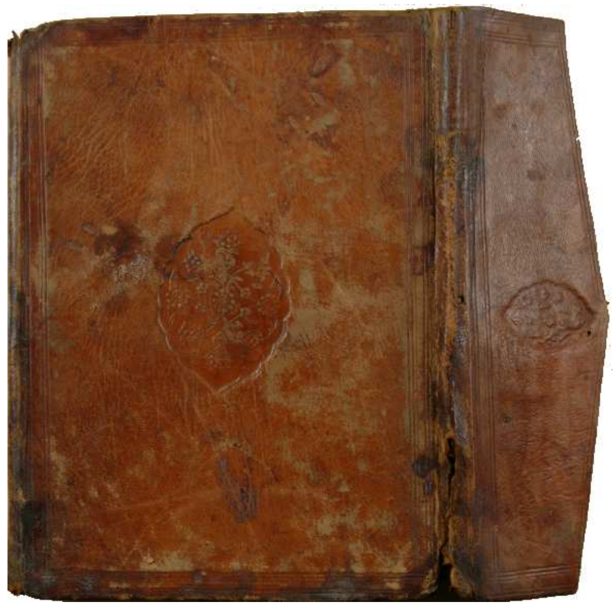


Figure 3.6: Cover of the Prayer manuscript

> leather covers, with flap, each ornamented with indented medallia (as illustrated in Figure 3.6). There is simple 'Unwān in black and red within triangle of red lines in folio 1.

The watermark used in this manuscript paper is tre lune (three moons), with a letter 'C' as countermark – an initial or symbol indicating the paper-maker's name, appearing opposite the main watermark on the other half of the mould and usually smaller than the watermark. Each pair of pages is bound together, which permits a complete design of the watermark to appear clearly. We used these data in our approach presented in Chapter 5. An example of this manuscript with watermark (zoomed and enhanced) is in Figure 3.7; sample complete illustrations are in Pages 161 – 162.

3.1.5 The 'West African' copy of the Qur'ān

This manuscript is also held by the Special Collections at the Brotherton Library of the University of Leeds (MS Arabic 301), and is a complete copy of the Qur'ān. It carries neither date nor other information of origin, but the script used is west African, called 'Sūdānī Maghribī'.

The manuscript was described by Ebied [45] and Brockett, a description is taken from the latter:

> fol. 332 (163 bifolios, 6 folios); 220–230 × 160–167.5mm; written area 150–160 × 100–ll0mm; 16–20 lines per page; laid paper; bold Ifrīqī hand

(a) (b)

Figure 3.7: Sample from the Prayer (zoomed), (a) Reflected, (b) Transmitted

in shiny black ink, with diacritics in black, vocalisation in red, and hamzat al-qaṭ’ in yellow; sūra-titles in the same hand but in red, with diacritics and vocalisation in black; marginal decorations in red, brown, yellow and black; 4 larger decorations in ‘earthy’ yellow, reddish brown and black (ff. lb, 8lb, 163a, 246a); strong, leather loose-cover binding, stained reddish brown, with dark brown (almost black) associated with the tooling, ending in an envelope-flap and strap for fastening; the whole contained in a rigid suede-leather satchel, with a triple flap, thongs and straps (as illustrated in Figure 3.8); no date.

The manuscript contains the tre lune watermark, which appears in different variations, one reason for which may be twin moulds for paper-making (see Section 2.1). Another reason may be movement of the watermark along the mould [24]; the wire forming the watermark seems to be attached to the mould improperly – some pages have the largest crescent rotated by a large angle. See Figure 3.9 for a sample of this manuscript, together with variations in the tre lune watermark (zoomed and enhanced).

The countermark used is the letter pair ‘C L’, with two variations, which proves that twin moulds were used in paper-making. Part of the manuscript also has the tre lune with human faces (three moonfaces) watermark with the ‘Andrea Galvani Pordenone’ countermark. See Pages 163 – 166 for full illustrated samples of this manuscript.

The manuscript is not dated, but with the help of watermarks and countermarks, the manuscript is estimated to have been written mid 19th century, between 1836-80 [24], because the countermark corresponds to the Venetian Andrea Galvani firm, providing

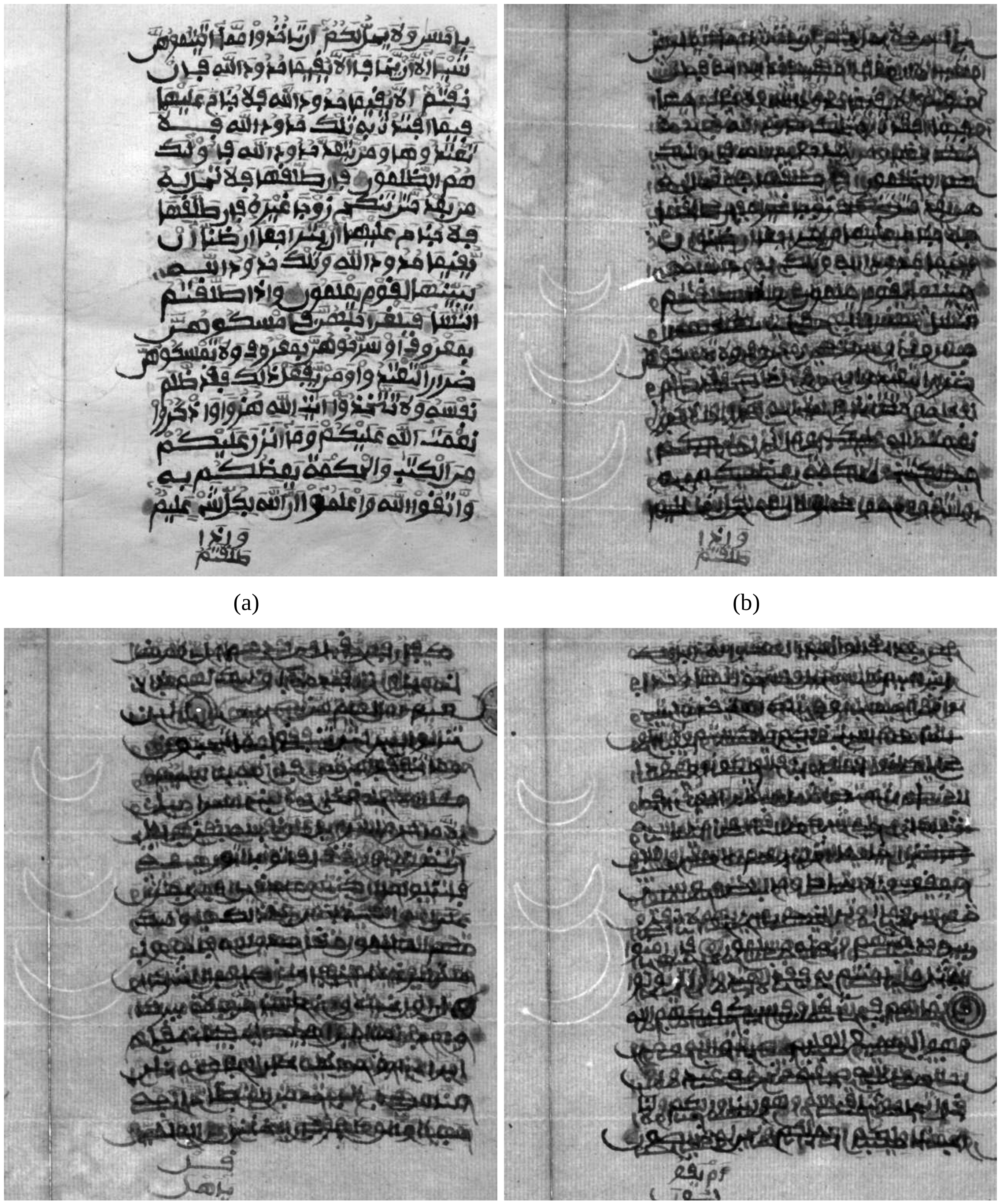
Figure 3.8: Cover of 'West African' copy of the Qur'$\bar{a}$n

1836 as the earliest paper-making date. Such paper was used in Egypt and western Sudan until 1880. Brockett suggested that the manuscript date is closer to 1836 rather than 1880, because the first use of three moonfaces watermark in Egypt was in early 1840s [147], and so around this date in western Sudan. This manuscript was also used as input data in our approach described in Chapter 5.

3.2 Digitisation procedures

The digitisation system used for capturing reflected and transmitted images was made available by the Interdisciplinary Centre for Scientific Research in Music (ICSRiM) [101]. This system is mounted using a stand with lights by Kaiser Fototechnik [85]. We used a FUJIFILM FinePix S1 Pro camera [51] in capturing our images. The system uses a light sheet for back-lighting: this is a thin foil of light with even homogeneous illumination behind the paper, used to visualise the watermark pattern. Each paper sheet is captured three times, reflected images of front and back, and a transmitted image (which captures the details of paper structure, including the watermark, together with laid and chain lines).

The camera comes with capturing software, which permits simple transfer and viewing of captured images, controlled from a PC via a USB connection with the camera. The camera uses Super CCD (Charge Coupled Device) image sensor technology and a 'Nikon



(a) (b) (c) (d)

Figure 3.9: Sample from the 'West African' copy of the Qur'ān (zoomed), (a) Reflected, (b) Transmitted, (c) Variation of tre lune watermark (twin watermark), (d) Another variation of tre lune

F' lens. It captures images with a resolution up to 3040 × 2016 pixels (6.13 megapixels). Full specifications of the camera, its functions and shooting software are in [51]. The 'Mahdiyya' Qur'ān, individual manuscripts, and University of Leeds paper were captured at a resolution of 258dpi, while the Prayer and the 'West African' copy of the Qur'ān were captured at 220dpi. During the digitisation process, it is important to position pages as consistently as possible: this will be important in locating watermarks using the approach presented in Chapter 5.

3.3 Data description: watermark and paper qualities

This Section discusses characteristics of the paper and watermarks of manuscripts presented in Section 3.1. The paper bearing the University of Leeds logo watermark has a uniformly textured background, and even illumination along the sheet. The watermark pattern is partially impaired by a background pattern, which cannot be clearly seen. Results of using this paper are shown in Figure 4.26 in Section 4.3.

Individual musical and handwritten manuscripts have interference caused by writing and other defects. Thin pen strokes were used in writing on paper (i.e., radius of the nib), the background is not uniform, and the paper used is thin. Watermarks (and laid and chain lines) appear clearly in most of the paper. This type of data was used successfully to extract watermarks as presented in Chapter 4; output may be seen in Section 4.3.

The 'Mahdiyya' copy of the Qur'ān was the most complex data we investigated. These data are challenging for several reasons:

- Its importance as a complete handwritten collection of the Qur'ān.
- The paper sheets and writing strokes on recto and verso are thick.
- The background is not uniform.
- The watermark patterns are not clear and of poor quality.

All these characteristics present high interference with watermark patterns.

The Prayer and the 'West African' copy of the Qur'ān were also challenging. They are valuable artefacts, and also have thick pen strokes, thick paper (but not thicker than the 'Mahdiyya' Qur'ān), but watermarks are clearly visible. Part of the 'West African' copy of the Qur'ān has poor watermark quality, especially the three moonfaces watermark. Both Qur'ān copies and the Prayer data were successfully used to locate watermarks in our approach described in Chapter 5. The paper type used in both Qur'ān and Prayer manuscripts is laid paper.

Chapter 4

A bottom-up approach

4.1 Introduction

Challenging pattern recognition and extraction problems are often approached in two independent ways:

- *Bottom-up* approaches, in which individual basic operations of the system are specified in detail, and are then connected to build larger sub-systems, which are joined together to form the main or top-level system.

- *Top-down* approaches, when an abstract or overview of the system is derived and mapped onto observation, and then divided into specified sub-sections, these are then further divided until detailed basic operations are specified [154].

In this Chapter we consider the former strategy and derive a process that pre-processes, highlights the watermark, and removes foreground and background interference. After this, the segmentation stage offers the localisation and extraction of watermark pattern and chain lines.

This sequential approach is demonstrated on a range of inputs and shown to be successful: it has limitations, however, which we also demonstrate, which lead to a complementary top-down approach discussed in Chapter 5.

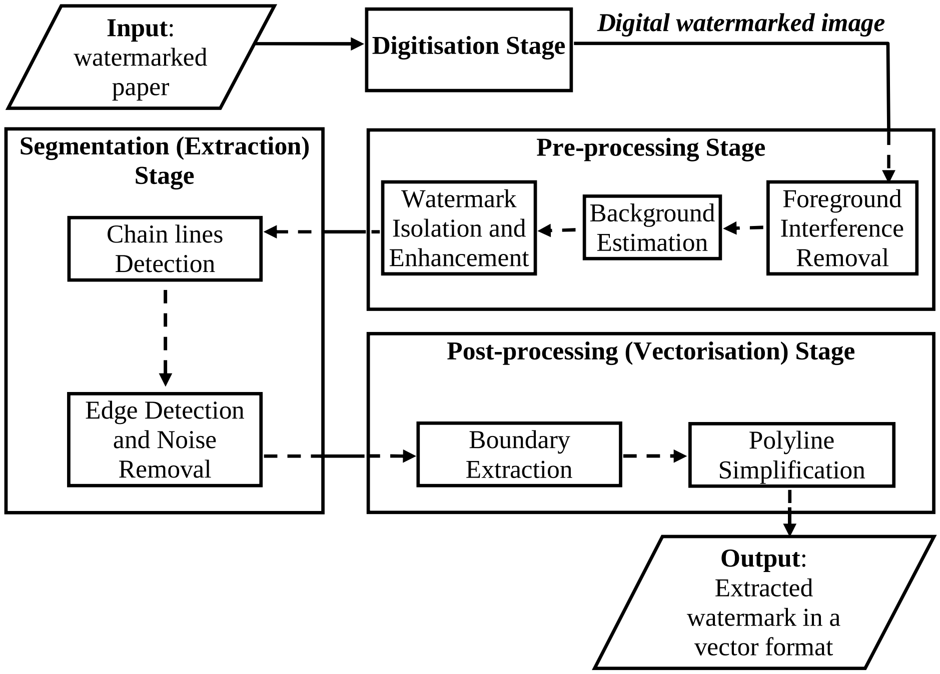


Figure 4.1: Flow chart of the bottom-up watermark extraction approach. Digitisation is described in Chapter 3 and vectorisation in Chapter 6.

4.2 Paper-based watermark extraction

This approach operates in two main stages:

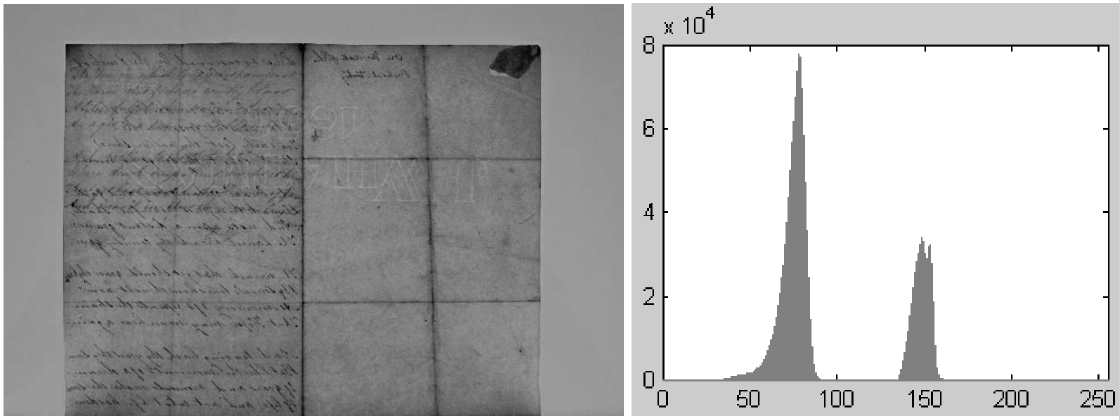
Pre-processing Image processing is applied to highlight the watermark and remove foreground and background interference. This is an important stage that provides the key advantage to this system since it handles typical noise and recto and verso writing and markings.

Segmentation The localisation and extraction of watermark pattern and chain lines.

A further post-processing stage is described in Chapter 6, in which a graphical representation of the segmented watermark is created as a vectorised description.

An overall flow chart of this approach with various stages is illustrated in Figure 4.1.

The overall process time depends on the PC machine speed and memory, complexity (the amount of interference caused by writing ink and other defects) and size of the image. It generally requires around two minutes with image size of around 1500×1000 pixels, with a Pentium 4 PC of 2.8GHz speed and 1GB RAM.



(a) (b)

Figure 4.2: Input backlit image and its intensity histogram. The watermark is presented on Page 154. This document is part of the works of Henry Litolff [14]. The text is readable on Page 153.

4.2.1 Pre-processing

The pre-processing stage focuses on highlighting and isolating the watermark from other digitised contents of the paper using morphological operations [63]. The digitised image normally consists of the paper (in the centre) with a border region due to the lighting sheet during digitisation. For better estimation of dynamic thresholds, the pre-processing stage starts with the localisation of the region of the paper in the image by analysing its grey-level distribution. Figure 4.2(b) illustrates how the distribution of the pixels of the paper region is separated from the surrounding border, since it is brighter. This area is removed by histogram thresholding; we pick the threshold as the highest intensity value in the first area (95 in this example). All intensity values above this threshold are set to 0. See Figure 4.3 for the transmitted (backlit) image (of Figure 4.2(a)) after border removal. A larger illustration of this sample image is on Page 154.

A series of steps is then applied in order to extract the watermark design by separating the image into a number of layers. Firstly, foreground interference, such as writing ink, is removed by producing an intermediate image I_a with the background and the watermark. Next, the non-uniform background of the image (e.g., paper texture, noise, folding marks, etc.) is estimated as I_b. After that, the difference image of I_a and I_b is produced $I_w = I_a - I_b$, which contains the watermark (and some residual noise).

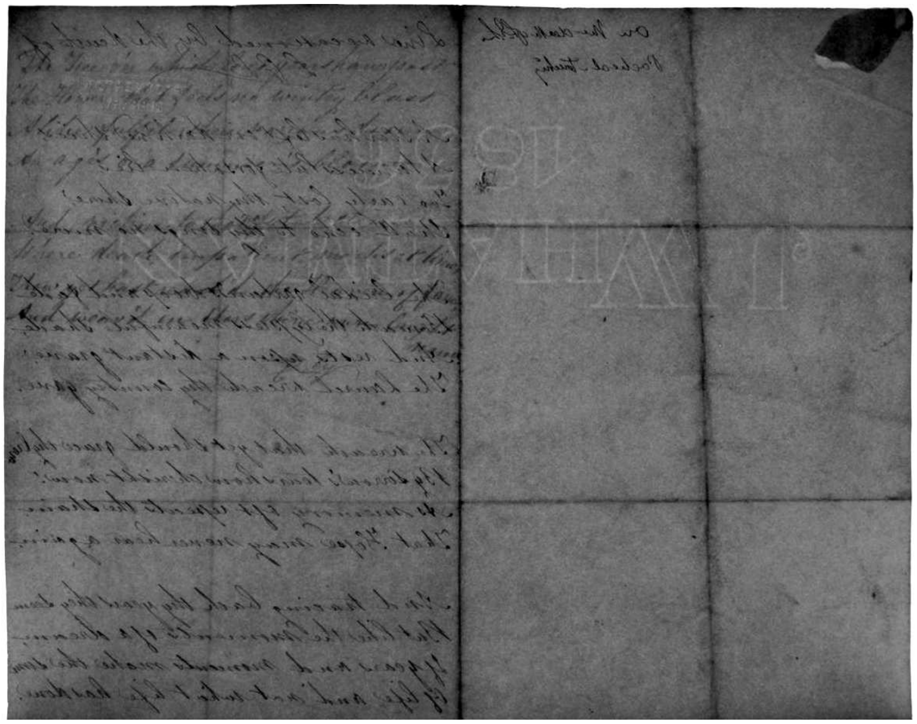


Figure 4.3: Backlit image after border removal

4.2.1.1 Foreground interference removal

In order to extract the watermark pattern, it is necessary to minimise, as far as possible, interference caused by the obstructing writing ink. In the examples we present (as in Figure 4.3), the writing ink is black, so the darkest pixels identify the writing. Also, in this type of data, writing features, either on recto or verso, are thinner than the watermark features, this fact motivated us to use morphological operations to suppress this interference. We devised a combination of morphological dilation ($C = A \oplus B$) and erosion ($C = A \ominus B$) operations, where A and B are the image and the structuring element respectively [63]. These operations are effective in writing removal, because they have the advantage of removing small black holes or gulfs represented by such features [122].

The size of structuring element B used in dilation to remove such interference is critical – choosing a non-suitable structuring element size will affect the clarity of the watermark pattern and make it blurry, as illustrated in Figures 4.7(e) and 4.7(f). The motivation behind this approach is to determine this parameter automatically to permit optimal content processing without time-consuming manual intervention. The following steps (illustrated in Figure 4.4) explain this approach:

1. Applying a contrast stretching process [63], so the darkest pixels will take a zero intensity value (as illustrated in the histogram distribution in Figure 4.5);

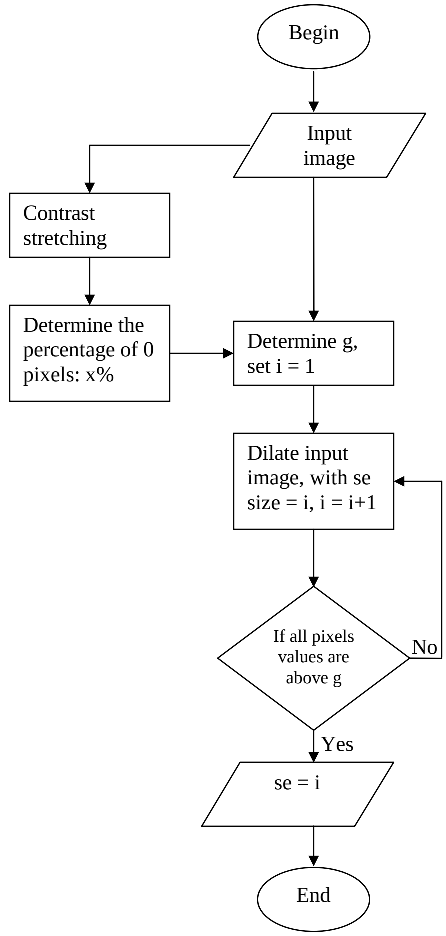


Figure 4.4: Flow chart of the foreground removal approach

2. Determining the percentage of such pixels: x%;
3. Within the original image, determine the grey level g such that x% of pixels are at intensity g or less;
4. Dilate the input image, starting with structuring element of size 1, and increasing the size until all pixels values are above g (as illustrated in Figure 4.6);
5. The final structuring size is taken as the optimal value to remove foreground interference.

Example results of iterated dilation using this algorithm on the image in Figure 4.3 are illustrated in Figure 4.7 (enhanced for better visualisation) – the writing fades out with iteration. The dilated image is then eroded in order to clarify remaining image features (including the watermark) resulting from dilation. Figure 4.8 illustrates the intermediate result after this stage.

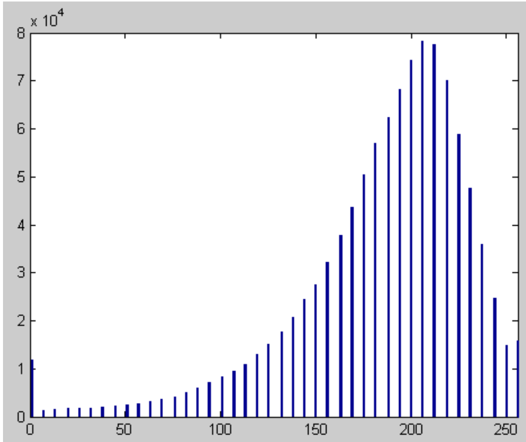


Figure 4.5: Histogram distribution of image in Figure 4.3 after applying contrast stretching

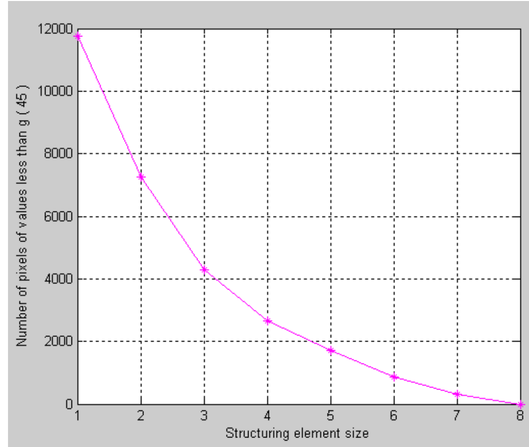


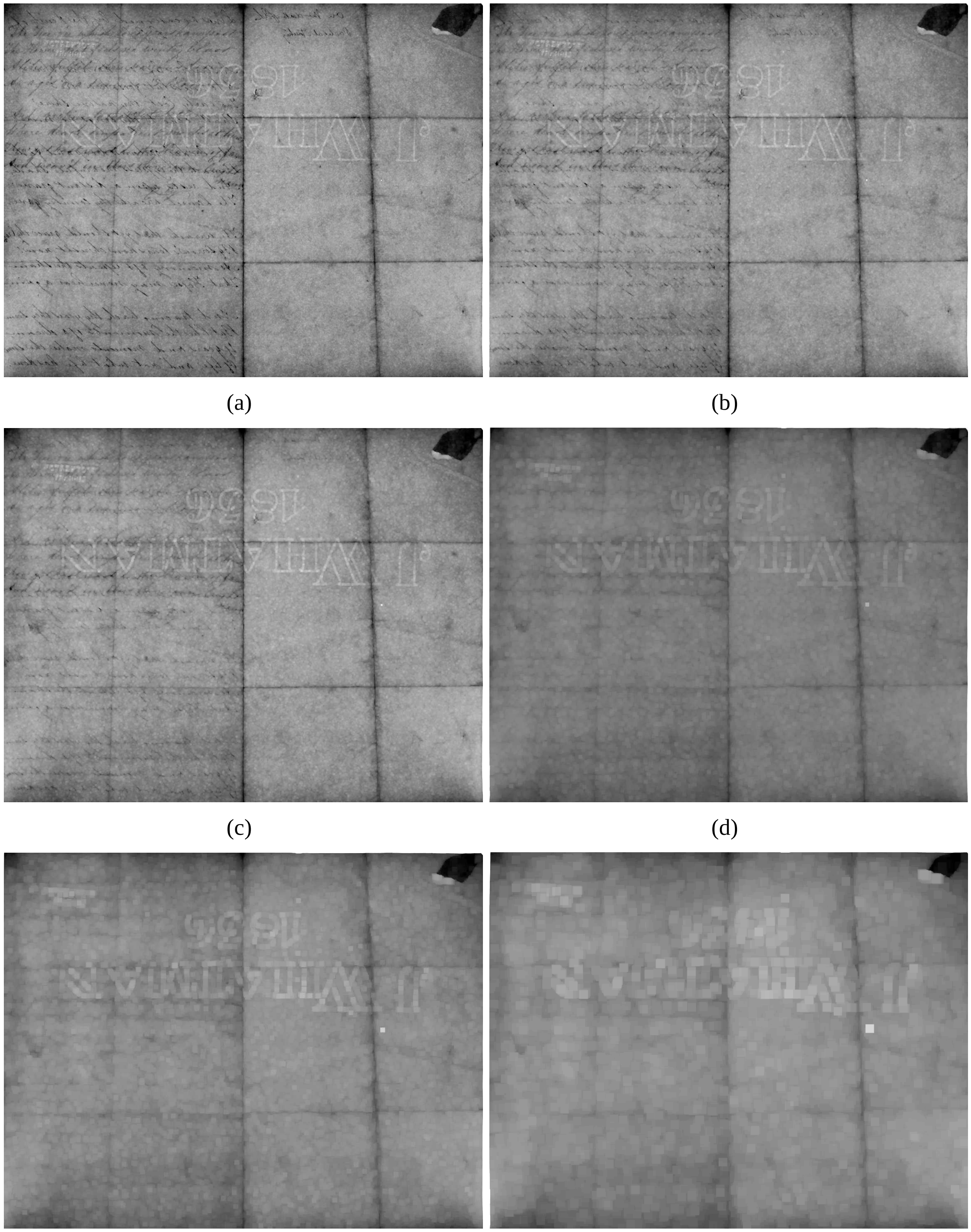
Figure 4.6: Number of pixels of values below g plotted against structuring element size

4.2.1.2 Background estimation

The next step focuses on the removal of non-uniform background. If the image does not have uniform illumination (i.e., some areas are brighter than others), it can be corrected by estimating and removing the background illumination, which is done by applying the morphological top-hat transform, defined as $TopHat = A - (A \circ B)$, where $\circ$ is morphological opening: $C = A \circ B = (A \ominus B) \oplus B$ [63]. A and B are the image and the structuring element respectively. Opening is useful for separating touching features, and removing small regions and sharp peaks.

This transform is applied because the opening operation removes image features that are completely contained in a structuring element. To estimate the image background, it is necessary to remove the watermark pattern by choosing a structuring element with a size that is large enough to cover a single feature of that pattern.

The automatic selection of this optimal size is an interesting challenge for this step, related works can be found in [152, 157]. However, they did not discuss this selection. One of the successful approaches is to estimate the width of the watermark pattern, and choose a structuring element size that is larger than this value; this estimation is now possible, especially after the removal of obstructing foreground features (e.g. writing ink). Granulometry [122] is used to determine the size distributions of features (objects or features: groups of connected pixels) in an image without segmenting each object. This is achieved by applying a series of morphological openings with structuring elements of increasing size. The sum of pixel intensity values in the output image after each opening is stored. See Figure 4.9.



(a) (b) (c) (d) (e) (f)

Figure 4.7: Iterated dilation of Figure 4.3, with structuring element size of (a) 1, (b) 2, (c) 3, (d) 8 (optimal), (e) 9 (the design starts to blur), (f) and 15 (the design is not clear)

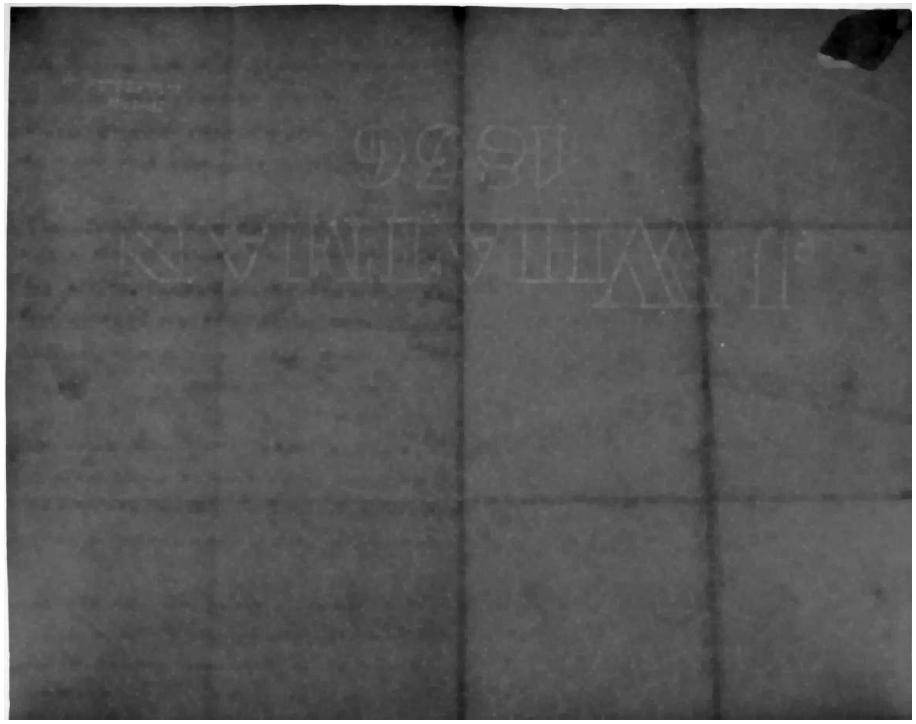


Figure 4.8: Backlit image after foreground removal: watermark is visible, and most foreground interference is removed

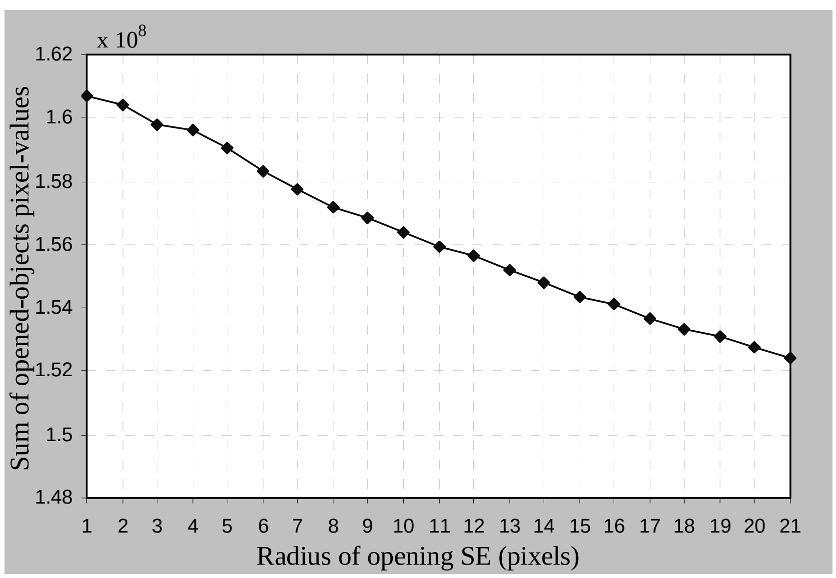


Figure 4.9: Cumulative intensities plotted against structuring element radius; original image in Figure 4.8.

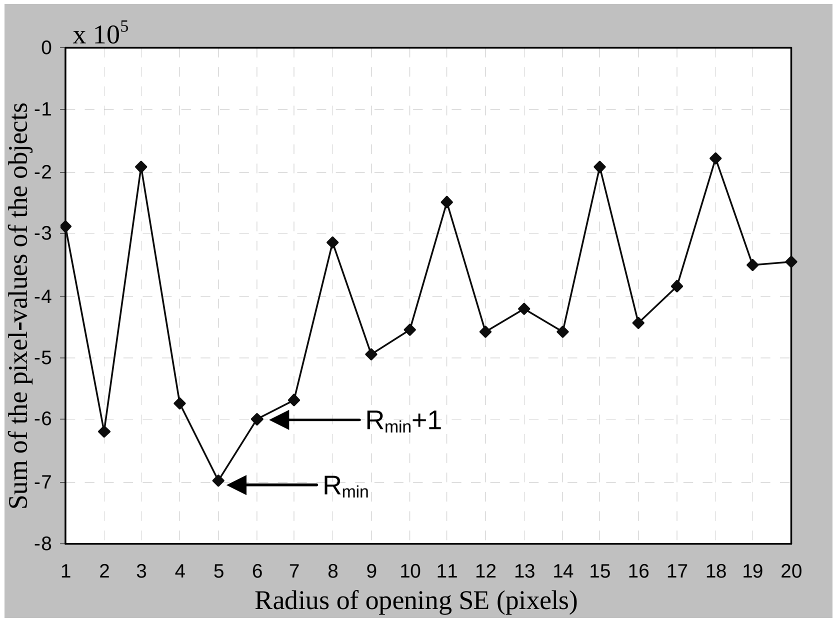


Figure 4.10: Granulometry (size distribution) of image objects: first differences of the plot in Figure 4.9

Taking the difference of total intensities (the sum of pixel intensity values) between two sequential openings will give the distribution of objects sizes at that scale. This definition is also referred to as the pattern spectrum of the image. Figure 4.10 illustrates the granulometry, or pattern spectrum, of image objects, which can be viewed as the first derivative of the intensity surface area distribution.

By investigating this distribution, a local minimum at a specific radius will indicate the existence of many image objects of that radius. The global minimum, R_{min}, will indicate the highest cumulative intensity of objects at that radius. The most suitable structuring element size for background estimation will have the value $R_{min}+1$; choosing a smaller size will not isolate the watermark pattern from the background. Figure 4.11 illustrates the estimated background.

4.2.1.3 Watermark isolation and enhancement

The pre-processing stage is finalised by subtracting the estimated background from the image after foreground removal. The result will have a uniform background; noisy regions such as folding should have been eliminated in this process. The signal for the watermark will then have less interference from foreground noise. However, the intermediate output after the differencing operation is low in contrast due to the numerical subtraction. To correct this, contrast stretching is applied for better visualisation and to

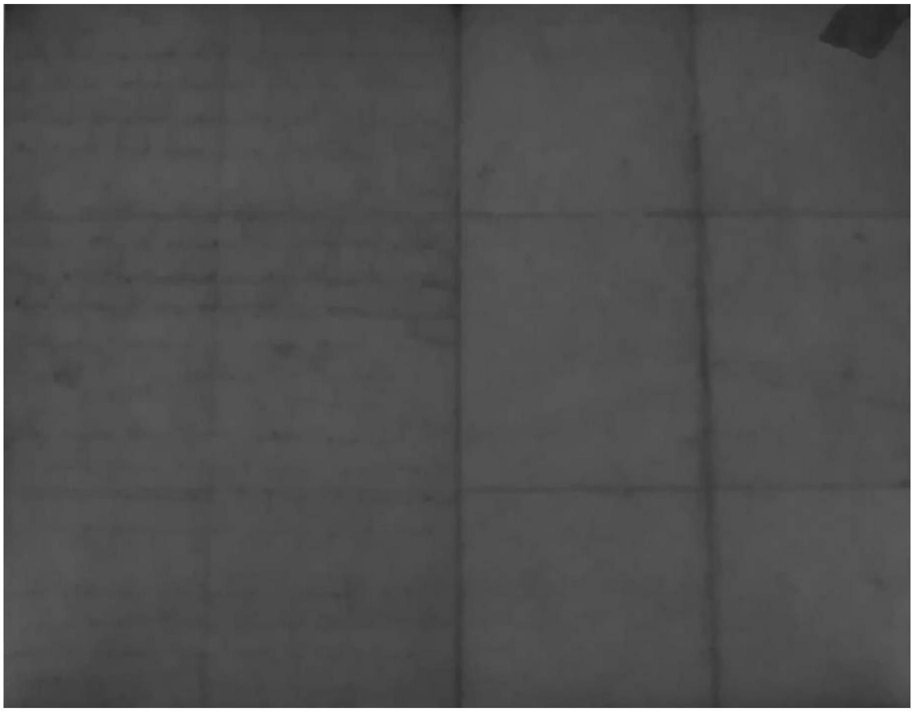


Figure 4.11: Estimated background of input backlit image shown in Figure 4.3

enhance the contrast of the image. See Figure 4.12.

4.2.2 Segmentation

As illustrated in Figure 4.12, the watermark became clear and easy to extract after the pre-processing stage. Its histogram, as illustrated in Figure 4.13 shows this possibility, it only contains 7 grey intensities in this example.

However, there is still some noise from the remaining foreground and background interference: thresholding this intermediate result can be effective to reveal the watermark, but still there is noise, see Figure 4.14(a). Stricter thresholding to remove more noise will affect the watermark signal, see Figure 4.14(b). The following sub-sections will discuss the detection and extraction of chain lines (described in Section 2.2), the location of the watermark area, and the extraction of the watermark pattern through this noise.

4.2.2.1 Chain line detection

As discussed in Section 2.3, chain lines can be very useful for the studies of paper identification: they can serve as fingerprint identification of the mould since such line sequences can be used to identify paper made from the same mould. A specific function of this watermark extraction system has been developed to detect and extract these lines.

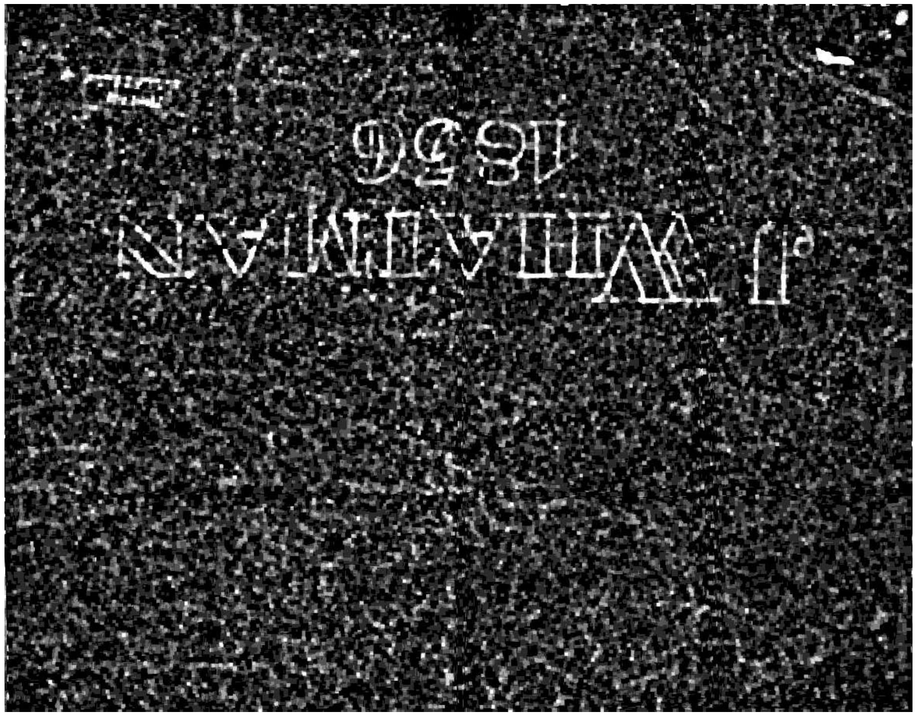


Figure 4.12: Intermediate result after pre-processing stages

The process of detecting chain lines in the image is performed using either the Hough or Radon transforms [91, 122, 136]. This process redraws the detected lines in case some of them do not appear due to the digitisation process, or because of paper folding and cutting. Furthermore, image skew can be also adjusted depending on detected chain lines, in case the paper was misaligned during digitisation.

This detection process can provide us the existence of chain lines, distance between sequential lines, chain line orientation, thickness of lines and the number of chain lines in the paper. The Radon transform computes projections of an image matrix along specified directions by computing line integrals from multiple sources along parallel paths by rotating the source around the centre of the image.

The Radon transform of Figure 4.15(a) is illustrated in Figure 4.15(b); detected lines (high peaks) were located when applying a projection of angle 1^o (equivalent to 181^o).

The detection process locates these lines using a manually selected threshold; detected lines are shown in Figure 4.15(c). This Figure illustrates that the transform detects the two edges of each chain line, and this facilitates the calculation of their thickness and spacing. Measurements are determined by finding the horizontal spacing (in pixels) in this image between sequential lines: small-sized spacings will provide the thickness of such lines, while large-sized spacings will provide the spacing between them.

The direction of the resulting image is then adjusted depending on the direction of

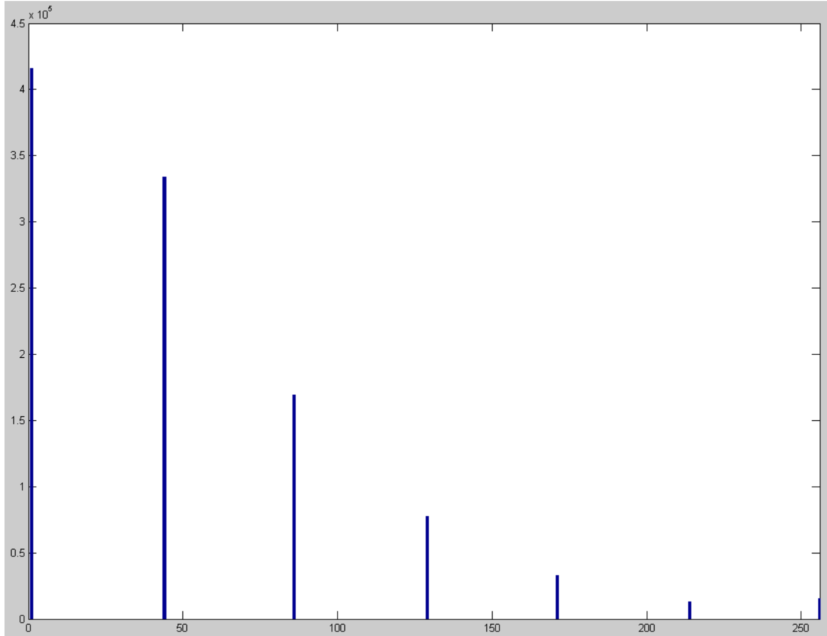


Figure 4.13: Histogram distribution of Figure 4.12

the chain lines; see Figure 4.15(d). This process differs from work presented in [141] as it detects chain lines at any orientation. It also has an advantage over [144] because it detects the thickness of chain lines, and does not need to detect all lines to redraw them.

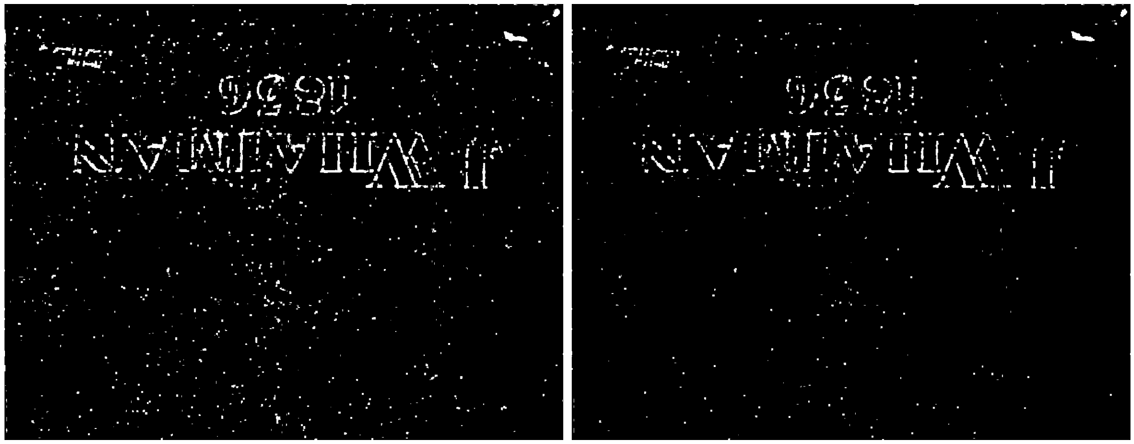
4.2.2.2 Locating the watermark

We are interested in determining automatically the window of the image in which the watermark lies. Despite the significant residual noise, images such as those in Figures 4.12 and 4.15(a) suggest that the signal of the watermark predominates and should be locatable under certain assumptions.

Considering Figure 4.16(a), we have experimented with projections in both x and y directions. The naked eye can detect the location of the watermark, which appears as peaks in x direction in this example. But locating these peaks still needs manual intervention, and it is difficult to locate small patterns, or patterns that are split along paper.

On the other hand, chain line suppression can be helpful in the localisation of the watermark: removing these lines has the advantage of highlighting the watermark area when applying the projection, especially in the y direction, because these lines are vertical and appear as large peaks.

Furthermore, the thresholded images (such as Figure 4.14) seem to demonstrate better signal to noise properties, and we have projected these in a similar manner, as illustrated in Figures 4.17 and 4.18. Visual inspection of the vertical projection easily betrays lo-



(a) (b)

Figure 4.14: Figure 4.12 at 2 thresholds.

cation of the watermark information, but this is less clear in the horizontal projection. Fortunately, deciding which of these directions to adopt (without the naked eye) is solvable by looking into the variance of each projection. By inspecting the projection data in the x direction, we find that the variance is large due to the high values of watermark features compared to other features, while in the y direction, it is low. In this case, we choose the projection where the variance is higher (x direction in this example).

The chosen projection data are then thresholded, using (for example) mean as threshold value – this can give a good localisation of the watermark, without the need for manual intervention.

As a conclusion, automatic watermark locating is possible, assuming that the watermark pixel intensities are high: the pre-processed intermediate result is thresholded, and the chain lines are suppressed. In this case, data projection will be able to reveal the watermark location.

4.2.2.3 Edge detection and noise removal

An alternative approach is to apply edge detection followed by the identification of noise image features and interior segments. A Canny detector [26] is used to locate edges; this method gave the best watermark design detection among a selection of edge detectors such as Sobel, Prewitt, Roberts, and Laplacian of Gaussian [63, 122]. These alternatives provided less shape detail, with more irrelevant image features. See Figure 4.19 for results after detecting edges.

A noise removal process is then applied. Small gaps between image features are eliminated by applying a morphological closing operation which reduces the number of

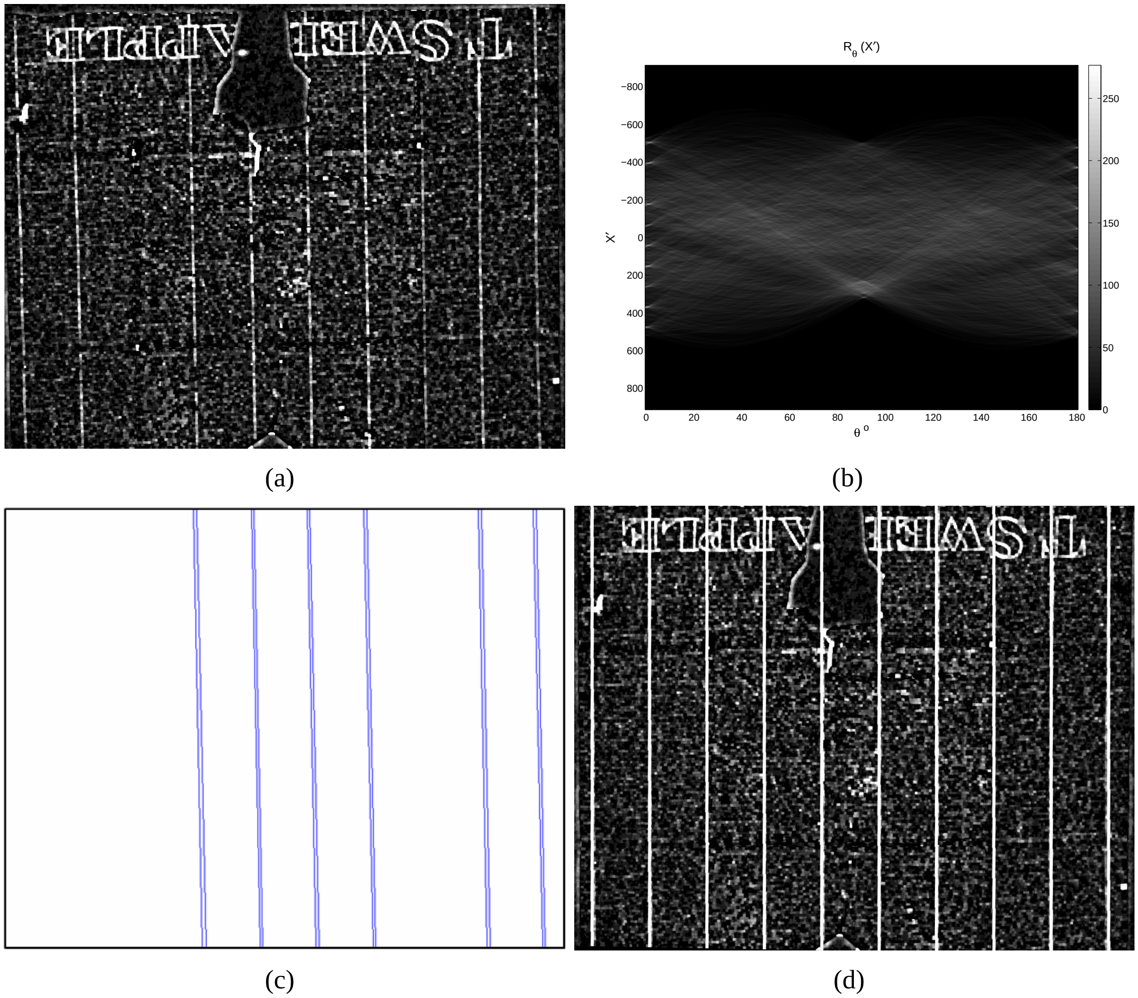


Figure 4.15: (a) Image before chain line detection, (b) Radon transform, (c) Detected lines, (d) Image after chain line detection

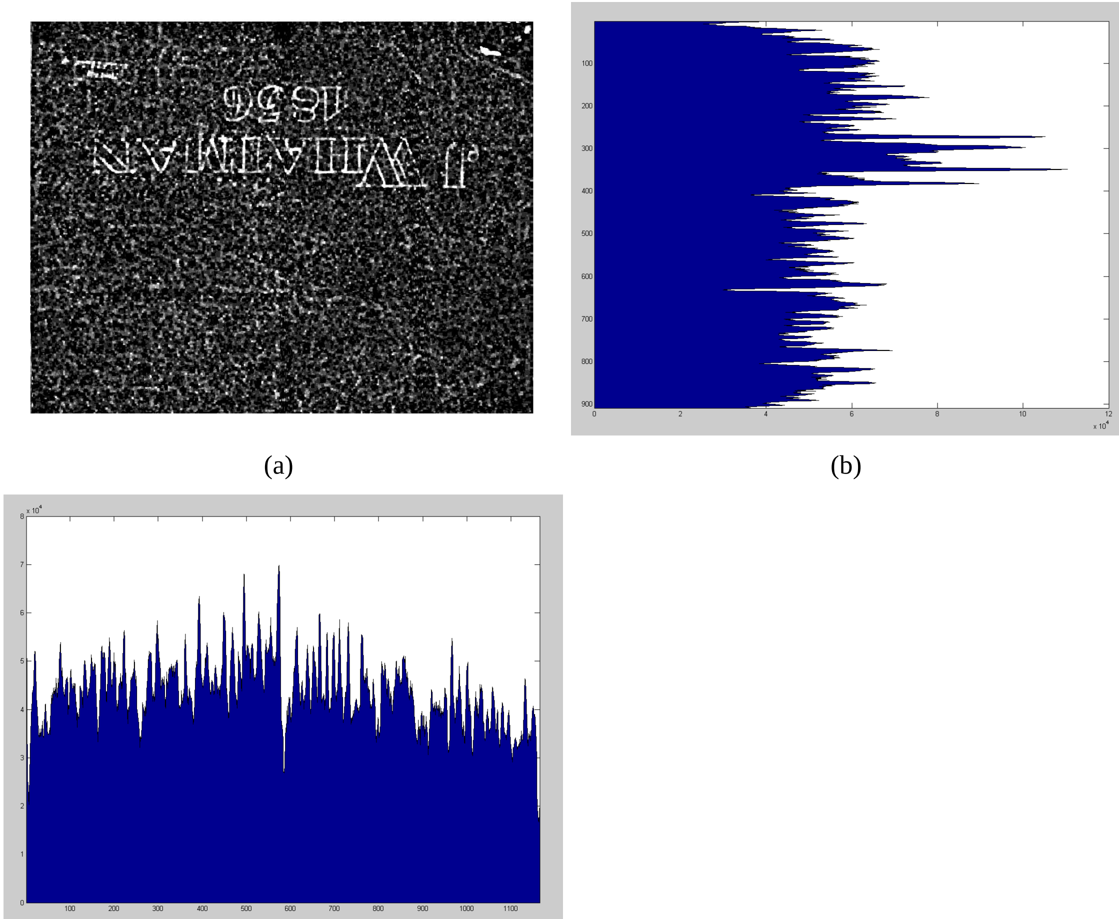
image features (and hence reduces processing time needed), see Figure 4.20.

Image noise is then located and removed. To do this, three assumptions were made: (i) Noisy image features are small-sized; (ii) Noisy image features are isolated; and (iii) Isolated, small groups of neighbouring image features are noise. Hence, three thresholds are used:

- t_1: object size (in pixels). Noise image features (objects) are mostly small, so only objects less than t_1 in size are processed. This speeds the noise removal process.

- t_2: object distance (in pixels). This threshold checks whether an object is isolated from other objects or not; if it is isolated by the given threshold; then it is assumed to be noise and removed.

- t_3: group of objects distance (in pixels). This threshold checks whether a group of



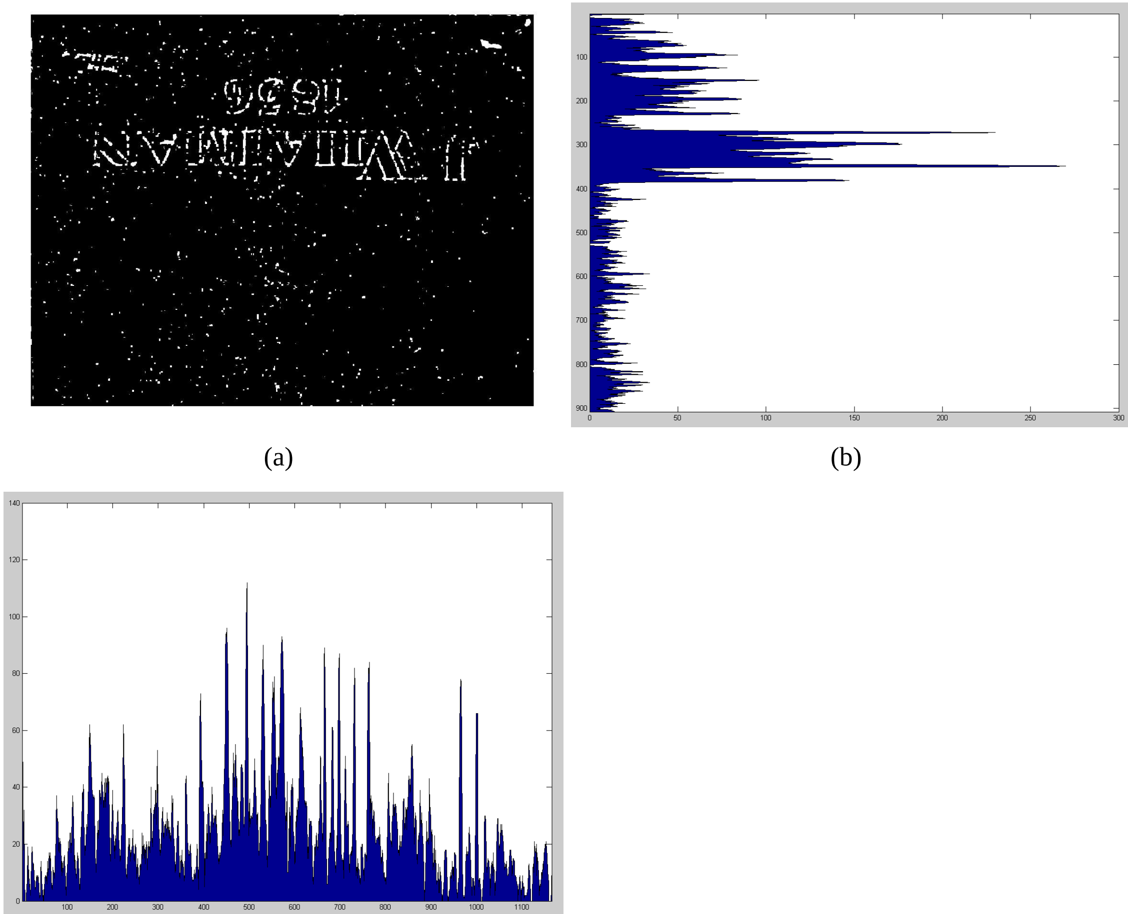
(a) (b) (c)

Figure 4.16: (a) Pre-processed image, (b) Data projection in *x*, and (c) *y* directions

neighbouring objects (objects close to each other) is isolated from other objects by a specified distance. If it is isolated; then it is assumed to be noise and removed.

Values of thresholds can be estimated by viewing the distribution of feature size versus number of objects as in Figure 4.21. These assumptions differ from the assumption used in [152], where they only remove image features of a size (in pixels) smaller than a specific threshold.

The result is then further improved by interior filling of small unwanted holes. The result after these stages is shown in Figure 4.22(a); another result with chain lines present is in Figure 4.22(b); results are rotated for better visualisation of the watermark.



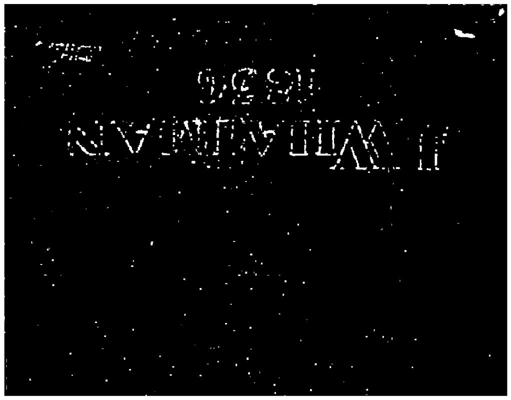
(a) (b) (c)

Figure 4.17: (a) Thresholded image, (b) Data projection in *x*, and (c) *y* directions

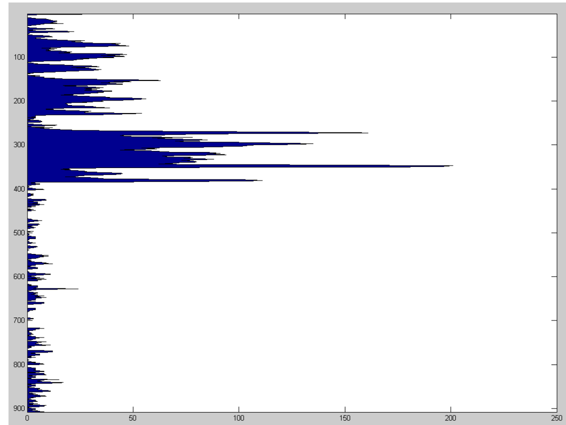
4.3 Results

This section presents several sets of watermark images to demonstrate the results and effectiveness of the approach. The system has been prototyped in MATLAB [134] with a specially designed graphical user interface to provide easy operation, especially for researchers unfamiliar with computer languages and programming, with default settings and the ability to handle manual intervention. The system can also be run in standalone mode, without the MATLAB environment. Results were obtained using an Intel Pentium 4 machine of 2.8GHz speed and 1GB RAM, under the Windows XP operating system.

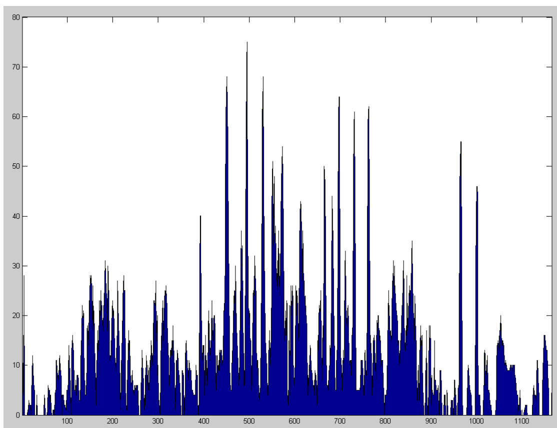
The main interface of the prototype has a window for the rendering of the input image and a set of controls on the right-hand panel. The prototype can be operated a step at a time to trace all the main processing stages. A full illustration of this interface can be found in Figure C.1 in Appendix C.



(a)



(b)



(c)

Figure 4.18: (a) Thresholded image, (b) Data projection in x, and (c) y directions

Figures 4.23, 4.24, 4.25, and 4.26 illustrate a selection of the results obtained with the current prototype. For each sample, we present the key processing stages with the digitised input image and the intermediate and final results. These manuscripts are taken from the works of Henry Litolff [14][1].

Figure 4.23(a) shows an example of a historical watermarked paper sheet with handwriting (ink) on recto and verso, noise and non-uniform background. It is obvious that the watermark and chain lines are brighter than other features in the paper structure – the watermark signal becomes clear in the intermediate result after removal of foreground and background interference as illustrated in Figure 4.23(b). Figure 4.23(c) demonstrates the output watermark pattern (zoomed for better visualisation) with the detected chain lines.

Another example of historical paper with low foreground interference is shown in

[1]Digitised with permission from the Special Collections of the University of Leeds Brotherton Library [123].

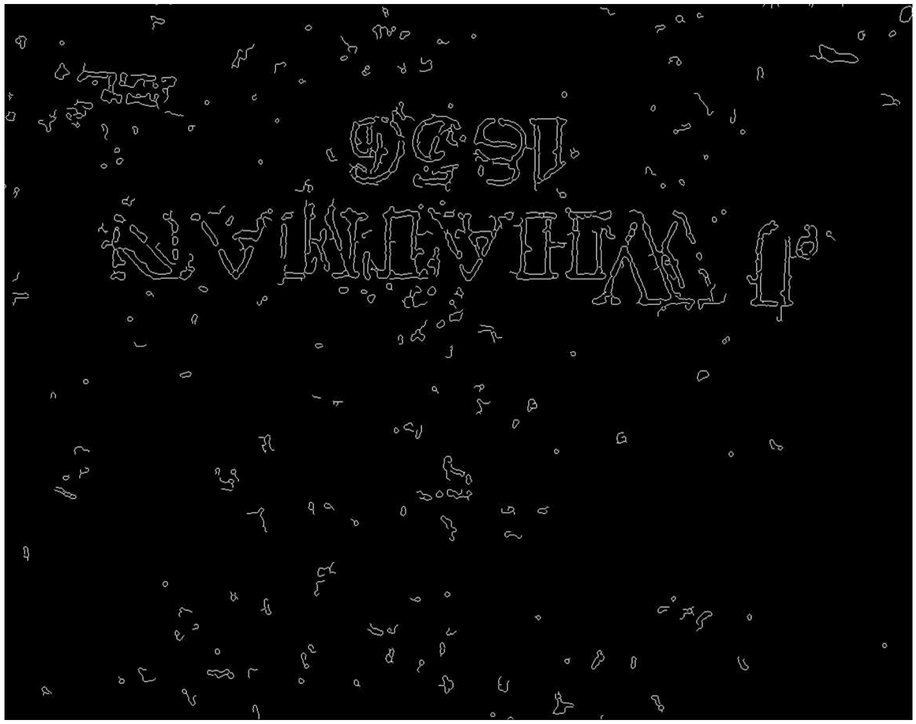


Figure 4.19: Intermediate result after edge detection

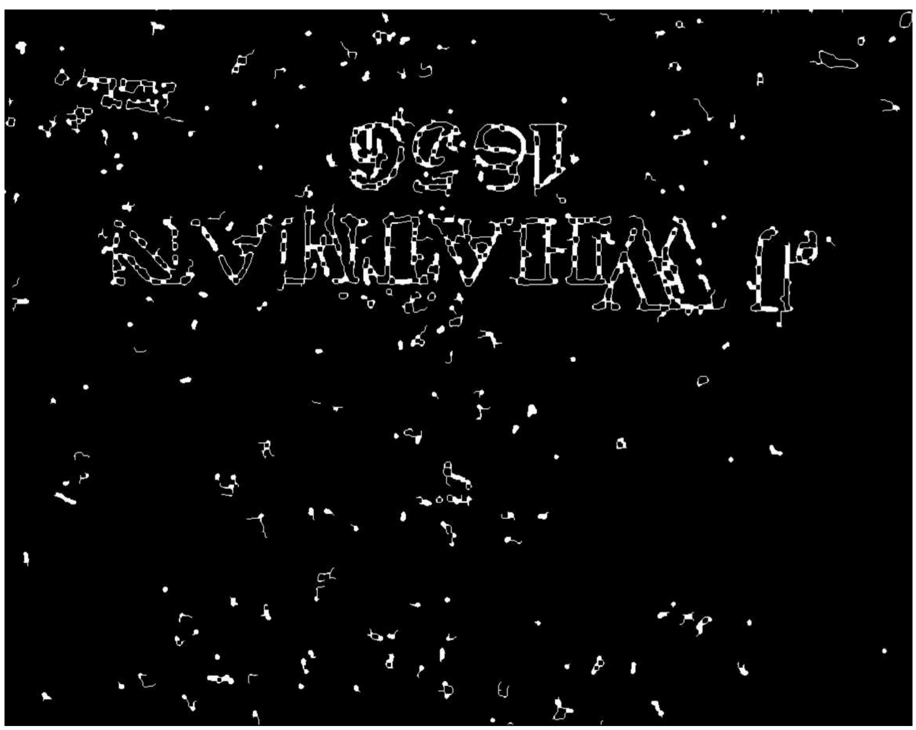


Figure 4.20: Intermediate result after applying morphological closing

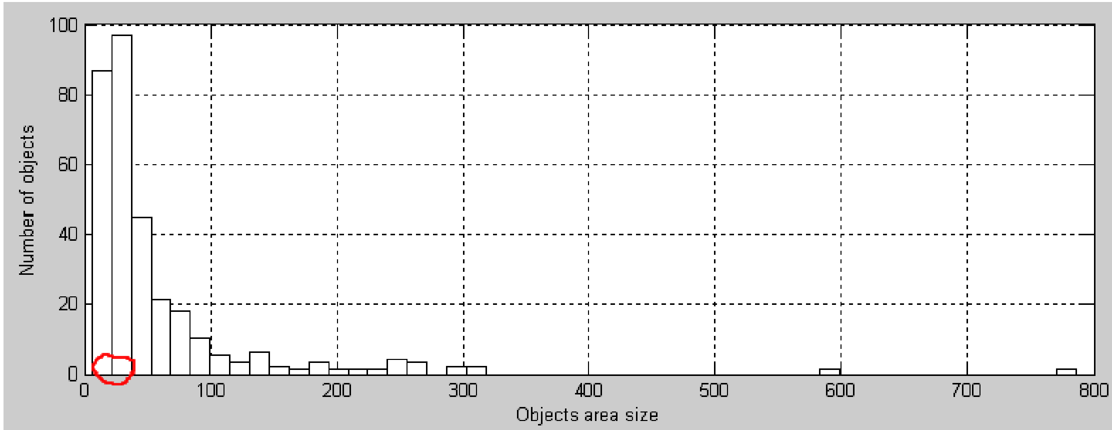


Figure 4.21: Estimation of noise removal thresholds – marked

Figure 4.24(a). The paper has a noisy background which obstructs the watermark design, but this interference was successfully removed after pre-processing as illustrated in Figure 4.24(b). The final output can be found in Figure 4.24(c); the segmentation is clean and contains only the extracted watermark pattern.

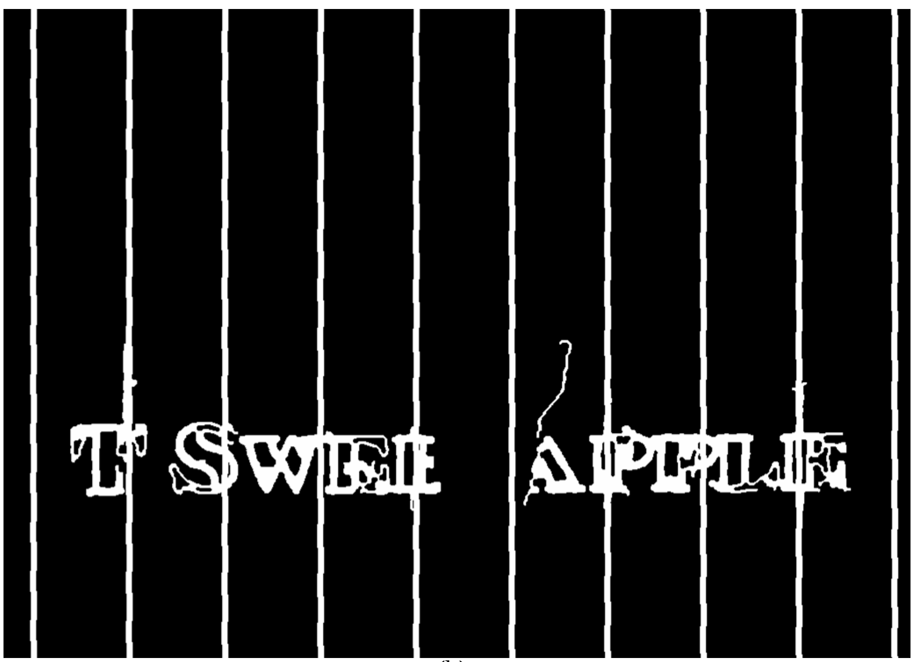
Figure 4.25(a) illustrates another example of historical watermarked paper, with a low watermark signal. This example is a musical manuscript with handwritten music notation, expressive symbols; text and signature, with both foreground and background interference (mainly hand-drawn horizontal stave lines). Figure 4.25(b) demonstrates the intermediate result after interference removal. The final result of the watermark design segmentation is presented in Figure 4.25(c).

An example of contemporary watermarked paper is shown in Figure 4.26(a) (enhanced for display). Here, there is no writing and it has a uniformly textured background. The watermark pattern is partially corrupted by the background pattern and cannot be clearly seen (by eye): hence the quality and completeness of the segmented watermark design is hindered as demonstrated in Figure 4.26(b). Figure 4.26(c) shows the segmented watermark design, and Figure 4.26(d) illustrates a vectorised representation, which is further described in Chapter 6.

4.4 Conclusion

This Chapter presented a prototype to extract paper watermarks in a bottom-up manner. This approach is generally capable of resolving a range of foreground and background interference, using only the transmitted (backlit) image for processing. It also presented the detection of chain lines and the dynamic adaptation of some of the necessary image

(a)



(b)

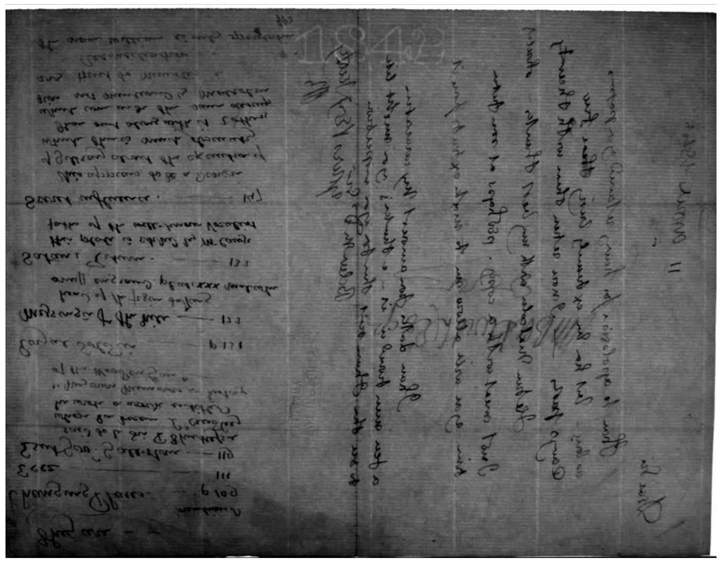
Figure 4.22: Results after segmentation

operations to automatically determine optimal parameter values.

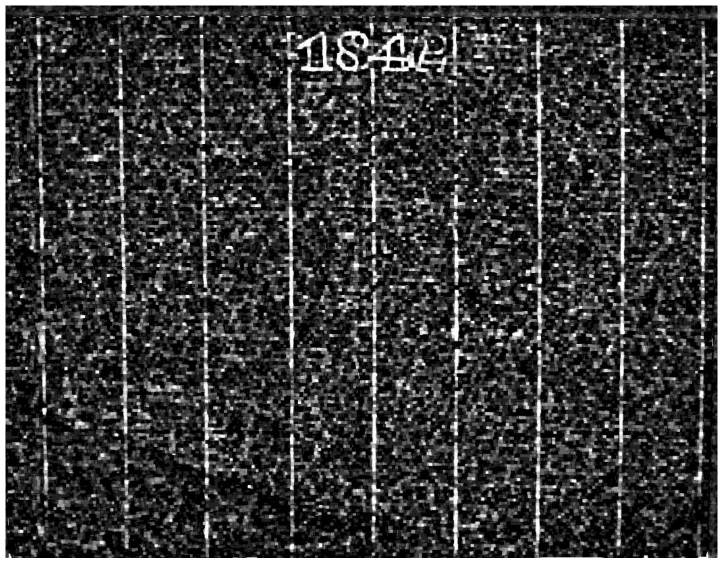
We also presented processing examples, sample results, and discussed applications from different sources, including old and modern watermarked laid and wove paper, and different types of writing, including graphical notation.

However, this approach is limited to the kind of data presented in Sections 3.1.1 and 3.1.2. These data are characterised by non-uniform background and thin pen stroke used in writing (i.e., radius of the nib). Clearly, any large region of dark interference cannot be supported. Datasets used are thin paper, with the watermark design clearly visible.

The morphological and edge detection algorithms are sensitive to parameters choices. We presented a number of algorithms to determine optimal structuring element sizes in dilation and opening operations, but other processes of this approach (e.g. edge detection) need manual parameter adjustment.



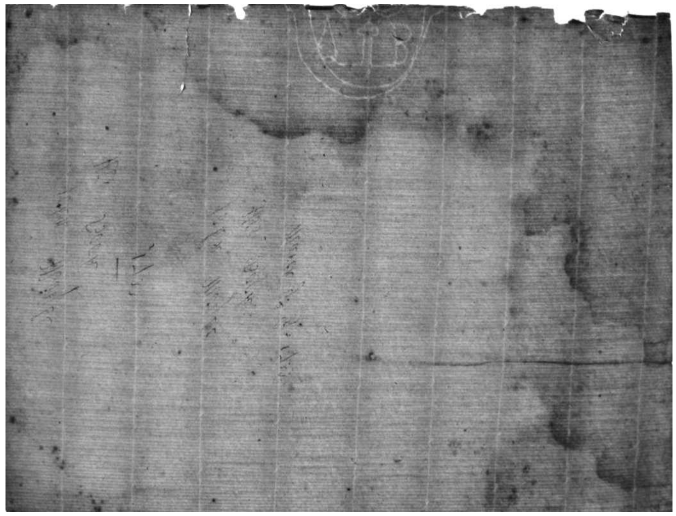
(a)



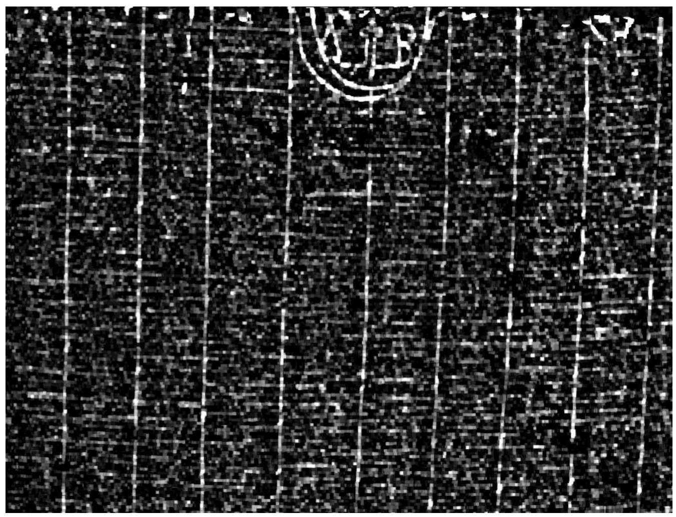
(b)

(c)

Figure 4.23: Sample input 1 with handwritten watermarked paper (a) input source image digitised with back-lighting, (b) pre- processed intermediate output, (c) segmented watermark design (zoomed)



(a)

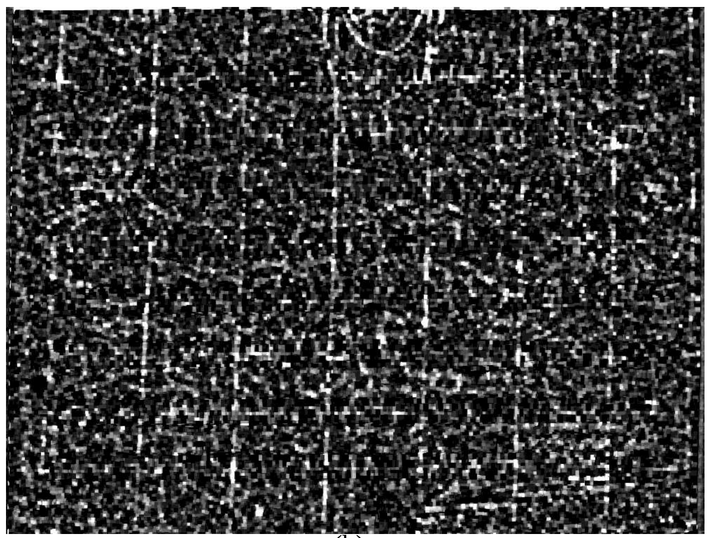


(b)

(c)

Figure 4.24: Sample input 2 with low foreground Interference (a) input source image digitised with back-lighting (b) pre-processed intermediate output, (c) segmented watermark design (zoomed)

(a)



(b)

(c)

Figure 4.25: Sample input 3 with handwritten music manuscript (a) input source image digitised with back-lighting, (b) pre- processed intermediate output, (c) segmented watermark design (zoomed)

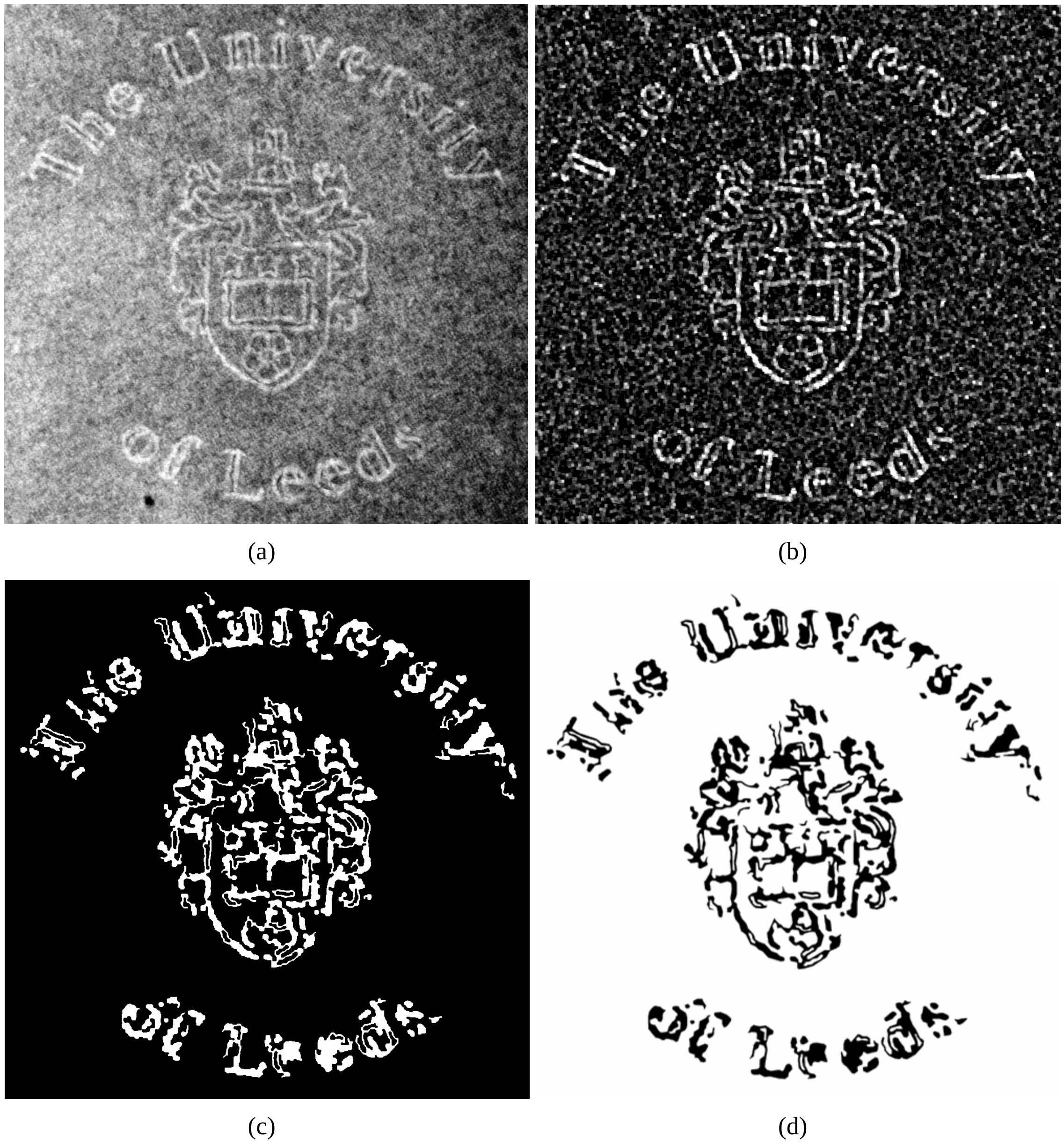


Figure 4.26: Sample input 4 with currently available watermarked paper (a) input source image digitised with back-lighting (enhanced for display), (b) pre-processed intermediate output, (c) segmented watermark design, (d) and its vectorised representation

Chapter 5

Watermark location via modelling back-lighting

5.1 Introduction

Chapter 4 presented a bottom-up approach which successfully locates different kinds of watermarks as presented in Section 3.1.2. These data are characterised by non-uniform background and thin pen strokes; the paper used in these data is thin and uniform, and the watermark design appears clearly. This results in low foreground interference and a strong watermark signal.

We now turn to the more challenging data presented in Sections 3.1.3, 3.1.4 and 3.1.5. These are complete handwritten collections of Islamic text: these data, especially the 'Mahdiyya' copy of the Qur'ān, are characterised by thick writing strokes on recto and verso, and the paper used in writing this manuscript is thick, and the watermark patterns are not clear. In summary, there is significant foreground interference, and a weak watermark signal. Hence the data is more difficult to process. However, it is important to support these artefacts due to their irreplaceable value[1].

This Chapter demonstrates the limitations of the bottom-up approaches in their application; this is no surprise. We proceed to introduce a top-down approach which has success with the more challenging data, and may well be more widely applicable. Our

[1]We have selected historical texts from the University of Leeds collection nominated for interest by a senior Arabic scholar [76]

approach attempts recto removal, followed by highlighting of watermark ‘hidden’ data. We also present a statistical approach to the location of watermarks from a known lexicon.

Throughout this Chapter, we will refer to images as upper case roman, I, and to pixels of images as lower case p: these will usually be multidimensional, and usually RGB.

5.2 Limitations of the bottom-up approach

We have deployed the algorithms of Chapter 4 to some of the Qur’$\bar{\text{a}}$n data (see page 160 for the original data). Figure 5.1 presents a representative sample of the result. Here, we can see that foreground (recto and verso writing) and background (paper textural features) still exist, and the watermark signal is very weak so it cannot be separated from surrounding interference.

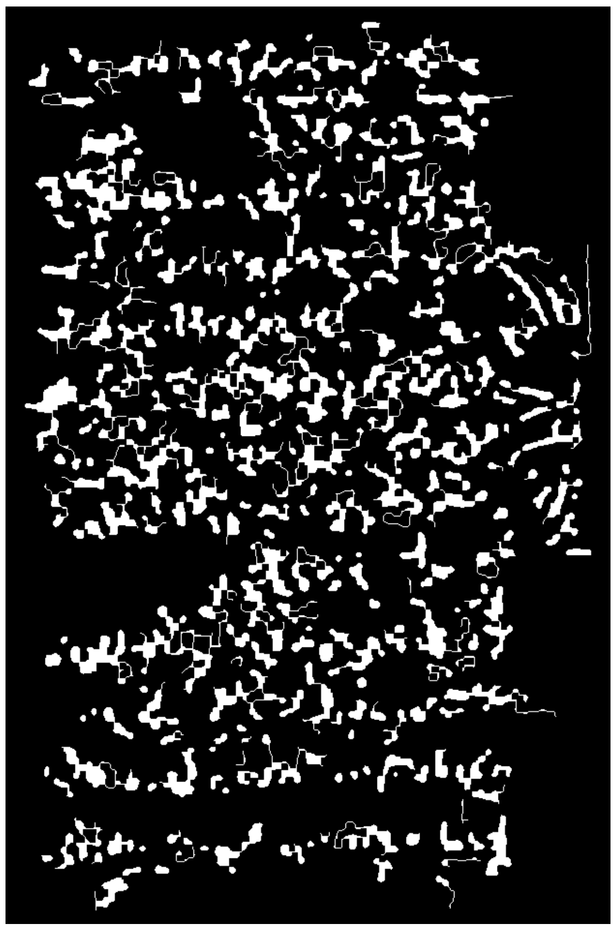


Figure 5.1: Result of applying bottom-up approach to the backlit image shown on page 160. A part of a double-headed eagle watermark is detectable by the eye at the centre of the right-side edge of that page.

This example illustrates typical limitations of the bottom-up approach that failed to extract the watermark pattern in these data. This is due mainly to the weak watermark signals.

5.3 Recto removal

5.3.1 A model of back-lighting

In this application, we are presented for each page with a recto scan, and a co-registered backlit scan. Figure 5.2 shows just part of an example page which illustrates well the range of problems – part of an existing watermark (fully illustrated in Figure C.8 in Appendix C) is visible to the eye, as is the range of other information the images contain. The non-uniformity of the paper surface is characteristic, and many pages suffer from damage of further kinds.

Figure 5.2: Left: part of a scanned recto; Right: corresponding backlit image – the watermark can be seen faintly at the right. These data are taken from the ‘Mahdiyya’ copy of the Qur’ān presented in Section 3.1.3.

To proceed, we assume a model of the effect of back-lighting that is illustrated in simplified form in Figure 5.3. The RGB vector detected at a particular pixel is dependent on the paper properties (absence or presence of watermark or other manufactured feature), recto features and verso features. In an ideal world, blank unfeatured paper (labelled ‘A’ in the Figure) would always produce the same output, but we do not have to assume that the same is true of inked regions (e.g., ‘B’), paper features, or combinations thereof.

For clarity, we shall define at this point a feature to be *visible* if it is visible on the recto – thus, recto writing and other paper features visible to the reader. Other features betrayed in the backlit image (watermark, verso writing, dirt on the verso face etc.) we

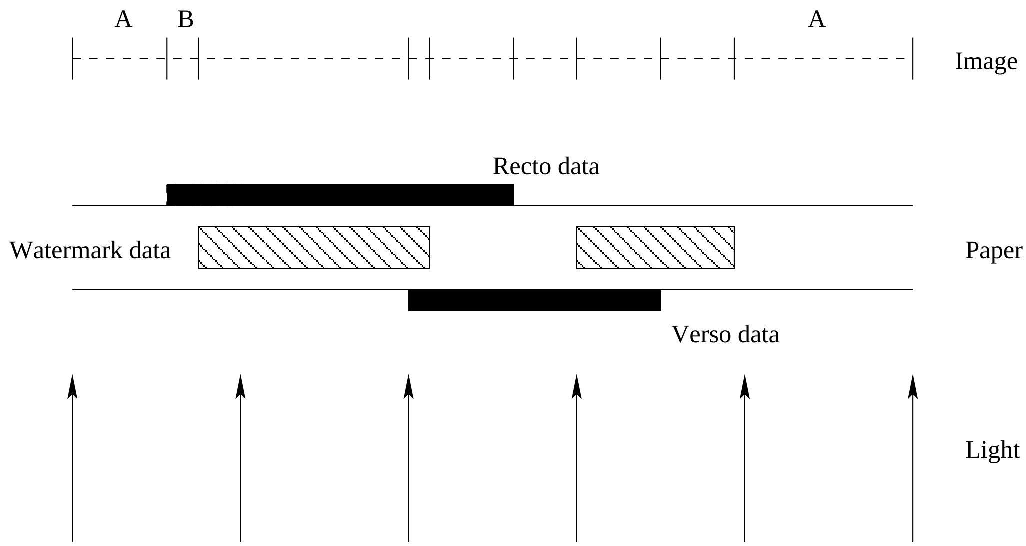


Figure 5.3: The model of back-lighting. The paper is lit from below (up-arrows) and the image (dotted line) sensed above; data may be received from blank paper, or some combination of recto, verso, or 'interior' features. The vertical lines along the image indicate points at which the received signal *may* change: at 'A', we are detecting blank unfeatured paper, at 'B' recto data inscribed on it. Of course, recto and verso inscriptions need not be uniform, nor need watermark features, and there may be many other influences as well, including dirt and noise.

shall collectively call *hidden*. Backlit pixels at which no hidden data are evident we shall call *uncorrupted*.

In fact, the noise and damage that we experience produces significant variations across all regions that we might wish to be internally homogeneous, as is clear from Figure 5.2. This however is not critical – what we can exploit is the difference between pixels that represent just blank paper or recto features, and those representing verso or other features, such as internal ones.

5.3.2 The trivial case: null recto

Consider momentarily a blank, unfeatured page which we scan as image S and backlight as image B, and define an image D in which pixels are given by the difference between their detected backlit intensity (in B), and the intensity we might *expect* given the corresponding location in S. In the ideal case this page will be of uniform intensity (r,g,b) in S and, say, (ρ,γ,β) in B. We hypothesise some transform T which describes the back-lighting, and subtract $T(r,g,b)$ from the corresponding (ρ,γ,β) in B. We should see (0,0,0) at all locations. If there are paper or verso features (invisible in S), these will be revealed by this differencing process.

In fact, of course, regions are not uniform in intensity and blank paper will scan and back-light as a range of (r,g,b), (ρ,γ,β) vectors – these may, however, be expected to cluster reasonably tightly, and to be related to each other. If we define

$$\begin{aligned}(\mu_r,\mu_g,\mu_b) &= mean(r_p,g_p,b_p) : p\varepsilon S \\ (\mu_\rho,\mu_\gamma,\mu_\beta) &= mean(\rho_p,\gamma_p,\beta_p) : p\varepsilon B \end{aligned} \tag{5.1}$$

then a simple approach is to seek a linear relationship

$$(\rho_p,\gamma_p,\beta_p) \approx A((r_p,g_p,b_p)-(\mu_r,\mu_g,\mu_b)) \; + \; (\mu_\rho,\mu_\gamma,\mu_\beta) \tag{5.2}$$

for some 3×3 matrix A that models the back-lighting. Lighting effects are often subtle and it is most unlikely that the effect we observe will indeed be linear, but we proceed with this simplification on the understanding that it is applied only to pixels that are ‘similar’, and in the ideal case identical.

In the event that there are no internal or verso features, we can derive an optimal A by considering Equation 5.2 for all pixels p as an over-determined system and ‘inverting’[2]

$$A = [(\rho_p,\gamma_p,\beta_p)-(\mu_\rho,\mu_\gamma,\mu_\beta)][(r_p,g_p,b_p)-(\mu_r,\mu_g,\mu_b)]^{-1} \tag{5.3}$$

Then, for the simple case of a blank page,

$$D = (\rho_p,\gamma_p,\beta_p) - A((r_p,g_p,b_p)-(\mu_r,\mu_g,\mu_b)) - (\mu_\rho,\mu_\gamma,\mu_\beta) \tag{5.4}$$

and we will expect significant differences from $(0,0,0)$ to betray hidden information.

This procedure is illustrated in a trivial case in Figure 5.4 which shows S, B and D for a blank page with a simple verso inscription, and Figure 5.5 which illustrates a watermark extracted by the same process. In these figures, ‘intensities’ (which may be negative) have been linearly mapped to the range $[0,255]$.

In the event that we expect the image to contain hidden features, this approach lends itself to an immediate improvement. Assuming that there exist uncorrupted features in B and the relative size of watermark features is small, we shall expect the watermark to exhibit a high magnitude response in D, and the uncorrupted areas to be low (ideally 0). Therefore, we may recompute A by reducing the set of pixels from which it is derived to

[2] A linear algebraic operation straightforwardly available in libraries provided by, e.g., MATLAB [134].

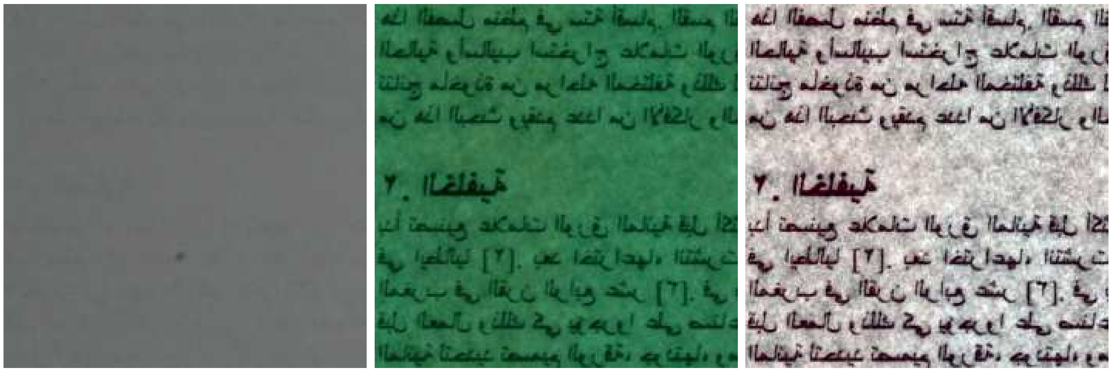


Figure 5.4: Scanned, backlit and differenced images (left to right) – the verso is clearly revealed. The difference has been contrast stretched for display.

Figure 5.5: Scanned, backlit and differenced images (left to right) – the watermark is clearly revealed. The difference has been contrast stretched for display. This image is a part of the full illustrated paper shown in Figure B.1 in Appendix B. This document is taken from the works of Henry Litolff [14], digitised with permission from the Special Collections at the Brotherton Library of the University of Leeds [123].

those we expect to be featureless; thus, Equation 5.3 may be re-employed;

$$\begin{aligned} \hat{D} &= \{p : |D_p| < T\} \\ A_{new} &= [(\rho_p, \gamma_p, \beta_p) - (\mu_\rho, \mu_\gamma, \mu_\beta)][(r_p, g_p, b_p) - (\mu_r, \mu_g, \mu_b)]^{-1}\ ,\ p \varepsilon \hat{D} \end{aligned} \quad (5.5)$$

where $|D_p|$ is a measure of the magnitude of the difference vector at p – Euclidean length is an obvious choice. Choices for the threshold T are discussed in section 5.5.2. This procedure is open, of course, to iteration in attempting only to compute A from pixels which are uncorrupted.

5.3.3 The general case: paper with recto inscription

We shall expect most scans to carry recto material and so the preceding assumptions about a 'blank piece of paper' are invalid. Nevertheless, the approach is sound if we can apply it to pixels of S that are similar in intensity. This is straightforwardly achieved by clustering the data of S in RGB space, and deriving a matrix A for each such cluster. Formally;

1. Using K-means [122] or similar, cluster the RGB data of S into a partition of K_1 clusters $C_1, C_2, \ldots, C_{K_1}$. These clusters may have spatial coherence, and may not.

2. For each cluster C_i derive a matrix A_i according to Equation 5.3, where p is restricted to C_i (not the whole image).

(The iterative refinement approach of Equation 5.5 is applicable to each such cluster).

At this point we do not discuss a suitable value for K_1. Choice of the 'optimal' number of clusters is a widely considered problem [47,114], and usually it is desirable to minimise K_1, thereby leading to a more compact data encoding. Here, the problem is somewhat different: the more clusters we define, the better the subtraction process is likely to perform, provided the matrices A_i are approximating uncorrupted pixels, and the model of Equation 5.5 is not that of hidden, or verso features. This issue is considered further in Section 5.5.2.

5.4 Watermark location

The foregoing procedure shows good success at erasing recto features – Section 5.5 provides some illustration of this. In pursuit of specific features we might now make some further assumptions: in particular, we might (usually) expect verso inscription to be dark relative to paper and so the components of relevant pixels in D to be negative: setting such components to 0 will have a beneficial effect on enhancing the signal due to, e.g., watermarks.

Nevertheless, the nature of data with which we are dealing is still extremely difficult. In Chapter 4 we have extracted watermarks without prior knowledge of their pattern, but this is, at this stage, ambitious. We simplify the next stage by assuming we know a set of possible or likely watermarks, and seeking their occurrence. This is not unreasonable as a task;

- For a given document, foreknowledge may well provide a set of plausible paper manufacturers and dates.

- Since a precise (or indeed complete) representation of the watermark is not necessary in what follows, an interactive phase may invite a user to outline candidates roughly in a small number of trial pages.
- Watermarks often occur as *near*-identical twins [126]: our approach will find such twins and allow a later refinement to determine which of the pair is actually seen.

Figure 5.6: An example 'difference' image; On the left, a version contrast stretched for display; on the right the same image colour coded according to the cluster that the pixel belongs to in S.

The output of the differencing phase contains very significant noise in addition to information of value; Figure 5.6 illustrates an example from our dataset. The presence of watermark fragments of value is clear, as is the spatial distribution of data as a result of the clustering in Section 5.3. In particular, the information of interest is not among the strongest responses, and simple thresholding approaches are unlikely to assist. On the other hand, pixels of the watermark are similar in RGB intensity, and to exploit this we re-cluster the D image.

Using K-means again, we now generate K_2 binary images $D_1, D_2, \ldots D_{K_2}$ by partitioning D – Figure 5.7 illustrates some of these for the example of Figure 5.6. Suitable values for K_2 are considered in Section 5.5.3. It will be clear that some of these images will contain binary patterns that are good representations of fragments of the watermark (in particular, the 'background' will), while others may not. We proceed by selecting informative fragments of the watermark and seeking a binary match in each of these partitions of D. Figure 5.8 illustrates two such fragments from the watermark of Figure C.8 in Appendix C.

'Matching' here is a binary templating task which is misleading to approach in the customary cross-correlation manner. Instead, we proceed for a given template (watermark

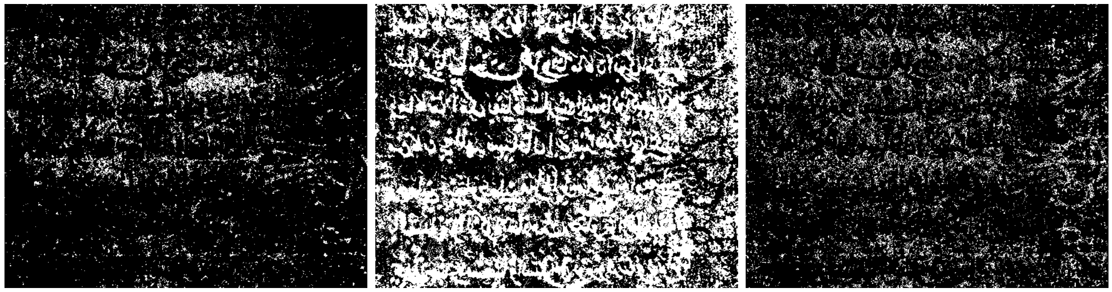


Figure 5.7: Three clusters derived from the difference image shown in Figure 5.6. Note that these clusters contain valuable information of the watermark design.

Figure 5.8: Two fragments of the double-headed watermark shown in Figure C.8 in Appendix C.

fragment) W_i by assuming it contains N pixels, of which w_i are 1's (implicitly, $N - w_i$ are 0's). Now when the template is offered at a particular offset in the image D_j, we count the number of pixels that match (both 1's or both 0's) and interpret this 'score' in the light of what may be expected in noise. If at this offset in D_j there are d 1's within the bounding box of the template, and these are chosen randomly, we have an instance of sampling without replacement to which the hyper-geometric distribution is applicable [94]. If at template offset p we write

$$u(p) = \{\text{No. pixels at which both template and image are 1, or both 0}\} \ ,$$

then (see Appendix A)

$$\begin{aligned} \mu(u(p)) &= N + 2\frac{w_i d}{N} - (w_i + d) \\ \sigma^2(u(p)) &= \frac{4 w_i d (N - w_i)(N - d)}{N^2 (N-1)} \end{aligned} \tag{5.6}$$

(both mean and variance clearly depend on the properties of the template fragment and the position in the image).

Now in seeking plausible locations for the fragment, we are interested in significant deviations from the mean we might expect to see in noise $\mu(u)$, where significance might be measured with respect to the standard deviation $\sigma(u)$. Thus at pixel position p in image D_j we will compute

$$m(p) \;=\; \frac{u(p)-\mu(u(p))}{\sigma(u(p))} \tag{5.7}$$

Herein, high positive responses will represent plausible match positions unless D_i is the background, in which case we would seek strong negative responses (since the template will be inverted). An example result $M_i = m(p)$ is illustrated in Figure 5.9.

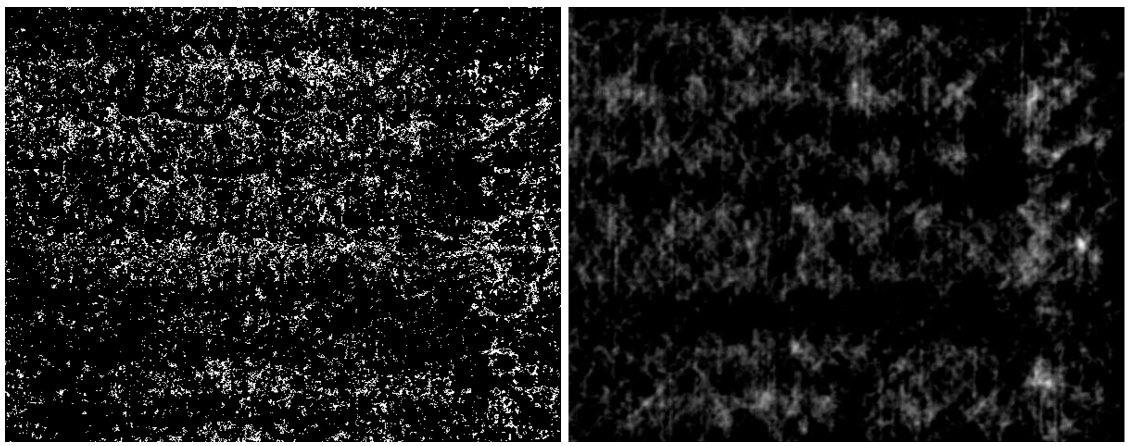


Figure 5.9: On the left an image D_i in which the watermark fragment shown in Figure 5.8 (left) is sought. On the right, the response M_i, given in Equation 5.7.

At this stage we can straightforwardly accumulate the M_i;

$$M = \sum_{i=1}^{K_2} M_i \tag{5.8}$$

Significant peaks in this array will now represent evidence for the fragment in the original image; how we interpret 'significant' here is considered in Section 5.5.3

In fact, we have valuable additional evidence from second, or further, fragments of the watermark: applying this procedure for each such fragment we can exploit their known geometric relationship in inspecting peaks in the M array, these relations are explained in Section 5.5.3.

5.5 Results and discussion

5.5.1 Introduction

We have tested this approach with data presented in Chapter 3, concentrating on samples from the 'Mahdiyya' copy of the Qur'ān of 346 pages, since itis the most challenging data among other manuscripts we have. The following sections will give example results of our approach, together with discussions and considerations of parameter selections used.

An evaluative measure is of use in judging levels of success, and we have chosen to use the signal-to-noise ratio (SNR) [122] of known data in a small number of samples.

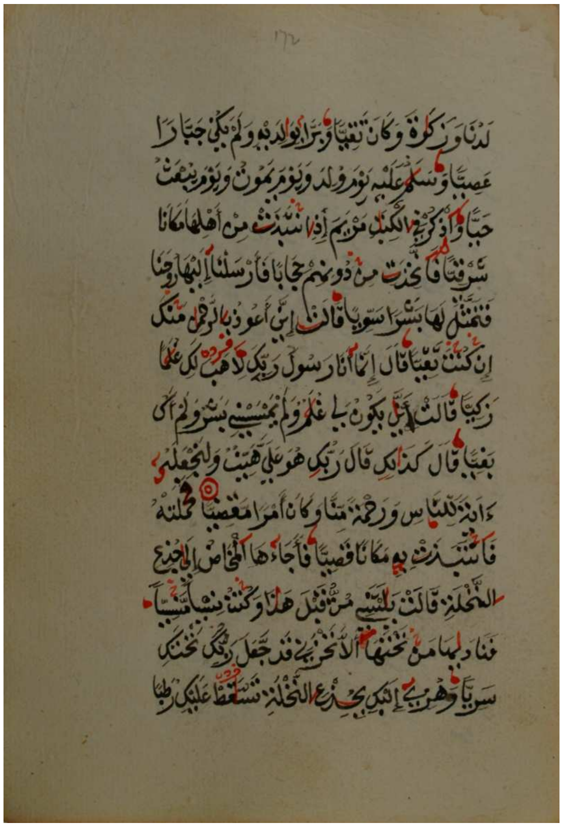
Supposing a watermark and its position to be known, we can split the image pixels into two groups: watermark features W, and all others which we regard as noise N. Then SNR may be calculated as

$$SNR = \frac{\sum_{i \varepsilon W} x_i^2}{\sum_{j \varepsilon N} x_j^2} \tag{5.9}$$

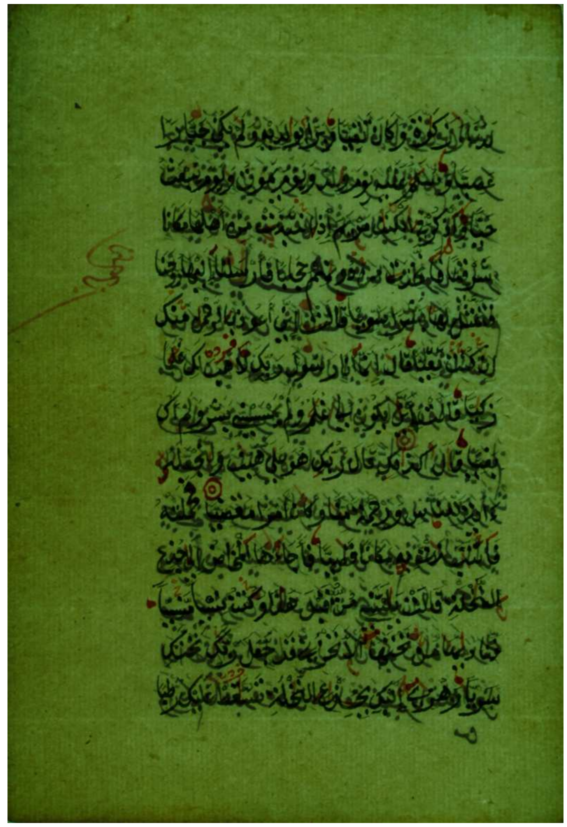
Here, x denotes the mean RGB value of each pixel. In all the experiments of measuring SNR, the known watermark features W are located in the image, and the square values are calculated for each of W and N to find the SNR. Note that the watermark is considered here to be a binary feature, and the calculation is performed with respect to the entire image. This is based on the fact that all the watermarks considered in this thesis are wire watermarks: in the alternative case of shadow (light and shade) watermarks, each pixel could be labelled with a non-binary representation, but we have not explored this here.

SNR may be measured over the whole image or a smaller window for the part that contains the watermark signal only. In the latter case, the SNR values will be higher, since there will be less corrupting noise. Either 'windowed' or 'whole' image SNR measures can be used in our experiments. We have chosen to use the latter measure, because it provides a measure of noise over the whole image. To illustrate, our experiments try to remove the recto features, the process of recomputing transform A improves the whole image SNR by merely removing further recto features. The 'whole' SNR approach helps making these effects obvious.

Figure 5.10 shows full illustrations of input scanned (reflected) and backlit (transmitted) sample images taken from the 'Mahdiyya' copy of the Qur'ān. This sample was chosen to clearly illustrate the high interference caused by recto and verso writing, and to show the difficulty of observing the watermark due to its low signal.



(a)



(b)

Figure 5.10: Full illustration of an input scanned and backlit images

5.5.2 Recto removal

As discussed in Section 5.3, we compute a transform matrix A that approximates the intensity effect of back-lighting; this is then used to remove all recto information in a differencing operation. Using the simple computation of A (Equation 5.3), Figure 5.11(a) illustrates the distribution of differences for a sample image pair: the differenced image is RGB, and we computed here the average of the RGB channels. We might expect high differences to correspond to hidden, bright features in the backlit image B (region X on the horizontal axis), and small differences (region Y) to be due to uncorrupted pixels. Dark features in B, such as verso writing, will manifest as negative differences (region Z).

This histogram shows the distribution of verso, uncorrupted, and watermark features. This distribution is non-symmetric, with verso features appearing prominently as negative; low magnitude pixels are modal, suggesting that the transform was good enough to model the back-lighting. High magnitude pixels in this distribution are relatively small in number, and represent the watermark and other hidden features.

Adopting the approach outlined in Section 5.3, we have refined the matrix A by iteratively recomputing the pixels from which it is derived. We have selected these pixels as those between the means of positive and negative observations in the differences (m_1, m_2). This is a simple way of trying to restrict the computation to uncorrupted areas of the image in the light of the distribution being non-symmetric. Figure 5.11(b) illustrates the distribution after this iteration has been conducted; observe that region Y in this new distribution is narrowed, while regions X and Y (which hold verso and hidden features) were pushed to right and left directions respectively. This improvement increased the effect of minimising recto interference, and enhanced the watermark feature.

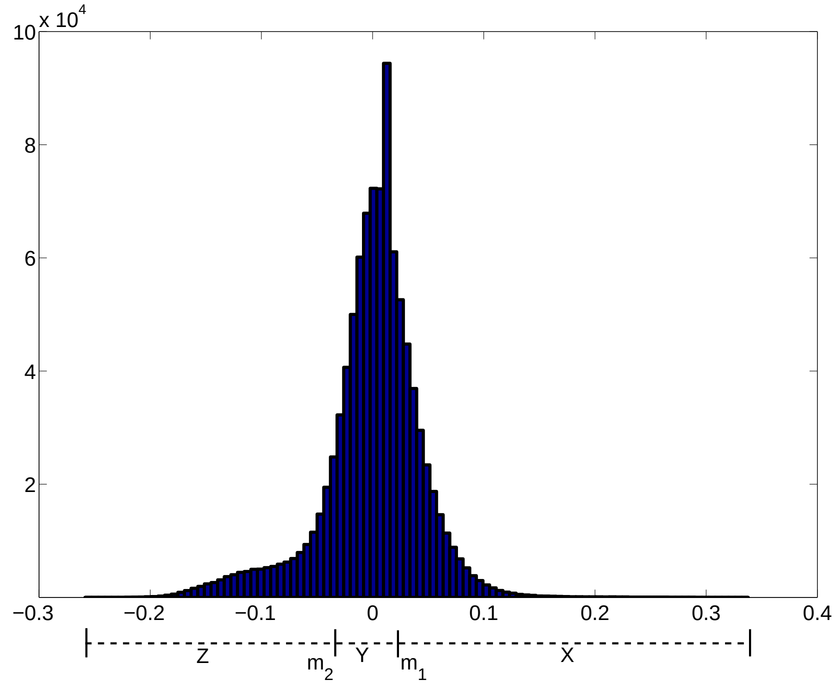
Having foreknowledge of the watermark, it is possible to draw its distribution before and after improving A. Figures 5.12(a) and 5.12(b) illustrate such distributions; we can see that pixels intensities were increased after iterating A – this highlighted and strengthened the watermark signal.

It is not clear in the general case whether the iteration will converge or when it should be halted, but we can demonstrate its beneficial effect from data with known ground truth. Figure 5.13 shows the SNR for such an example as the matrix A is iterated, showing that – as anticipated – the signal improves. In this case, the watermark signal keeps improving until a specific iteration, at which point there is convergence. SNR experiments were run on 30 randomly chosen sample pages.

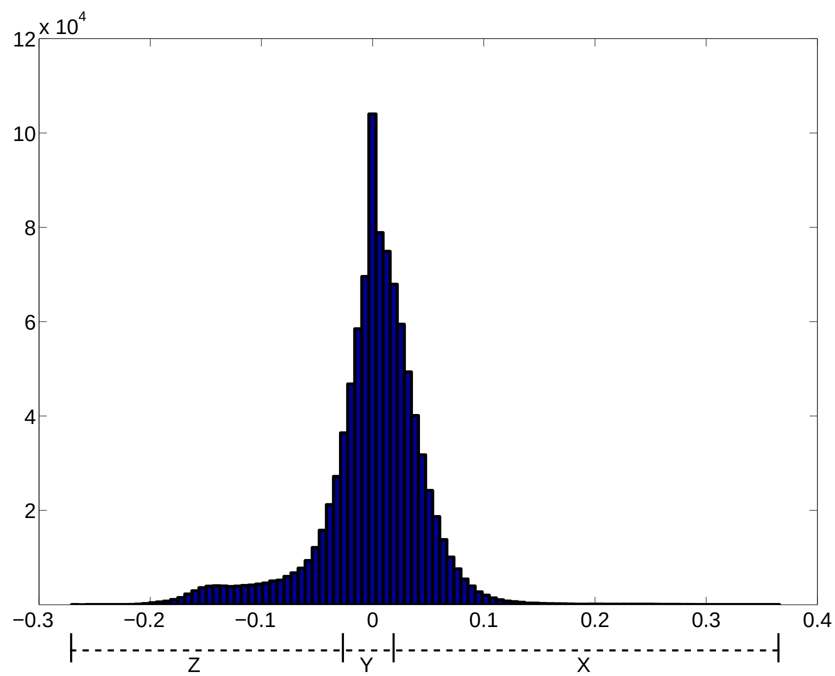
In the unknown case, SNR cannot of course be measured: Figure 5.14 plots the Frobenius norm [62] (a scalar that gives a magnitude measure matrix elements) of the difference between successive iterations of A (plotted for each cluster of intensities), suggesting that this mirrors adequately the signal improvement we wish to see.

We therefore adopt a convergence criterion that iterates until the matrix A stabilises (so the Frobenius norm of the difference between successive iterations becomes 0). This convergence depends upon the set of pixels being used to compute A becoming fixed at some stage. In all experiments, we have tried on a variety of datasets this has proved to be the case, but we cannot claim this will always be so. Therefore, when processing future datasets, a proposed solution is to iterate the process for a finite number of iterations: this number can be chosen experimentally by looking at the convergence cases in the datasets we examined. An acceptable approach is to inspect the Frobenius norm of the difference between successive iterations, and pick the iteration with the minimum value as the suitable stopping point. In perfect conditions, this minimum value will be (0), which is what we have observed in all test cases.

To observe the change in recomputing the transform, the initial matrix A, and after 10



(a)



(b)

Figure 5.11: Histogram distribution of image *D*, (a) before, and (b) after improving transform *A*

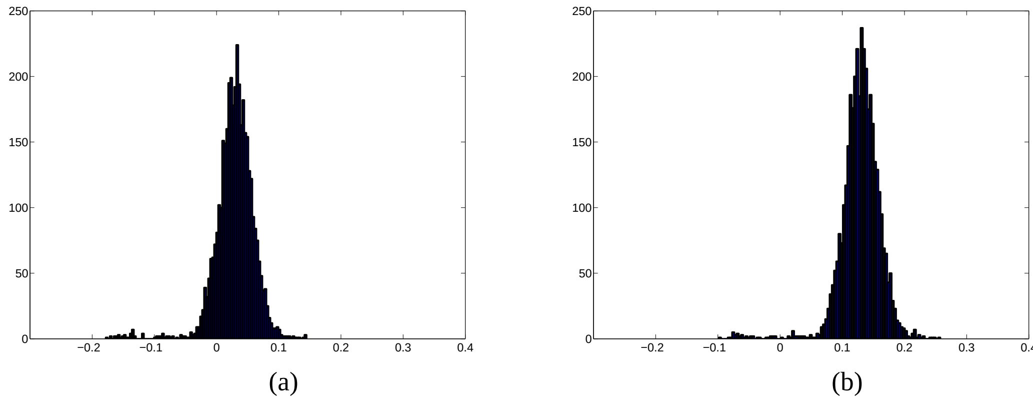


Figure 5.12: Histogram distribution of watermark features in D, (a) before, and (b) after improving transform A

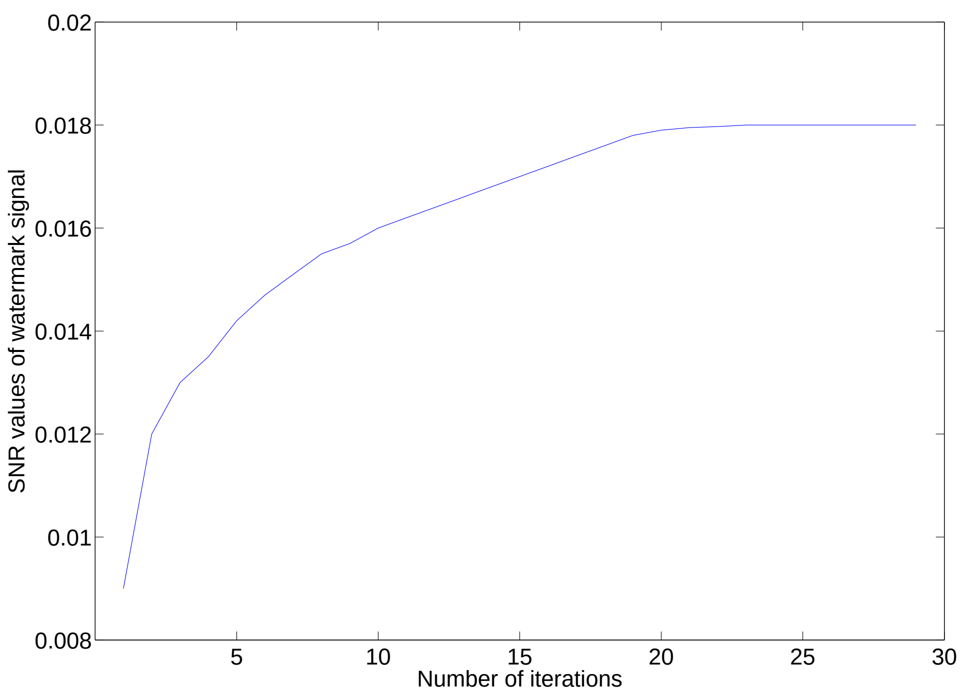


Figure 5.13: Evolution of SNR as transform A is iterated

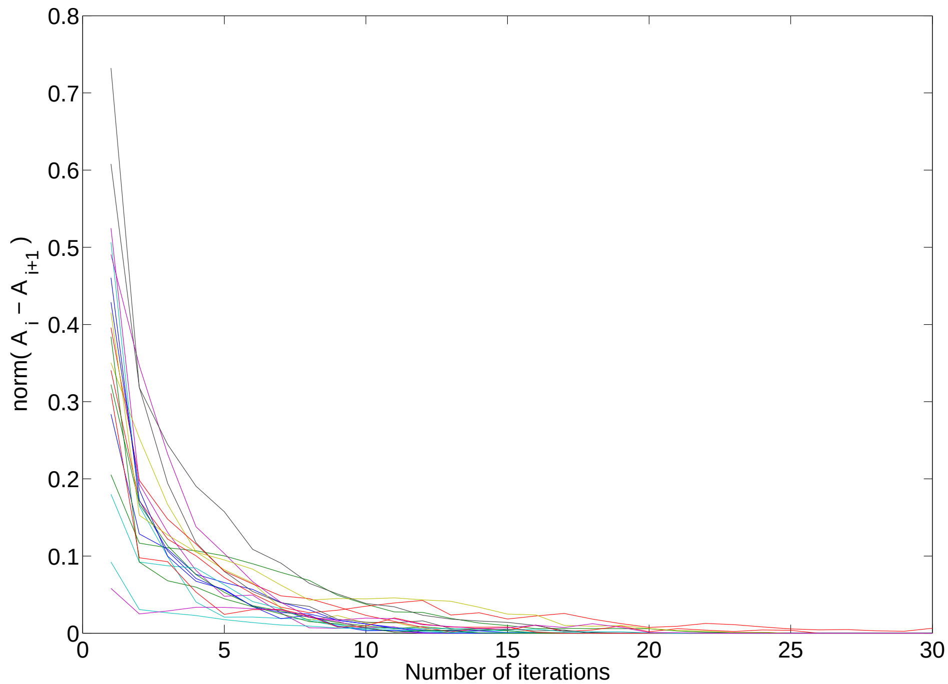


Figure 5.14: Frobenius norm of the differences in iterated values of A – each line denotes a specific cluster

and 30 iterations, for a specific cluster, are

$$\begin{bmatrix} 0.315 & 0.513 & -0.419 \\ 0.208 & 1.113 & -0.796 \\ -0.013 & 0.213 & 0.084 \end{bmatrix} \begin{bmatrix} 0.374 & 0.92 & -0.637 \\ 0.28 & 1.888 & -1.189 \\ 0.036 & 0.312 & 0.027 \end{bmatrix} \begin{bmatrix} 0.396 & 1.006 & -0.639 \\ 0.323 & 2.025 & -1.19 \\ 0.048 & 0.357 & 0.027 \end{bmatrix}$$

We can observe the change of the transform A as the iteration proceeds: the values of first and second column (red and green channels) has increased, while the third column (blue channel) has decreased. These observations vary among different clusters – for example, the initial values of A, and after 10 and 30 iterations, for a different cluster, are

$$\begin{bmatrix} 0.706 & -0.023 & -0.318 \\ 0.587 & 0.146 & -0.387 \\ 0.084 & -0.242 & 0.367 \end{bmatrix} \begin{bmatrix} 0.946 & -0.109 & -0.165 \\ 0.901 & -0.050 & -0.061 \\ 0.245 & -0.505 & 0.626 \end{bmatrix} \begin{bmatrix} 0.989 & -0.136 & -0.134 \\ 0.965 & -0.086 & -0.017 \\ 0.273 & -0.521 & 0.645 \end{bmatrix}$$

Here, the values of the first and third columns have increased, while the second column has decreased.

A particular parameter of this procedure is the number of RGB clusters K_1 defined in the reflected image S. Consideration of the 'best' number of clusters to seek via, e.g., K-means has received extensive attention in the literature [47, 114] – usually a trade off is sought such that this number satisfactorily captures the nature of the original data (i.e., K is 'high enough'), while allowing the centroids to represent the data with as little noise as possible (i.e., K is 'low enough'). Plotting clustering cost (usually summed distances from data to centroids) against K (see, for example, Figure 5.15), informally one seeks the point of diminishing returns where the cost starts to decrease very slowly: the L-method of Salvador [114] is a well-known approach.

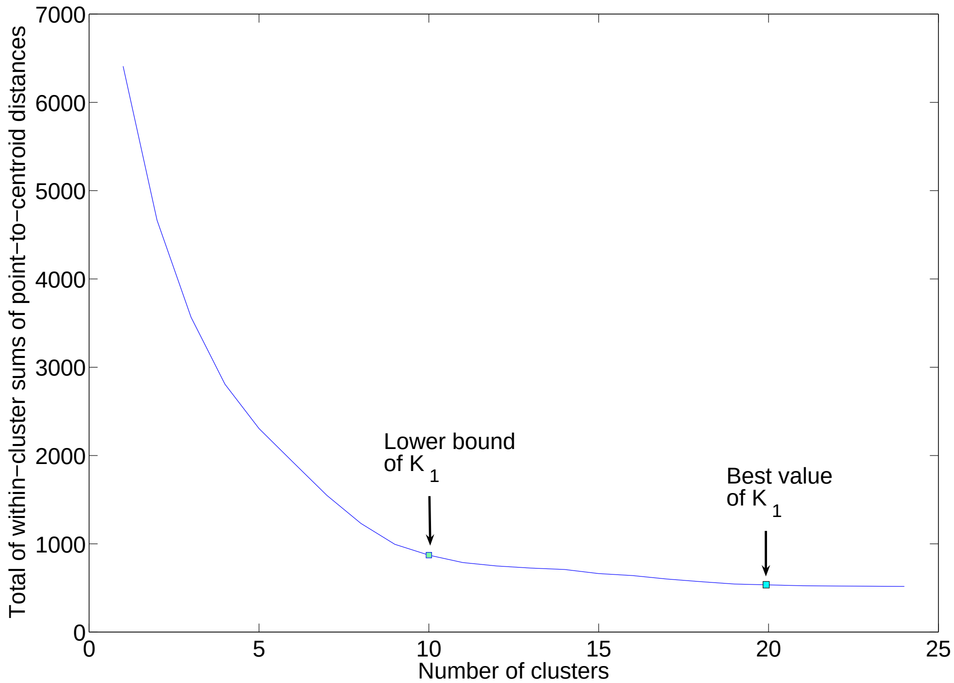


Figure 5.15: Distribution of number of clusters vs. clustering 'cost' in image S

The problem here is different: the more clusters we define, the better the subtraction process is likely to perform. However, we run the risk of developing clusters in which the watermark features will be numerically dominant.

To avoid this, a solution is proposed to estimate the best value of K_1. We know that watermark feature pixels are relatively bright. Based on this, we choose a lower bound for K_1 using the L-method approach [114] (see Figure 5.15), and iterate it until reaching an unacceptability criterion.

We have knowledge of the mean of image B

$$(\mu_\rho, \mu_\gamma, \mu_\beta) = mean(\rho_p, \gamma_p, \beta_p) : p \varepsilon B$$

and can similarly compute a mean from B for each cluster $C_1, \ldots C_{K_1}$

$$(\mu_\rho^i, \mu_\gamma^i, \mu_\beta^i) = mean(\rho_p, \gamma_p, \beta_p) : p\varepsilon C_i, i = 1, \ldots, K_1$$

We then compare the image RGB mean $(\mu_\rho, \mu_\gamma, \mu_\beta)$ with every cluster RGB mean value $(\mu_\rho^i, \mu_\gamma^i, \mu_\beta^i)$, seeking none of these to be 'large'. There are many ways of doing this: by experiment we discover that the condition

$$\mu_\rho^i > \mu_\rho \text{ AND } \mu_\gamma^i > \mu_\gamma \text{ AND } \mu_\beta^i > \mu_\beta$$

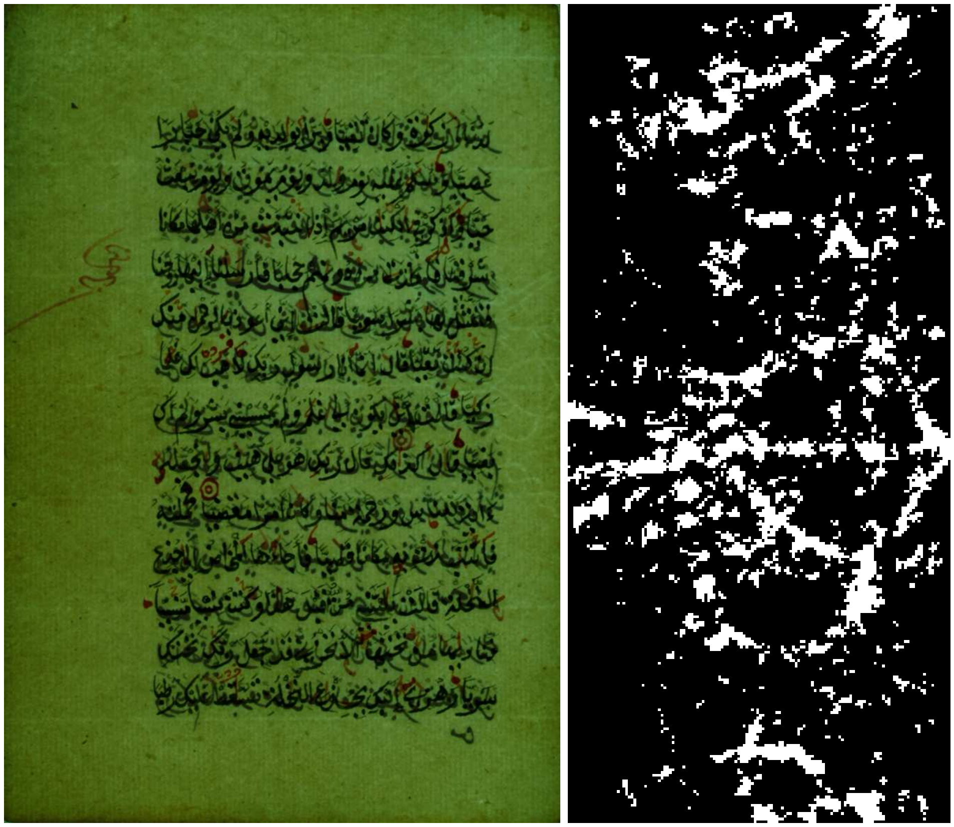
is sufficiently strict. Should a cluster channel mean exceed the global one on all three colour channels, we decrement K_1 and accept it as the value with which to proceed.

Figure 5.16 illustrates a backlit image B and one of the clusters C_i when clustering with $K_1 = 21$. Part of the watermark is very evident in this cluster. For these data, $(\mu_\rho, \mu_\gamma, \mu_\beta) = (69, 98, 29)$, while $(\mu_\rho^i, \mu_\gamma^i, \mu_\beta^i) = (91, 129, 53)$ – higher than the image mean for each component. This indicates that in this case K_1 should be less than 21, and we find a satisfactory result with 20 (indicated in Figure 5.15).

Having foreknowledge of the watermark design and its position, we can verify the applicability of the preceding algorithm. At each iteration, we consider the pixel locations of each cluster in B, and compare them with the location of the known watermark. If most pixels of a single cluster represent watermark features, then we decrement K_1 and compare it with the best K_1 obtained from the algorithm. This verification was successful with 30 chosen randomly test pages.

Characteristically, for the difficult data of the 'Mahdiyya' Qur'$\bar{\text{a}}$n, starting values of K_1 chosen by the L-method were in the range 9-11, and the final chosen values using our algorithm were in the range 20-25 clusters. The difference in range between the two approaches is obvious: our approach provided better clustering of intensities, and hence better subtraction results compared to lower values of K_1.

An example of a cluster distribution of a sample input S is in Figure 5.17(a), and a transformed image of S is in Figure 5.17(b). The number of RGB clusters here is 20: we can see how clustering reflects the variation of features. It is clear that background features vary from one region to another. This variation, together with the existence of recto features, makes transforming each cluster separately necessary to model the back-lighting.



(a) (b)

Figure 5.16: (a) Backlit image B, (b) Pixels of a specific cluster within B (displayed in white, with all others erased to black for display). Part of the watermark is seen to predominate in this cluster.

5.5.3 Watermark location

As discussed in Section 5.4, for our data, the differenced image D can be further improved by setting negative pixel values (which correspond, for example, to verso features) to 0 – we set a pixel value to 0 if any of its RGB channels is negative. Figure 5.18 shows the resulting D, enhanced for better visualisation. Observe here that the watermark signal becomes stronger, while the interference of recto and verso features become low, because these features now have low magnitude pixel values.

While the watermark features are partially evident here, we are still at the mercy of very considerable noise. We have sought to find a partial segmentation by clustering to K_2 centroids the RGB data in D; this time the L-method [114] is a suitable approach. Figure 5.19 shows a plot of cost against K_2 and the derived number of clusters (here 10) – characteristically with the hard data this number is in the range 8-10 clusters. Figure 5.20 illustrates the cluster distribution of D: the zoomed window shows that these clusters do

(a) (b)

Figure 5.17: (a) Clusters distribution of image S presented in Figure 5.10(a), using $K_1 = 20$, (b) Transformed image of S

successfully pick out watermark features (in addition to many noise and other artefacts).

When applying the matching process, selecting significant peaks in the accumulated response M (equation 5.8) is important in locating the watermark fragments. We propose a thresholding approach on this array and then selecting the centroid – or weighted centroid – of regions that pass it.

This approach, with well-chosen templates, seems to have promise but is often troubled by noise, and this leads to the existence of many significant peaks for every fragment. A simple approach to find the exact watermark location is by exploiting the fragments' known geometric relationship (distance and rotation angle) in inspecting these peaks. In other words, we will be seeking co-occurrences of peaks in accumulated M arrays that match the known geometric relationship of the fragments.

In thresholding the accumulated array M, one approach is to determine the mean response μ and the standard deviation σ, and seek a suitable multiplier s, thresholding at $\mu + s\sigma$. We have sought to set s on the basis of a known dataset. Firstly, the response M is found for each watermark fragment in each of the sample data. Then s is speci-

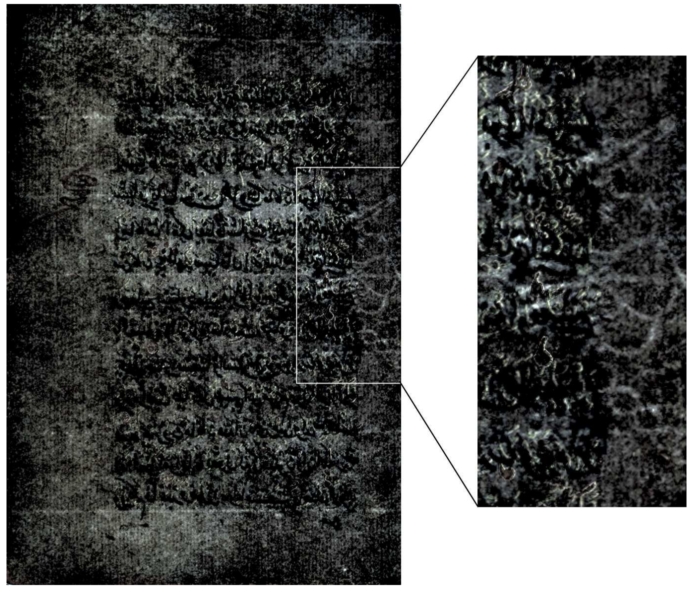


Figure 5.18: Differenced image D. A watermark fragment is visible in the right hand margin.

fied by finding (manually) the exact location of the watermark fragment in the histogram distribution of M, and determining the value $\mu + s_k\sigma$ at that location. Finally, we pick the ‘reliable’ s as the minimum of all s_k values. Figure 5.21 illustrates the selection of s (marked) using a sample set of different M responses. This procedure indicates that $s = 6$ is a suitable value.

Figures 5.22, 5.23 illustrate this response M for two watermark fragments, where dots denote significant peaks, and squares as their centroids – zoomed for better viewing.

After choosing the centroids of significant peaks for each fragment, we find the geometric relations (distance D and rotation angle θ, as illustrated in Figure 5.24) between each pair of these (a many-to-many relation).

Known geometric relations are inspected between significant peaks in a generalised Hough transform-like approach [122]. Figures 5.25(a) and 5.25(b) show the significant peaks in the accumulator response M for two fragments after matching. Geometric relations D and θ are found for each pair (p_i^1, p_j^2), where i and j indicate significant peaks for each fragment. Figure 5.25(c) illustrates the parameter space, where the cross-mark denotes the known geometric relation, and dots as the geometric relations between each pair. The closest point is taken as the best matching.

To find the best match, the summation of absolute difference between these values and the values of the known fragments (p^1, p^2) are determined:

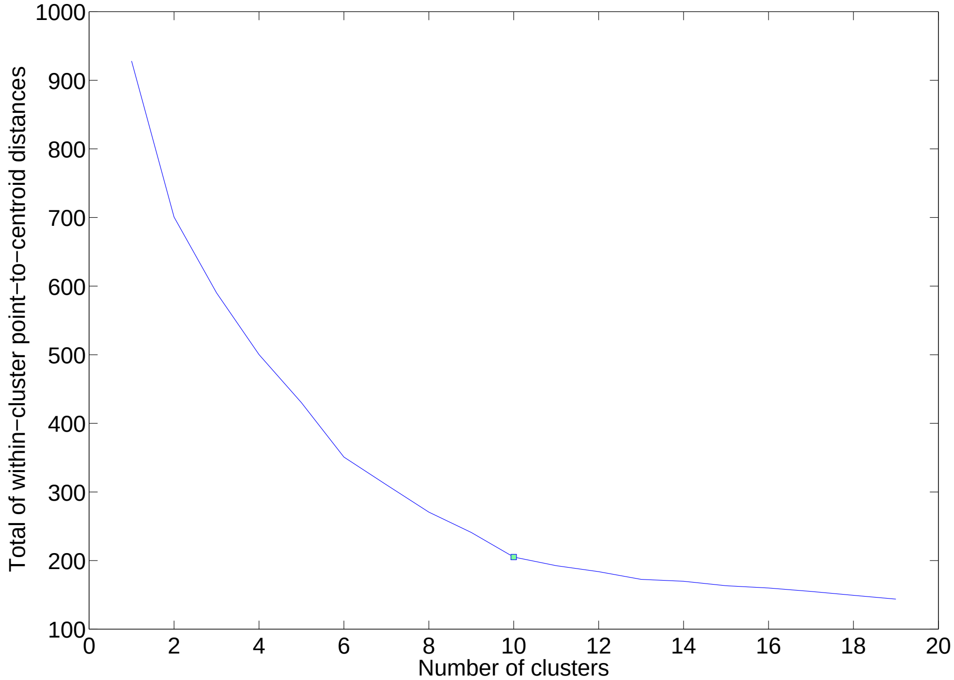


Figure 5.19: Distribution of number of clusters vs. summation of point-to-centroid distances in image D

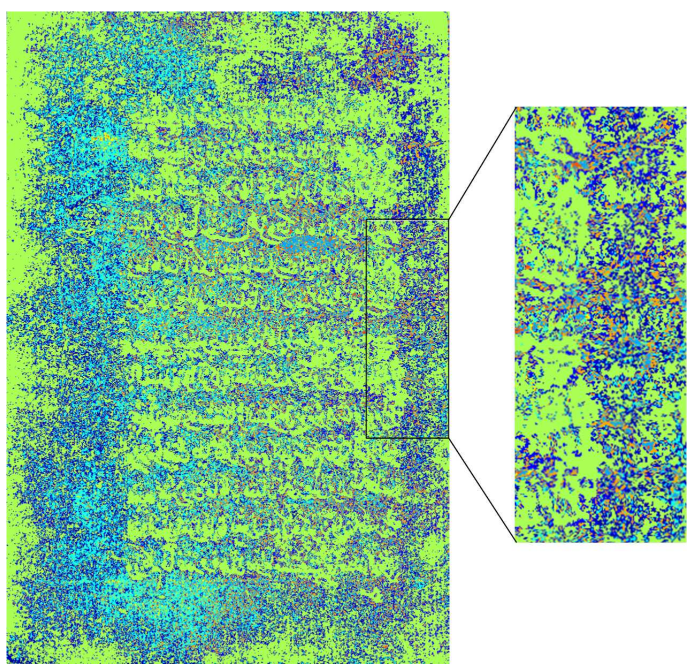


Figure 5.20: Clusters distribution of image D presented in Figure 5.18, using $K_2 = 10$, with watermark area enlarged on the right

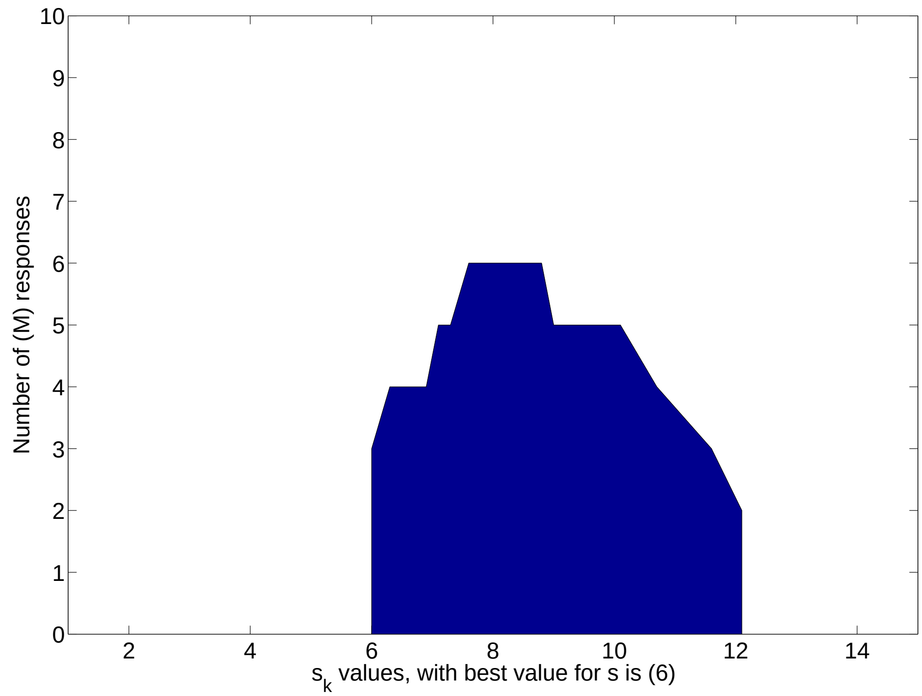


Figure 5.21: Finding best value for standard variation multiplier s

$$w(p_i^1, p_j^2) = |D_{(p_i^1, p_j^2)} - D_{(p^1, p^2)}| + \lambda |\theta_{(p_i^1, p_j^2)} - \theta_{(p^1, p^2)}| \tag{5.10}$$

Here, λ recognises the different scale of the distance and angle contributions to this cost. In experiments we have performed, $\lambda = 1$ has been seen to give a satisfactory result, and we have not explored this choice deeply. The weight w is calculated for all peak pairs, and the minimum, w_{min}, is taken as the best possible match. w_{min} is compared with an acceptability threshold t. This threshold has been determined by inspecting sample test data of different, known, watermarks. From experiments, we found $t = 10$ to be an acceptable choice. If w_{min} is less than t for a specific pair, then this pair is chosen as the possible best match.

In the event of there being three (or more) fragments $(p^1, p^2, p^3, \ldots)$, the same procedure is applied for each fragments' pair: i.e., the relation values are calculated for all pairs. The reason for treating fragments as pairs and not all together is because (as observed in many cases in our experiments) one or more of the fragments may not be visible in the image due to a weak watermark signal. When treating fragments as pairs, the classifier will find the best match.

Further, in the case of three (or more) fragments, it may happen that there are two different best matchings for one fragment. Fortunately, conflicts can be resolved by finding

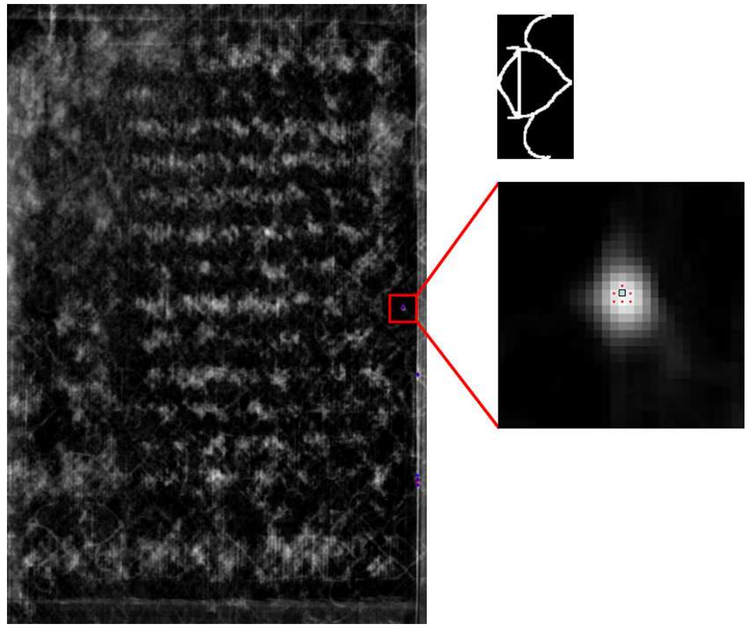


Figure 5.22: The accumulator M, with positions of significant peaks of 1st fragment $(s = 6)$, and its selected centroids, square-marked

the odd one out. For an example of three fragments (p^1, p^2, p^3), if the coordinates (x, y) of the best matching ϕ for the pairs are

$$\begin{aligned}
\phi(p^1, p^2) &= (600, 700), (700, 700) \\
\phi(p^1, p^3) &= (200, 100), (100, 100) \\
\phi(p^2, p^3) &= (700, 700), (500, 700)
\end{aligned}$$

then based on the matching coordinates of the second fragment, we can decide that the correct matching peak of the first fragment is located at the coordinates $(600, 700)$, the second at $(700, 700)$ and the third at $(500, 700)$.

Our classifier works well in recognising the watermark designs, even those of weak signal. Table 5.1 shows the retrieval results for four design parts, which represent a double-headed eagle watermark ‘E’, and a moonface-within-shield countermark ‘M’ used in the ‘Mahdiyya’ copy of the Qur’ān. The table shows excellent matching results – our classifier managed to find similar designs with a high percentage of true positives (correct matching), and no false positives.

However, there is still a small percentage of false negatives (missed matches). This is due to the threshold used to select significant peaks (s), because the watermark signal in these false negatives is very weak. A possible solution is to decrease the threshold to find the correct match, but this may affect overall results – decreasing s will result

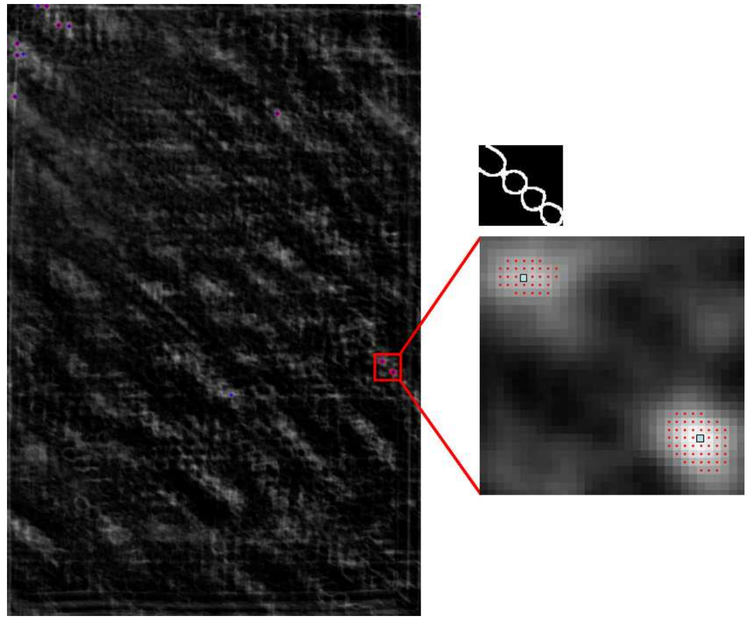


Figure 5.23: The accumulator M, with positions of significant peaks of 2nd fragment $(s = 6)$, and its selected centroids, square-marked

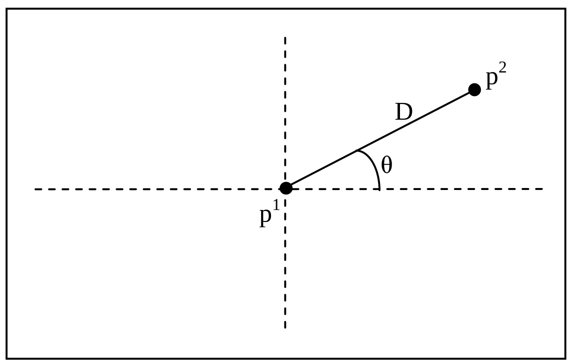


Figure 5.24: Geometric relations between a pair of significant peaks

in the appearance of many peaks. Even deploying the known geometric relationship of fragments will leave many false positives. Experiments show that decrementing s by 1 resulted in an average of 10% of false positives.

Figures 5.26 and 5.27 show the centroids of significant peaks of two fragments when choosing $s = 5$ instead of 6. In this example, it is obvious that there are many centroids compared to those of Figures 5.22 and 5.23. Consequently a false positive is generated, because there is more than one pair of peaks (p_i^1, p_j^2) which have geometric relations close to those of the original known fragments. We see that the choice of s is thus critical to results. On the other hand, having more watermark fragments will reduce this problem, since the number of significant peaks will be reduced by the geometric relations between

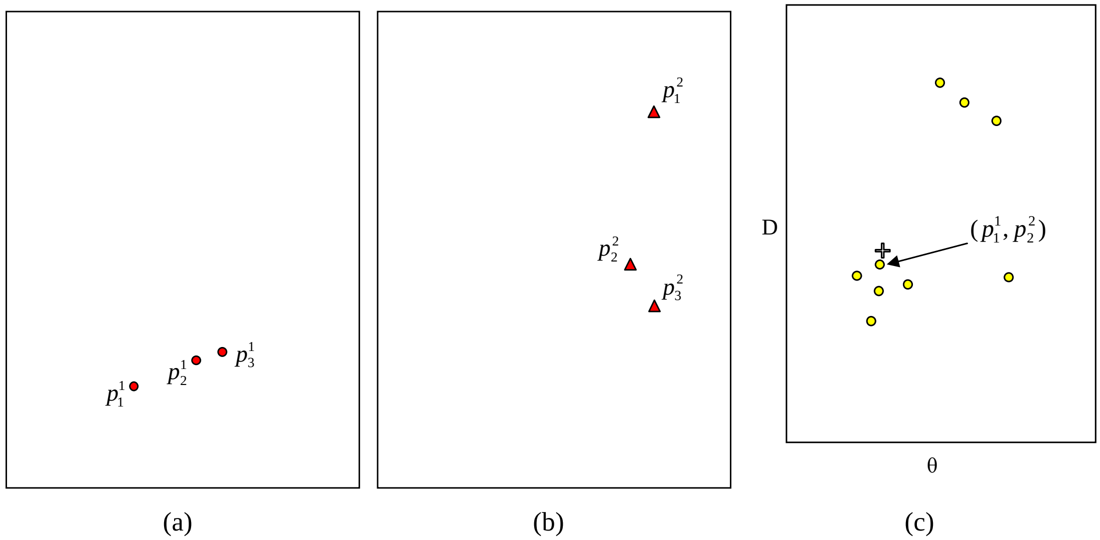


Figure 5.25: Locating best matches between fragments, (a) Significant peaks of 1st fragment, (b) Significant peaks of 2nd fragment, (c) Parameter space: the known relation is cross-marked, and pair (p_1^1, p_2^2) is the best match.

Table 5.1: Percentage of matching results for different watermark shapes (%)

Watermark	M (upper part)	M (lower part)	E (upper part)	E (lower part)
True positive	98.8	97.7	96.5	94.3
False positive	0	0	0	0
True negative	100	100	100	100
False negative	1.2	2.3	3.5	5.7

them, provided the watermark signal is not very weak. We experimented with selecting 3 more fragments for each watermark (so each design is represented by either 5 or 6 fragments). We found that the average percentage of false negatives was reduced from 10% to 3%.

We also tested our approach with other, simpler, datasets presented in Sections 3.1.4 and 3.1.5; it worked successfully, with 100% true positives, and 100% true negatives. This is no surprise, since the ‘Mahdiyya’ copy of the Qur’ān is the most difficult dataset we used. Success with these other datasets demonstrates that this approach has good applicability.

5.6 Watermark aggregation

Given a reliable watermark extraction algorithm, we can try to recapture with some accuracy the full original design by aggregating the registered images: the watermark signal

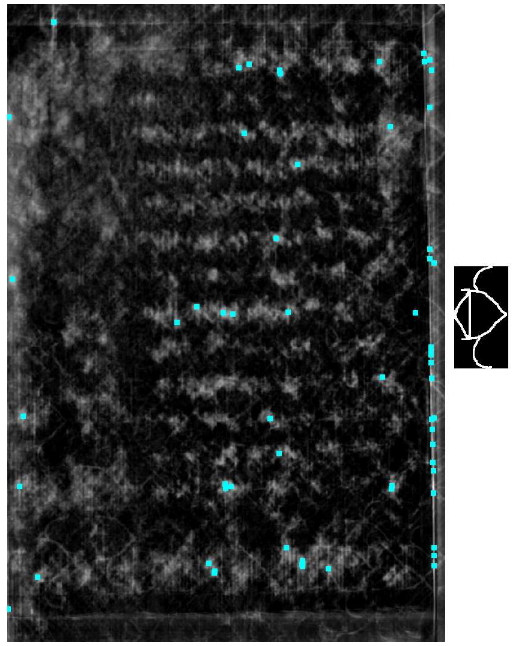


Figure 5.26: The accumulator M, with positions of peaks of the 1st fragment ($s = 5$), square-marked for display

should reinforce while all other features might be expected to be unpredictable (although maybe not random) in location, and so would not reinforce. Such an aggregation would be useful because

- It would allow the recapturing of a complete watermark even though only a fragment was used to locate it in the image.
- It would help distinguishing 'identical' from 'twin' watermarks, since it will help observing differences between these designs, when laid together, that could not be observed before.
- It would highlight and clarify chain lines, which are significant to scholars in paper studies.

We have performed this for a number of difference images (after nulling the verso 'signal' pixels), for a known watermark, and compared the result with ground truth to judge its quality. This comparison is via the SNR measure discussed in Section 5.5.1.

The value and interest of the aggregation procedure is well demonstrated by the following example, since it has revealed details of watermarks that we could not observe before. Figure 5.28 (also enlarged in Figures C.7 and C.8 in Appendix C) illustrates the superimposition of the double-headed eagle, and moonface-within-shield designs: we

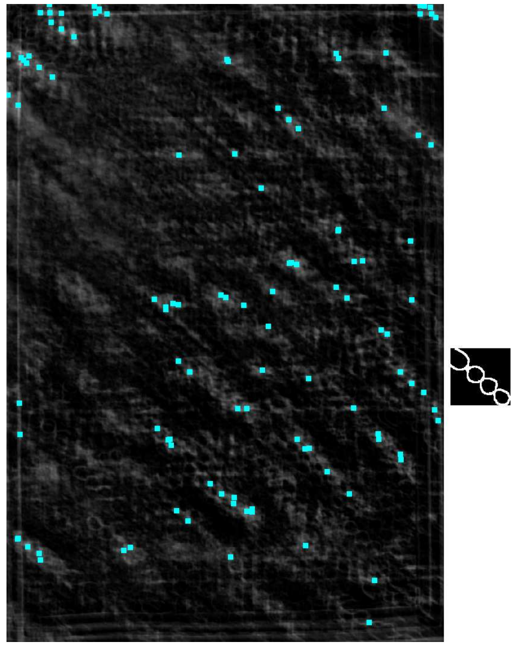


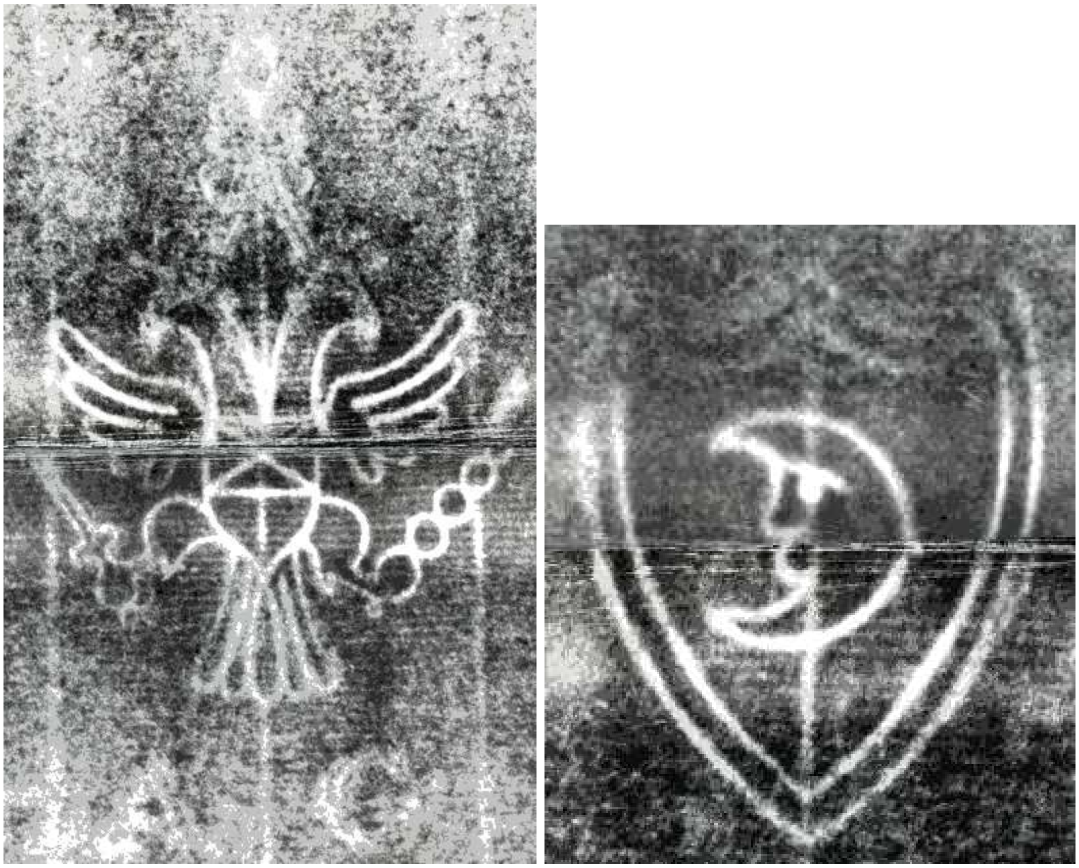
Figure 5.27: The accumulator M, with positions of peaks of the 2nd fragment ($s = 5$), square-marked for display

could not detect the ‘A G’ countermark below the eagle in single sheets before applying this process, and many details of the design become clear that cannot be detected in individual sheets. We can observe chain lines have developed high responses in the aggregated image. It was difficult to study these in individual sheets due to their weak signal.

The more superimpositions, the clearer the watermark details. Experiments confirm that adding more samples provides a better SNR than individual images until some convergence point. Figure 5.30 (solid line) shows SNR values of superimposing 2 and more differenced images D_k of the double-headed watermark.

It is clear that some parts of the superimposed watermarks in Figure 5.28 are brighter than others; lower quality areas are attributable to the [removed] presence of recto features, and the nulling of pixels associated with verso features. We experimented with neglecting ‘nulled’ pixels when performing the averaging. A result of the double-headed eagle after this step is in Figure 5.29(b): the variation in watermark brightness is reduced, however this affected the strength of the signal. We measured the SNR of the superimpositions, and found the values low compared to that achieved before, as illustrated in Figure 5.30 (dotted line).

The aggregation operation could also be very useful in the study of ‘twin’ watermarks, because when similar designs are superimposed together, it could be easy to identify the



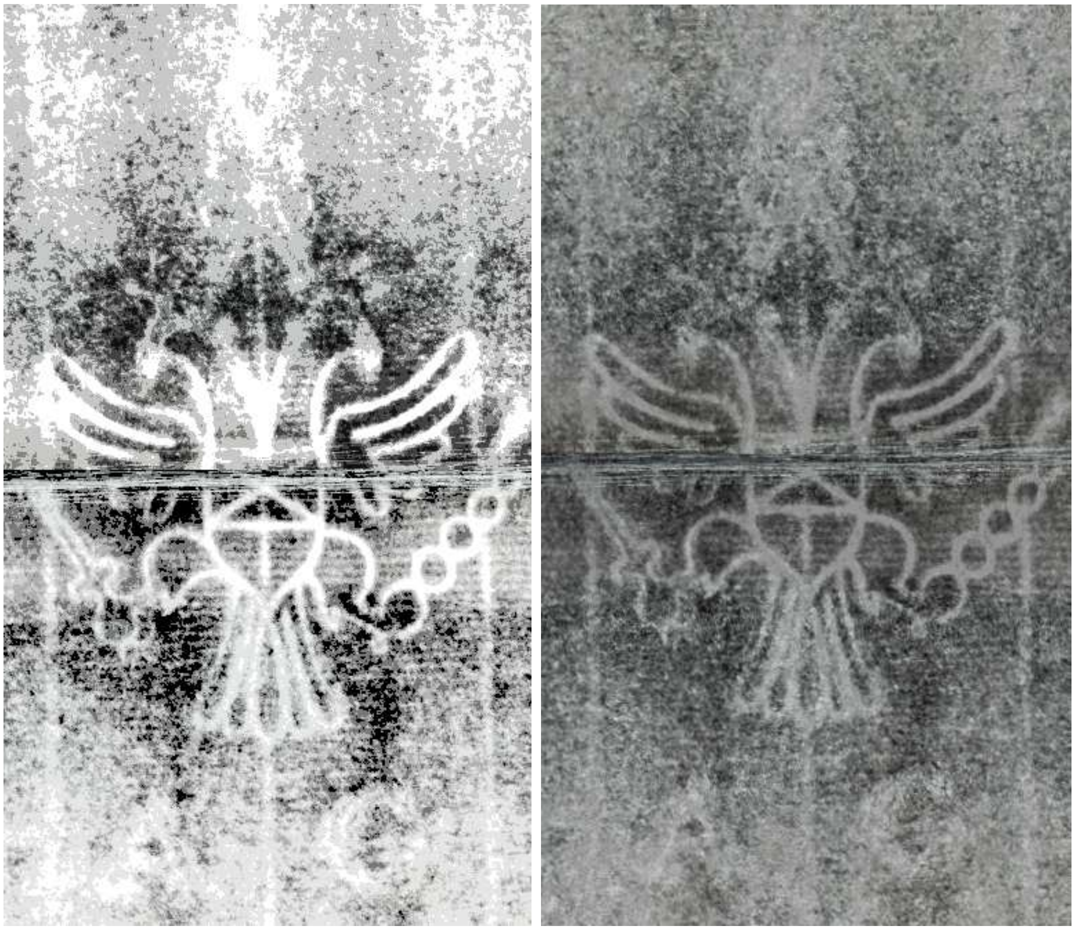
(a) (b)

Figure 5.28: Complete watermark designs used in the 'Mahdiyya' copy of the Qur'ān data. There are two, but paper was cut in two to form pages, giving in all four different patterns on *most* pages.

differences between them. To illustrate this, Figure 5.31 shows 3 trelune watermarks taken from different sheets (of the Prayer presented in Section 3.1.4); these designs have been coloured to highlight any differences that exist. Figure 5.32 shows the aggregation process: in this example, the first two watermarks were observed as 'identical', where the third shape was 'twin' – this is obvious by looking into the slight changes of the crescents' edges. This Figure is magnified for better visualisation.

5.7 Conclusion

This Chapter presented a model-based approach to locating watermarks in scanned documents; it managed to remove recto material successfully, and developed a statistical approach to locate watermark fragments from a known lexicon. Results show a very good ratio of retrieval correctness.



(a) (b)

Figure 5.29: Superimposed watermark design (a) with 45 superimpositions, and (b) after neglecting null pixels

The algorithm depends on some global parameters that control clustering and signal thresholding (from noise), and we have considered robust means of choosing these.

This approach has been used to locate watermarks in two nineteenth century copies of the Qur'ān and a Prayer [76]. Locating such 'hidden' material in this data is difficult, because these data are characterised by thick recto and verso writing, the paper used is thick, and the watermark patterns are not clear, resulting in high foreground interference, and a weak signal of the watermark shape. These data, together with individual manuscripts presented in Section 3.1.2, proved that this approach works with various sets of data of different attributes.

We further presented an aggregation of located watermarks that has been seen to enhance the detected detail. This operation is important as it can reveal subtle details in designs that are difficult to observe in single watermark designs. This procedure is very

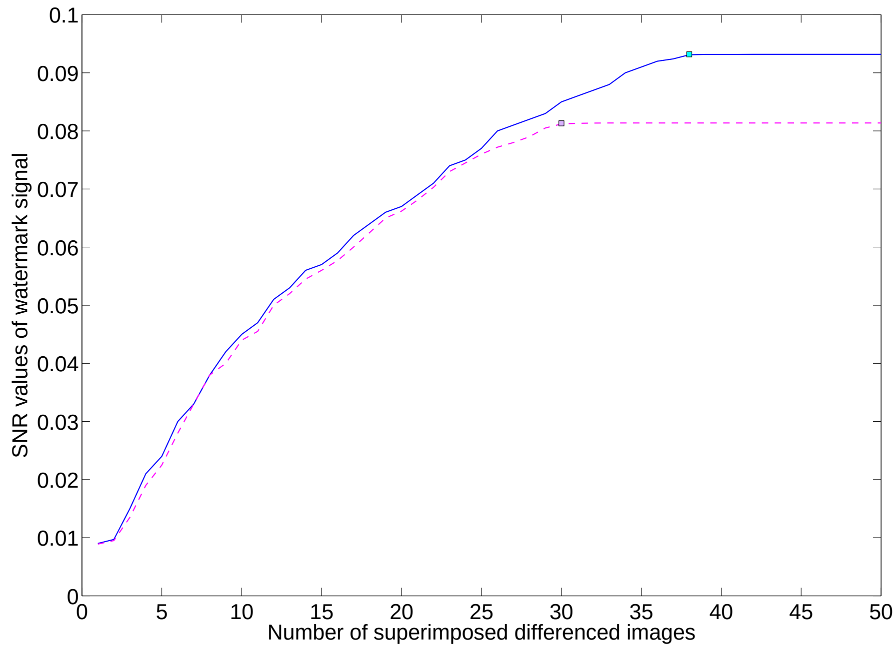


Figure 5.30: SNR values of superimposed differenced images D_i (solid line), and after neglecting null pixels (dotted line)

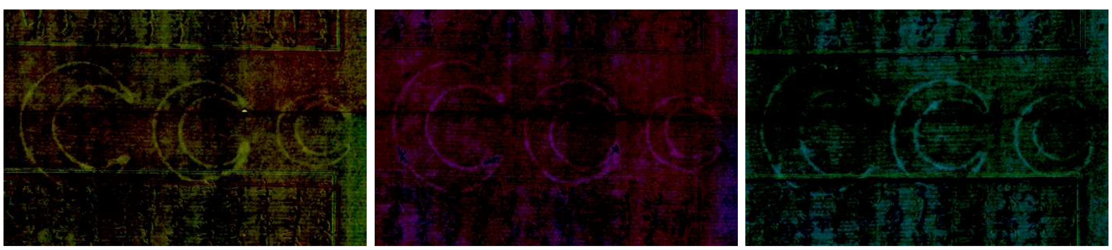
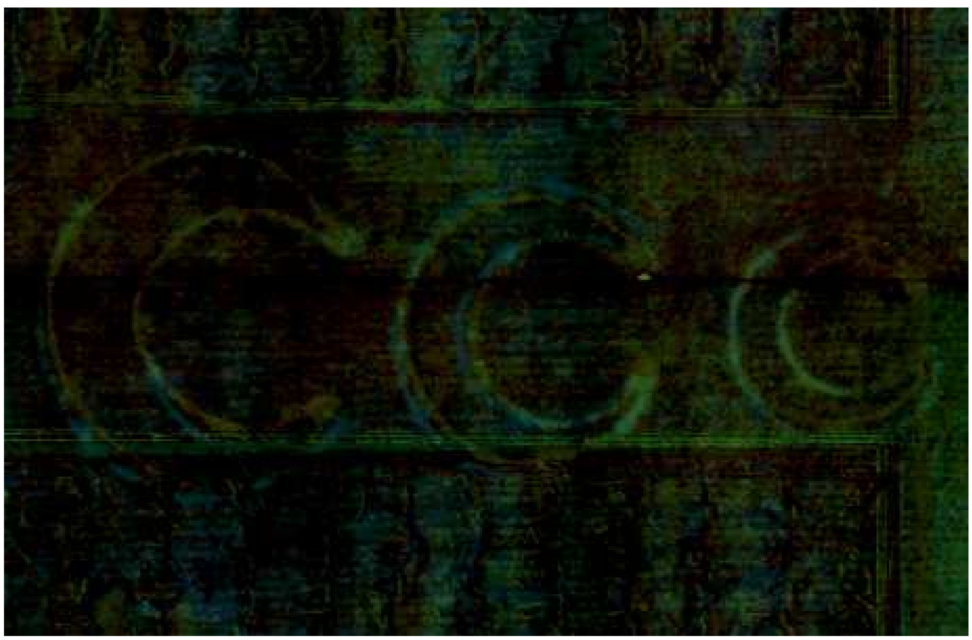


Figure 5.31: Three trelune watermarks in different D images, coloured in yellow, magenta and cyan respectively

useful in highlighting chain lines, which are very hard to observe in individual sheets. This operation could also be very useful in studying 'twin' watermarks, since it may be easy to identify the differences between designs when laid together.

This approach requires a foreknowledge of the watermark designs in order to proceed. In some cases this will not be an obstacle (it is being sufficient to have a set of watermarks of which the observed one is a member). Should this not be viable, our approach will succeed given a *part* of a watermark which may be outlined interactively on screen by a user as part of an initialisation phase. It is possible to conjecture an automatic approach to locate these designs without any previous knowledge of their structure – possible approaches to this are considered in Chapter 7.

(a)



(b)

Figure 5.32: Aggregated watermark designs of Figure 5.31, (a) the aggregation of first and second designs ‘identical’, (b) the aggregation of first and third designs ‘twins’.

Chapter 6

Post processing

6.1 Introduction

In this Chapter, we discuss further processing to the bottom-up approach presented in Chapter 4. This includes vectorising bit-mapped output images, and interactive applications to assist manual removal of defects and residual noise on the paper.

The post-processing presented here has particular advantages: it provides users with the necessary tools to edit and enhance extracted watermark patterns. The post-processed results are in vector representation and can be simplified, zoomed at large scales, and printed in high resolution.

The motivation behind offering vectorisation and interactive tools is to provide a simple and easy environment for different users. By design, these tools can deal with patterns interactively without any previous knowledge of using computers being necessary. For example, these tools can be helpful in the removal of unavoidable noise, and completing missing parts of the extracted designs.

6.2 Vector representation and simplification

At this stage, the bit-mapped watermark design output from the bottom-up approach is traced and converted to a simplified vector graphical representation – this offers a number of advantages, including:

Figure 6.1: Output after vectorisation

- Vector graphics are produced by a sequence of commands or mathematical statements, and a vector file is smaller than a corresponding bit-map.
- Vectors are resolution independent, meaning that they can be zoomed to any scale with quality preserved, without any degradation.
- This graphical description can be read and modified by a large range of tools (e.g. Notepad), and further may be printed with high quality at any resolution.

The boundary pixels of the watermark pattern are detected and extracted, and then converted to vector data. A vectorised watermark (of the pattern presented in Figure 4.22(a) in Section 4.2.2.3) is in Figure 6.1. Visually, the output consists of the same shape as in the segmented result, however, the shape of the watermark is now represented by a vector description and no longer in pixels.

Vector representations are open to simplification, in which the number of edges and vertices of a polyline is reduced, retaining only those seen as 'necessary'. This can make the representation far more accessible to editing and manipulation by different classes of user. We present here three polyline simplification methods that have been implemented.

Polyline variation : given a polyline P with n vertices, we compute the weight of each vertex v_i – "the vertex weight is a measure of variation of the polyline at the specified vertex. A simple measure of weight is based on three consecutive vertices, v_{i-1}, v_i, v_{i+1}" [44]:

$$w_i = \frac{Distance^2(v_i, segment(v_{i-1}, v_{i+1}))}{Length^2(segment(v_{i-1}, v_{i+1}))}$$

where $segment(v_{i-1}, v_{i+1})$ is the line segment connecting vertex v_{i-1} to v_{i+1}, and $Distance(v_i, segment(v_{i-1}, v_{i+1}))$ is the distance between v_i and $segment(v_{i-1}, v_{i+1})$. The vertex with the smallest weight in P is removed to obtain P', and the algorithm

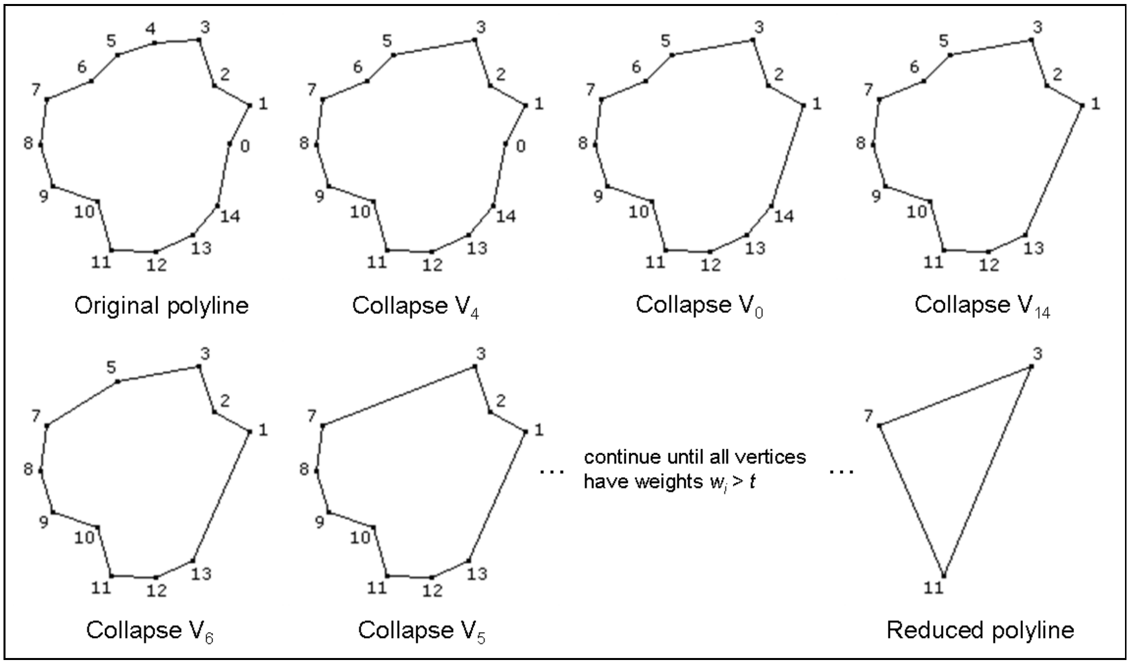


Figure 6.2: Description of the polyline variation simplification method. With permission from David Eberly [44]

is repeated on P' recursively. The process stops when the smallest weight becomes larger than a given threshold t. An example is given in Figure 6.2.

Vertex reduction “...a polyline vertex is discarded when its distance from a prior initial vertex is less than a minimum threshold $t > 0$. Specifically, after fixing an initial vertex v_0, successive vertices v_i are tested and rejected if they are less than t away from v_0, when a vertex is found that is larger than t, then it is accepted as part of the new simplified polyline, and becomes the new initial vertex for further simplification” [131]. Figure 6.3 illustrates this method.

Douglas-Peucker simplification [40]: This algorithm was later modified by Hershberger and Snoeyink [73] to reduce running time.

In this algorithm, “the two extreme endpoints of a polyline are connected with a straight line as the initial rough approximation of the polyline. Then, how well it approximates the whole polyline is determined by computing the distances from all intermediate vertices to that finite line segment. If all these distances are less than the specified threshold t, then the approximation is good, the endpoints are retained, and the other vertices are eliminated. However, if any of these distances exceeds t, then the approximation is not good enough. In this case, choose the point that is furthest away as a new vertex subdividing the original polyline into two shorter

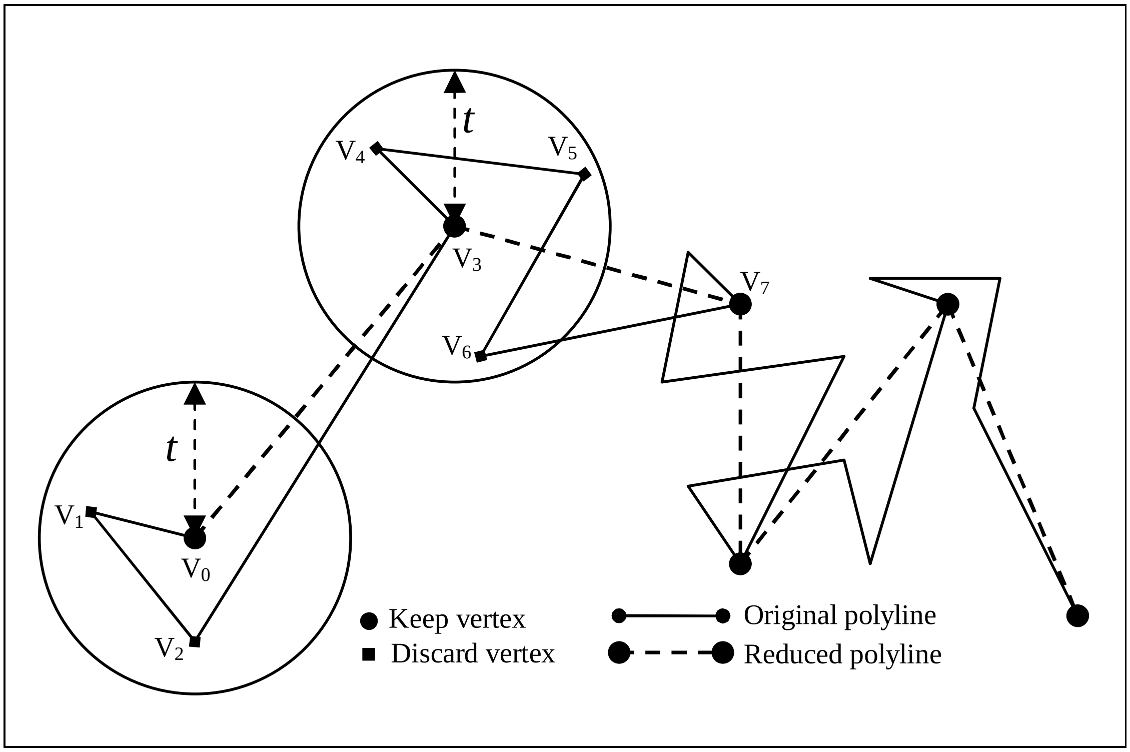


Figure 6.3: Description of the vertex reduction method [131]

polylines. This procedure is repeated recursively on these two shorter Polylines. If at any time, all of the intermediate distances are less than the *t* threshold, then all the intermediate points are eliminated" [131]. An example explaining how this algorithm works is in Figure 6.4; a more detailed explanation of stages is in Figure C.6 in Appendix C.

The resulting graphical representation is stored in SVG (Scalar Vector Graphics) vector file format [135]. This format provides wider accessibility through the web, contents of SVG vectors can be searched and indexed easily [72]. An example of vector simplification using the Douglas-Peucker Polyline simplification algorithm is in Figure 6.5(a), which shows the original exported vector without simplification. In this case 9332 vertices where used to represent the vector, while the simplified version illustrated in Figure 6.5(b) needed only 826, with a short processing time compared to the non-simplified version. From our experiments, the simplified vector has generally over 90% fewer data points compared to the original vector, which has the advantage of making the design easier to modify for interactive editing and enhancements.

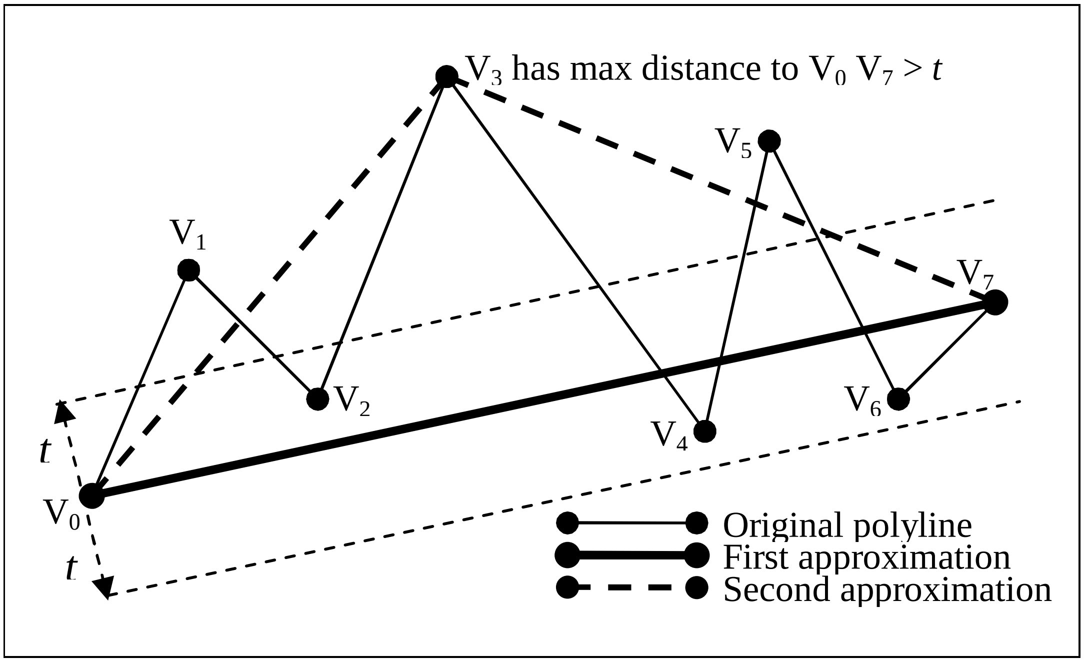


Figure 6.4: An example illustrates the Douglas-Peucker Polyline simplification algorithm [131]

6.3 Interactive enhancements

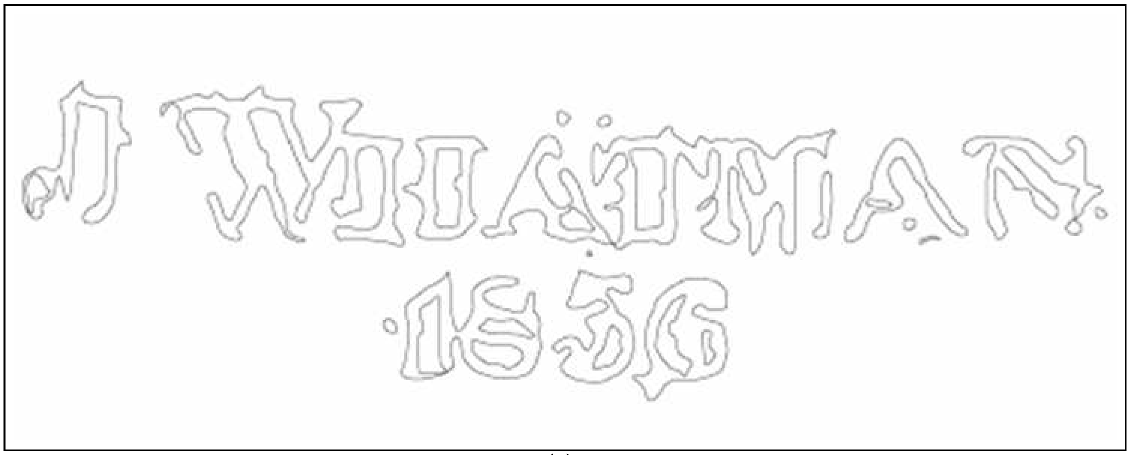
Much of the work in this thesis is motivated by the need of scholars with little or no experience in computing to work on documents of interest to them. Recognising that 'perfect' solutions are unlikely, particularly with more challenging inputs, it becomes useful to provide such scholars with an interactive means to work on watermarks. To this end, tools with simple interactive image and vector editing functionalities were also developed to allow manual removal of defects or residual noise on the paper.

A simple facility is the ability to view how image intensity data are distributed by looking at the image histogram distribution; an example is in Figure 6.6.

Tools were also built to apply semi-automatic interactive editing functions to binary images, and vectors in SVG format. The image editor (fully illustrated in Figure C.2 in Appendix C) is used to enhance image resulting from the segmentation stage. It includes four main functions:

1. Remove: to eliminate residual noise objects. This works by clicking on the object to be removed; an example is in Figure 6.7.

2. Connect: to connect two selected points together with a line of foreground pixels of an automatically adjusted width, depending on the objects behaviour in the area around selected points. Connecting functionality is illustrated in Figure 6.8.


(a)

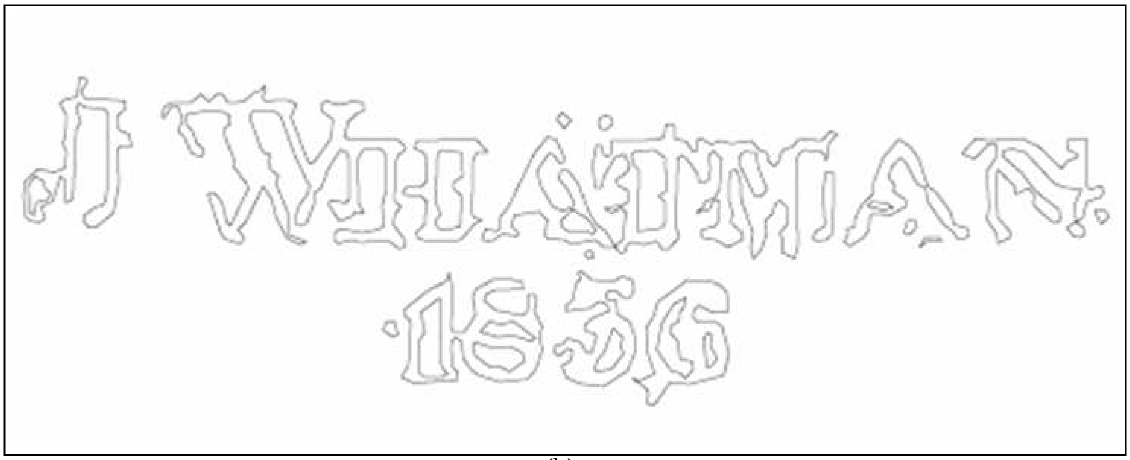

(b)

Figure 6.5: Vectorised watermark design, (a) without simplification (9332 vertices), (b) with simplification (826 vertices)

3. Disconnect: to isolate unnecessary and additional objects parts in order to remove them, by placing a line of background pixels between two selected points, see Figure 6.9.

4. Fill: to fill objects' holes by clicking on them; filling functionality is in Figure 6.10.

These functions are performed interactively with an easy-to-use graphical user interface. This editor is also equipped with basic functions such as: zoom, move, save, undo, redo, etc.

A further tool was built for vector editing (see Figure C.3 in Appendix C). Its main function is to remove unnecessary vector data points (vertices) and edges, and hence simplify the vector representation. This operation is performed by straightening the vector between two selected vertices; original data points are marked so that it is easier to select these points interactively. An explanation of the straightening process is illustrated in

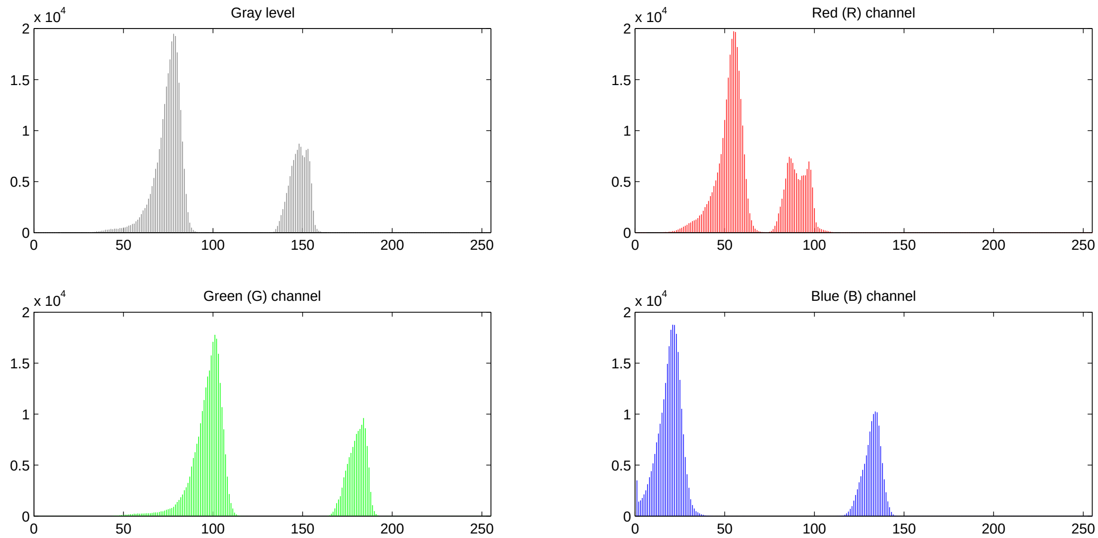
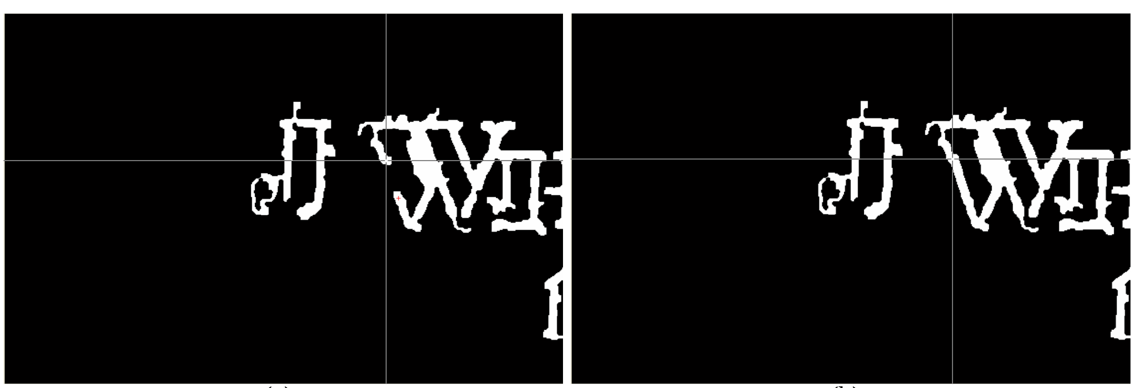


Figure 6.6: Histogram distribution of image grey level, and the RGB channels in separate plots

(a) (b)

Figure 6.7: Image editor functionalities, (a) before, (b) and after 'remove'



(a) (b)

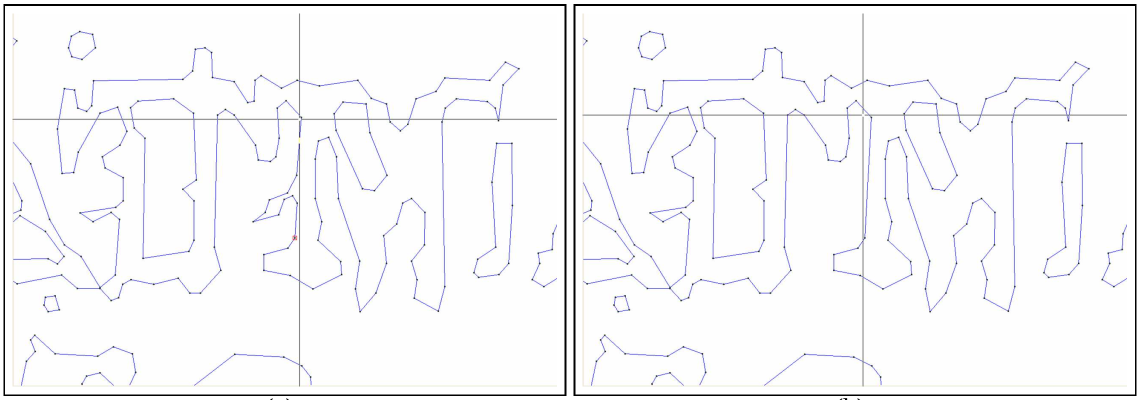
Figure 6.8: Image editor functionalities, (a) before, (b) and after 'connect'

(a) (b)

Figure 6.9: Image editor functionalities, (a) before, (b) and after ‘disconnect’

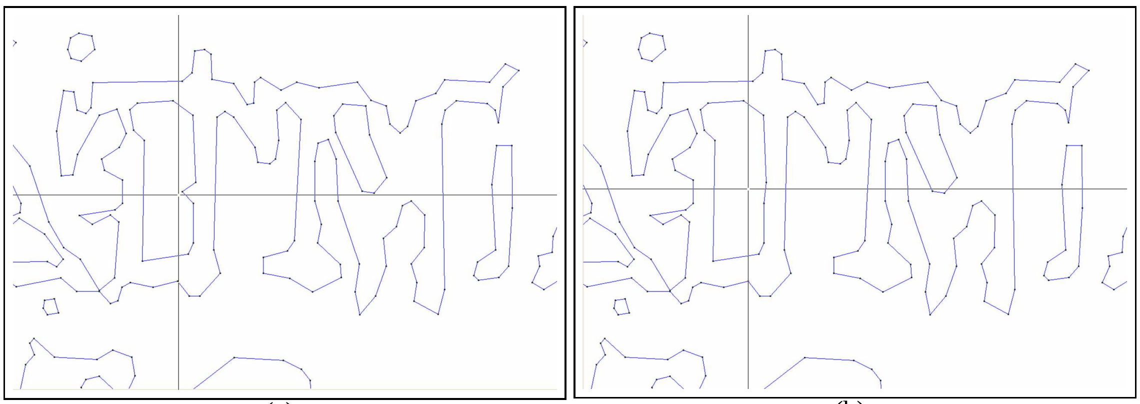
(a) (b)

Figure 6.10: Image editor functionalities, (a) before, (b) and after ‘fill’



(a) (b)

Figure 6.11: Vector editor functionalities, (a) before, (b) and after ‘straighten’



(a) (b)

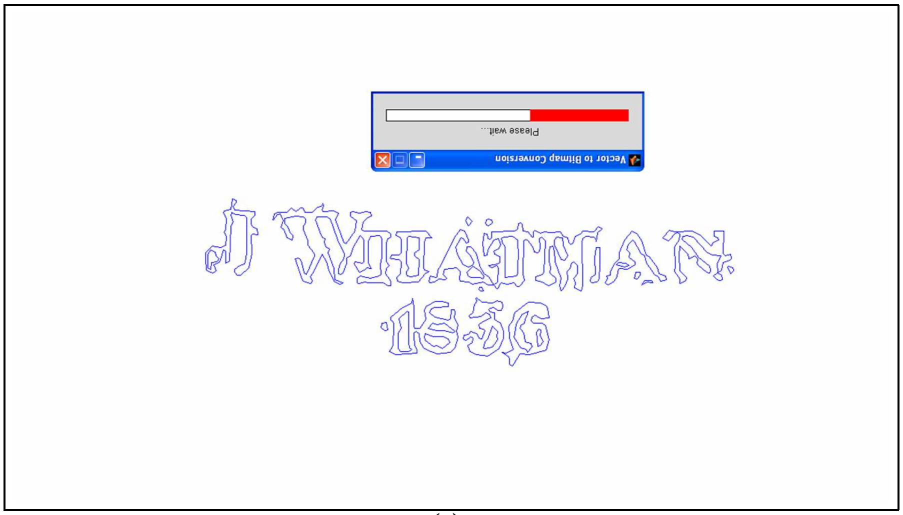
Figure 6.12: Vector editor functionalities, (a) before, (b) and after ‘remove’

Figure 6.11.

Another vector function is ‘Remove’, which works by interactively selecting the data point to be removed; an example is shown in Figure 6.12.

This vector editing tool can also change vector attributes, such as filling and stroke colours, and stroke width. A Vector-to-Bitmap conversion tool has also been implemented which converts the current vector to a bit-map image file; see Figure 6.13 for an example. The vector editor tool is also supported with basic functions as in the image editor tool.

Two further tools were built to view images and vectors (fully illustrated in Figures C.4 and C.5 in Appendix C). These include basic viewing facilities such as: browse, move, zoom, and save. The vector viewer can display SVG vectors without the need of external applications or plug-ins, while Internet browsers need a special plug-in [2] to view this vector format.



(a)

(b)

Figure 6.13: Vector-to-Bitmap conversion tool, (a) before, (b) and after conversion – illustrations are flipped for better watermark display

6.4 Evaluation

In all such processes, it is important to devise criteria to judge the quality of results after post-processing: an evaluation is necessary to determine to what extent the design was successfully extracted. Sometimes a ground truth of the watermark is available (for example, it may be found in one of the online databases, e.g. [4, 42, 48, 52, 69, 88, 98, 153, 156]), but if this is not an option we might, with comprehensive knowledge of the data, draw an exact image of the watermark design, and then compare it with the extracted one. In this procedure, the 'standard' so derived may well not be optimal, simply because it is drawn manually. Nevertheless, we contend it will be acceptable in the circumstances of our previous knowledge of the watermark designs. Results were also inspected by other users to judge accuracy and quality by eye.

To provide a basis for comparison, we asked six users to perform a manual tracing of watermark patterns from the input backlit images used in extraction to be the tracing source. Tracing is done digitally using a computer mouse, and ‘Paint Shop Pro’ imaging software [32]. The chosen users are experts in tracing by mouse, and familiar with this imaging software, and so we are confident the results are good enough to act as the basis of such a comparison.

Different watermark patterns were traced and compared with our extracted results: Figures 6.15 – 6.19 illustrate different watermark designs, along with the extracted and traced patterns. Similarity measures are in Table 6.1, and plotted in Figure C.9 in Appendix C.

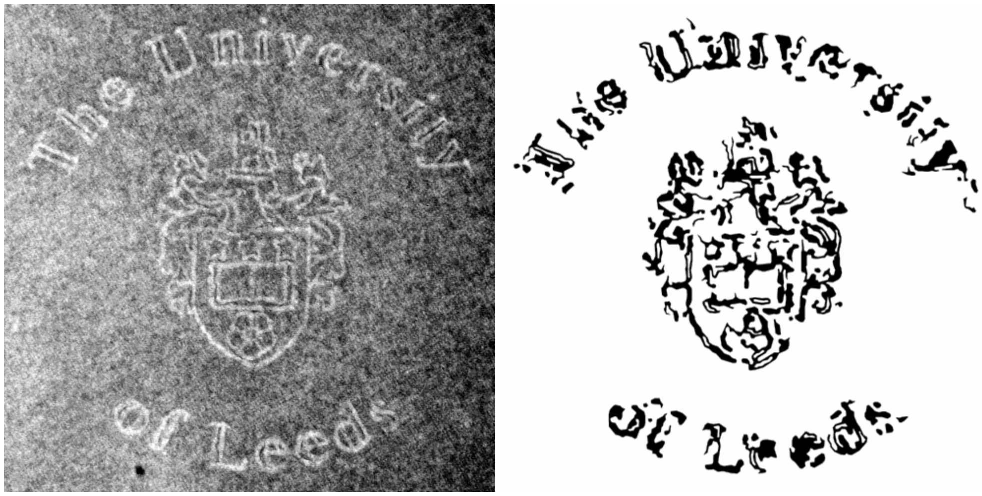
The similarity comparison is performed on a pixel-by-pixel logical AND basis: that is, similarity is counted if corresponding pixels in two designs are both white or black.

Table 6.1: Similarity comparison of extracted and traced watermark patterns (%)

Watermark	**Pattern (1)**	**Pattern (2)**	**Pattern (3)**	**Pattern (4)**	**Pattern (5)**
Extracted	90.1	87.5	90.3	82.3	68.4
Traced (1)	89.6	86.7	90.9	86.7	70.6
Traced (2)	87.8	82.6	87.6	83.3	56.7
Traced (3)	88.1	82.1	89.5	86.7	69.8
Traced (4)	89.4	84.1	91.0	86.0	65.1
Traced (5)	89.2	88.4	92.7	88.9	72.6
Traced (6)	92.5	88.2	92.5	89.0	71.0

The similarity table shows that in raw numerical terms, our extracted results are comparable and sometimes better than traced designs. Some of the traced designs were very good due to the accuracy of users, as shown in the last two rows of the table: users are more successful in tracing textual watermark patterns. On the other hand, our approach showed good results for extracting watermark drawings for some inputs, as illustrated in Figure 6.14.

We also considered a more qualitative criterion to decide whether an extracted watermark pattern is ‘good’ or not. We asked different users to judge (by eye) the goodness of an extracted pattern – this criterion is based on the original and extracted patterns only. As a result, all extracted patterns were accepted as ‘good’ except pattern (5) in Figure 6.19, which lacks much detail. This criterion verifies the usability of our extracted patterns.



(a) (b)

Figure 6.14: Input backlit image (enhanced for display) and extracted watermark design

6.5 Conclusion

This Chapter presented post-processing operations to convert the extracted bit-mapped watermark pattern to a vector graphics representation which can be zoomed at large scales and printed at high resolutions without any loss in detail.

Tools with graphical user interfaces were also presented to aid further interactive editing and enhancements, especially to users who are not experts in image and vector processing and programming. These tools can be helpful in enhancing the extracted watermarks, including the removal of residual noise features, and completing missing parts of the extracted designs interactively.

We presented an evaluation of the approach discussed in Chapter 4 and continued in this Chapter. We evaluated the approach quantitatively (by devising a similarity measure) and qualitatively (by judging by eye). Results of similarity comparisons show that extracted patterns are comparable and sometimes better than traced designs, which proves the potential applicability of the approach.

Users found tracing of textual watermarks easier than drawings; on the other hand, our approach showed promising results on both textual and geometrical patterns. Qualitative criteria were effective in deciding if extracted patterns are ‘good’ or not, and proved the viability of the approach.

However, the extracted vector designs are still far from perfect in their resemblance to original shapes, which are formed by twisted wires. Furthermore, the standard used in evaluation may not be optimal because it is manually drawn. Further, this approach is limited to data of the kind presented in Sections 3.1.1 and 3.1.2.

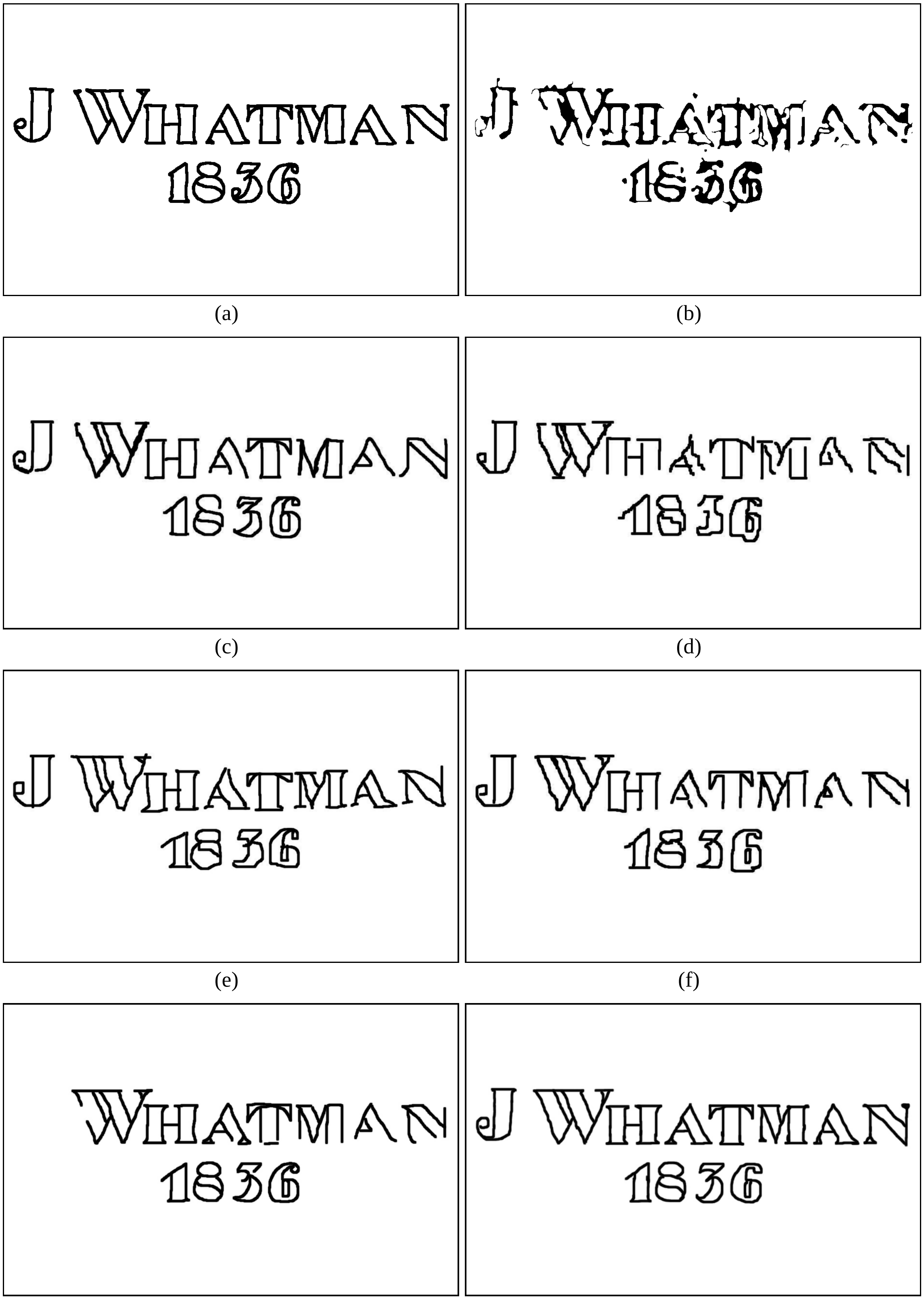
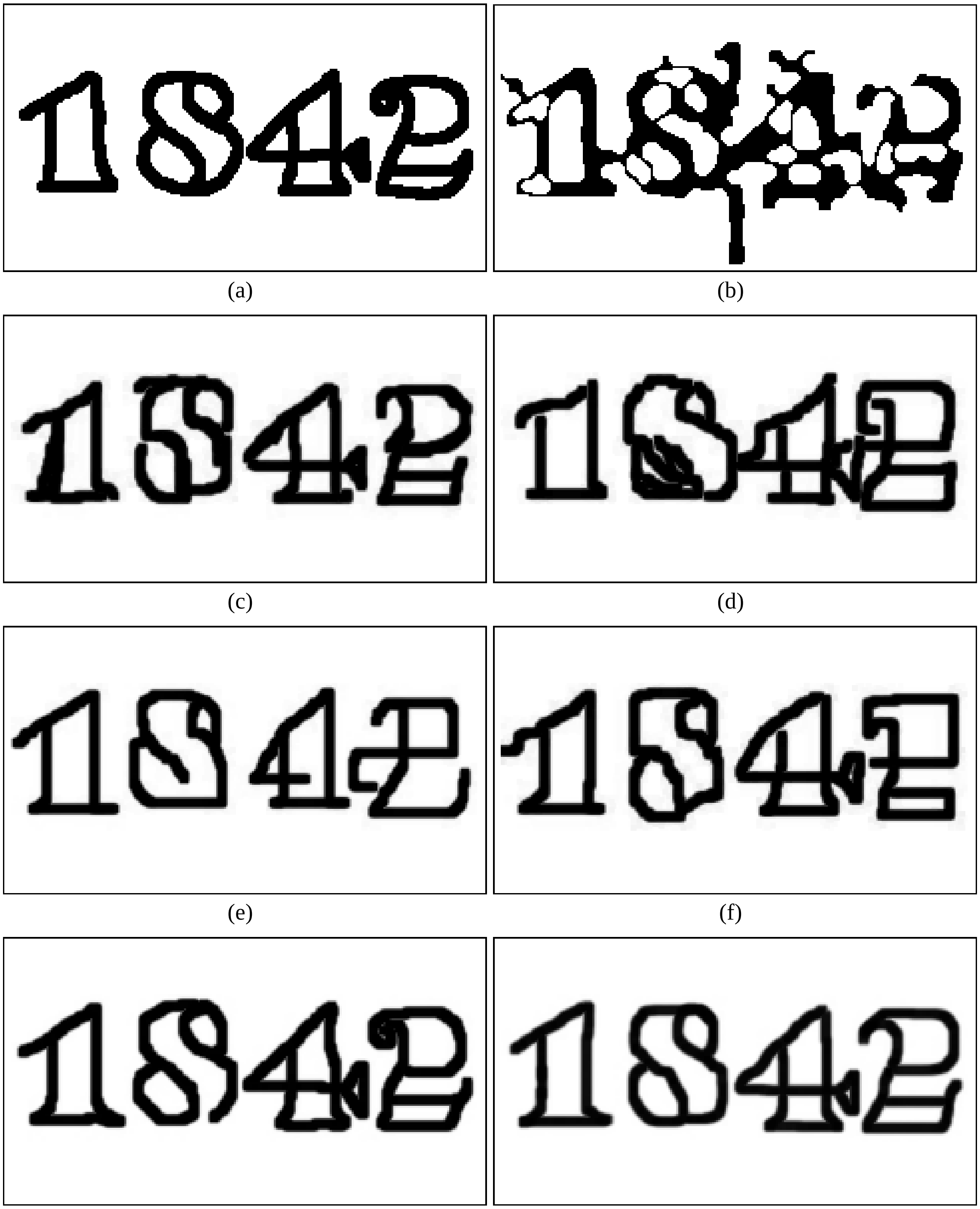


Figure 6.15: (a) Watermark pattern (1), (b) Extracted design, (c) – (h) Traced designs



(a) (b) (c) (d) (e) (f) (g) (h)

Figure 6.16: (a) Watermark pattern (2), (b) Extracted design, (c) – (h) Traced designs

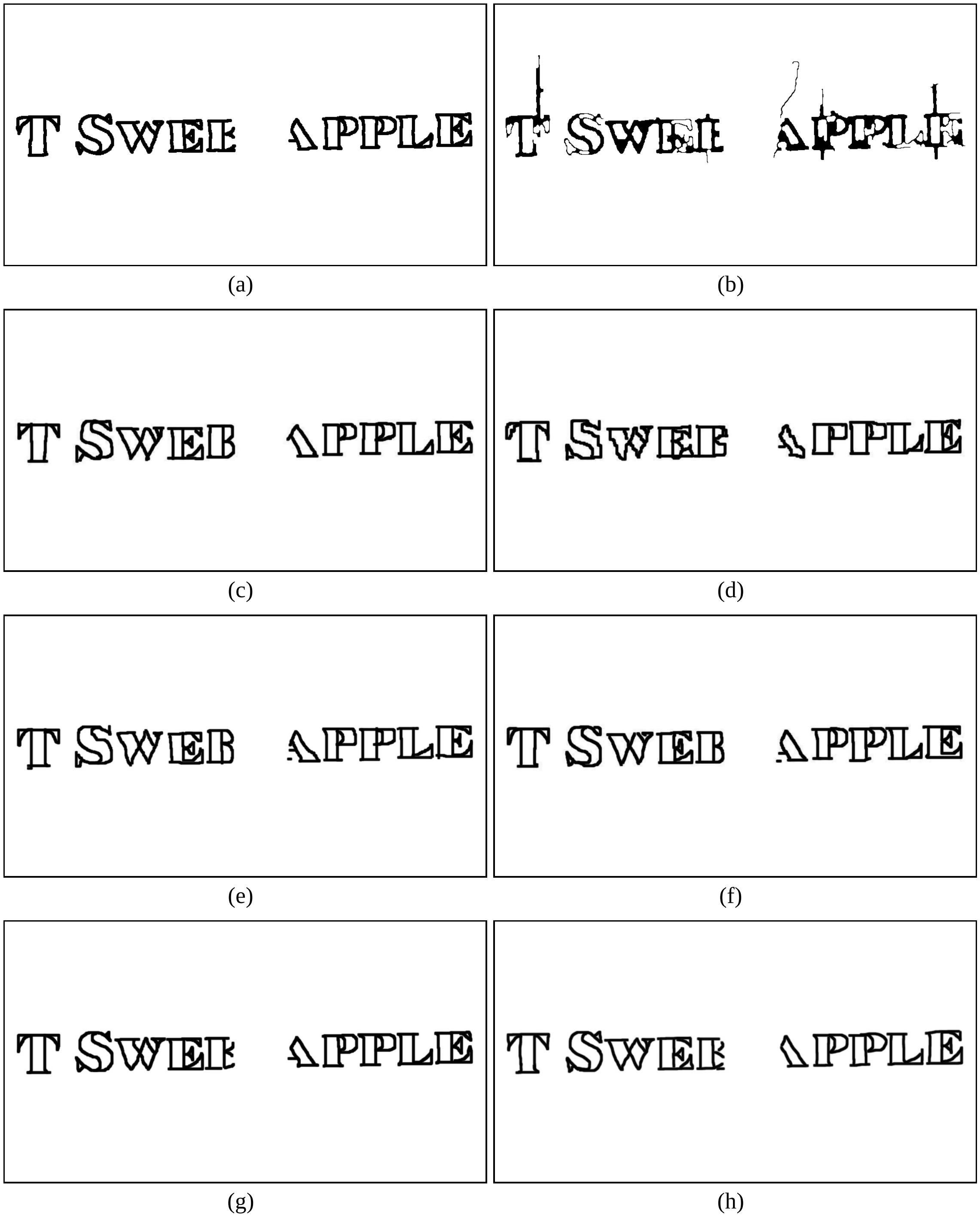


Figure 6.17: (a) Watermark pattern (3), (b) Extracted design, (c) – (h) Traced designs

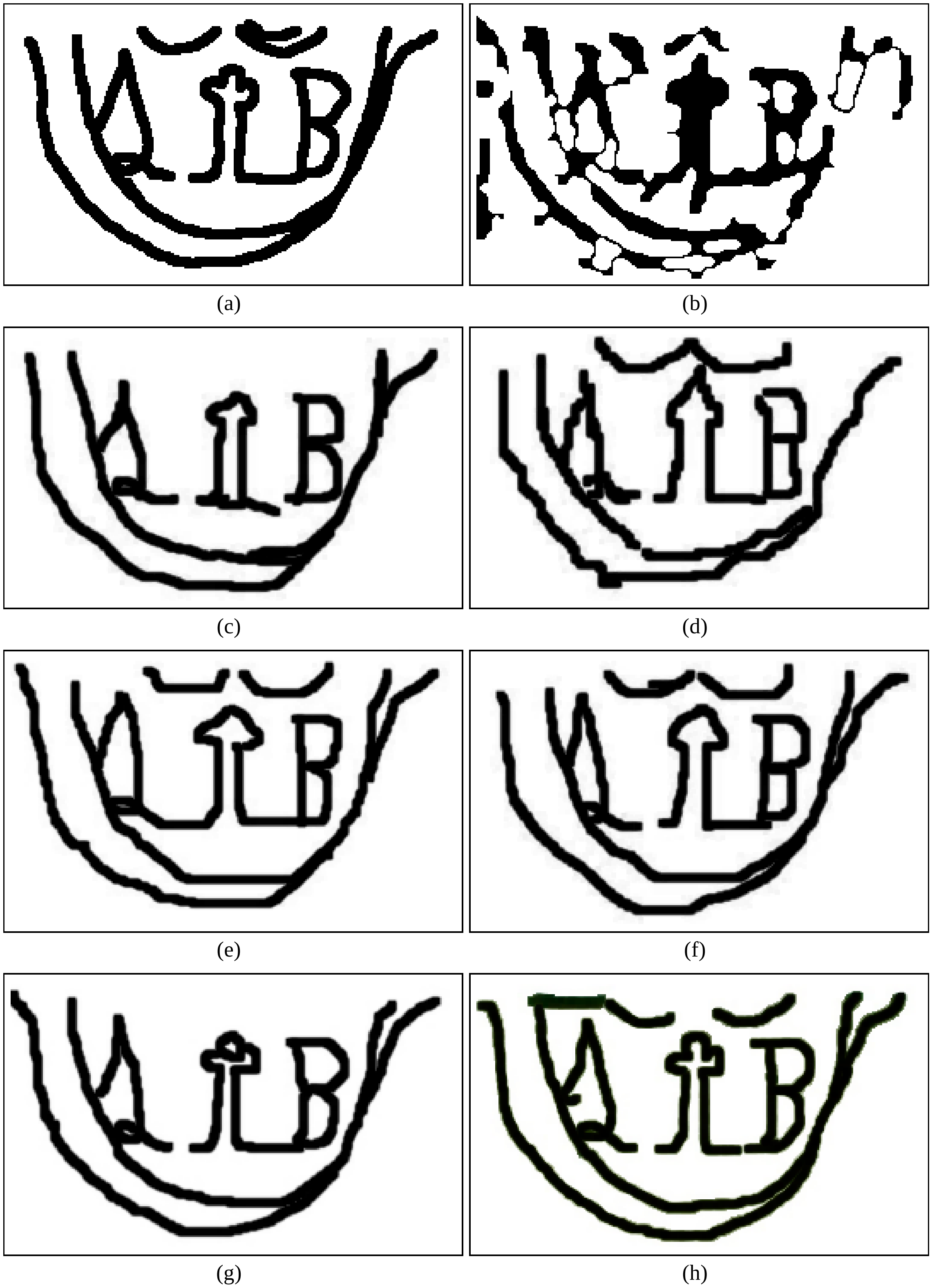
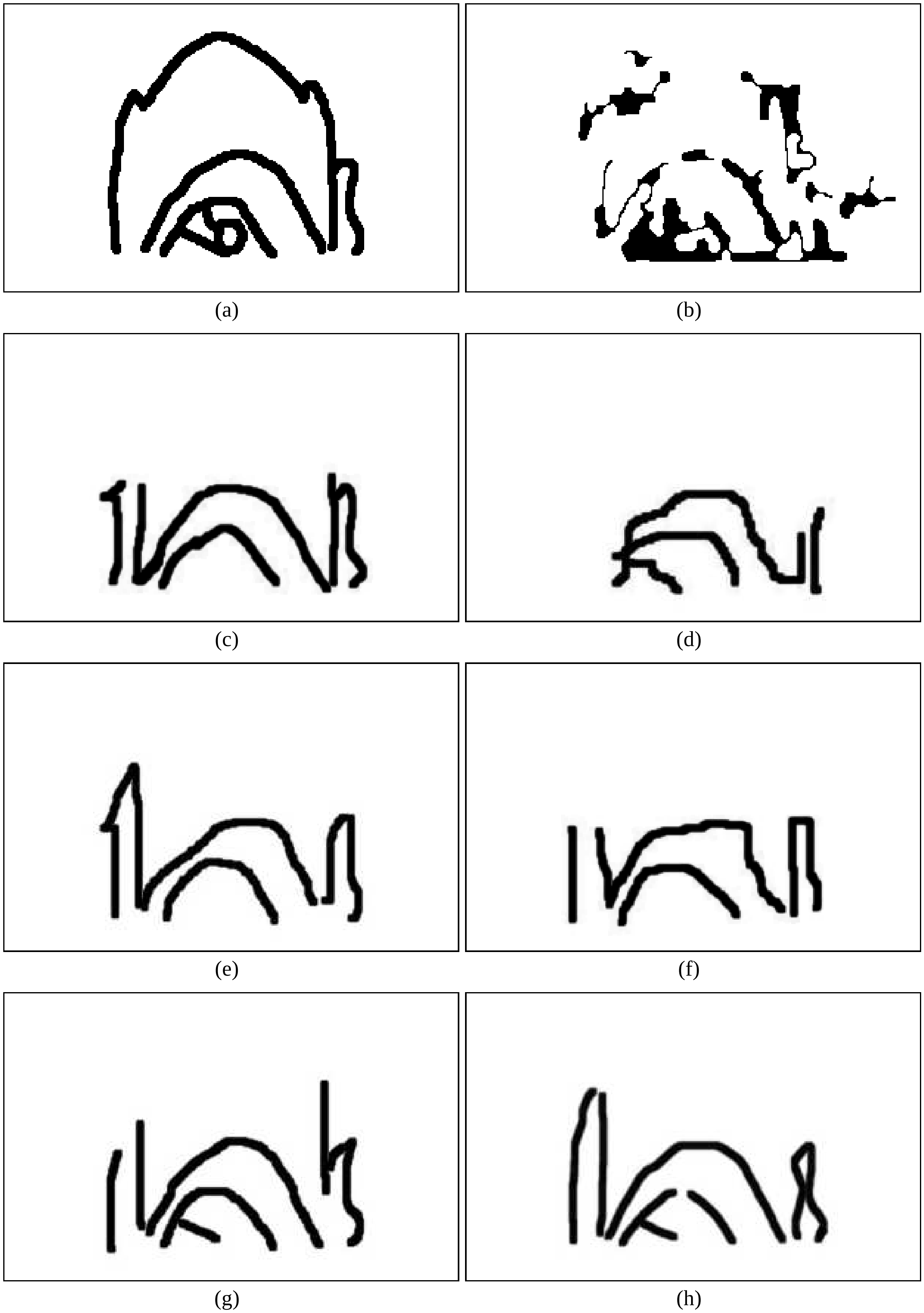


Figure 6.18: (a) Watermark pattern (4), (b) Extracted design, (c) – (h) Traced designs



(a) (b) (c) (d) (e) (f) (g) (h)

Figure 6.19: (a) Watermark pattern (5), (b) Extracted design, (c) – (h) Traced designs

Chapter 7

Conclusions and Future Directions

7.1 Summary of work

This thesis presented two different approaches to locate and extract watermarks in paper:

The bottom-up approach presented a prototype to extract paper watermarks using a sequence of image processing algorithms. This approach pre-processes images to remove interference and highlight the watermark, followed by segmentation, which achieves localisation and extraction of watermark patterns and chain lines. This approach was evaluated with human opinion: results of similarity comparisons are good, which proves the potential applicability of the approach. Extracted designs from the approach were exported in vector form, which can be simplified, zoomed at large scales and printed at high resolutions without loss in detail.

The top-down (modelling back-lighting) approach presented a model-based technique to locating watermarks in more difficult manuscripts; it managed to remove recto material successfully, and developed a statistical approach to locate watermark fragments from a known lexicon. Results show an excellent record of retrieval. The approach was extended to aggregate similar designs from different documents which enhanced watermark detail, highlighted chain lines, and distinguished 'twin' from 'identical' watermarks.

The bottom-up approach used only the backlit (transmitted) image for processing, while the modelling approach requires both reflected and transmitted images. These approaches can handle both types of paper – laid and wove – and worked well with wire watermarks of different shapes, including geometrical patterns. These approaches covered a wide range of manuscripts of various characteristics, including paper thickness, watermark visibility, noise distribution (paper structure, background illumination, etc.), and recto and verso inscription of varying thickness. Sample datasets and results were presented for each approach. Furthermore, these approaches can handle digitised images of dynamic resolution.

This research study succeeded in achieving its objectives. The thesis contributions may be summarised as:

Wider accessibility and distribution: This research will assist easier and wider accessibility of valuable historical manuscripts for scholars. This was achieved by establishing web-archives of the manuscripts used in the study [76, 77]. This prototype repository contains 18th and 19th century beautifully handwritten documents: two complete copies of the Qur'ān and an Islamic Prayer. Other manuscripts were from the works of Henry Litolff [14].

Preservation: Manuscripts were digitised using a back-lighting capturing system that captures not only the paper surface, but also the contents hidden beneath the surface of the paper, in particular the watermark designs. Digital preservation of these artefacts is important, particularly for collections that are fragile, which may suffer paper decay issues.

Interference removal: Approaches developed in this thesis managed to minimise different kinds of interference caused by writing on front (recto) and back (verso). In addition, there are often paper defects such as folding marks, paper texture, etc. The bottom-up approach removed this interference using various morphological operations, while the top-down approach modelled the effect of back-lighting. Both approaches managed such interference successfully.

Adaptive parameter selection: Both approaches considered dynamic adaptation of various processes to automatically determine optimal parameter values, including morphological operations, clustering and signal thresholding, and we have considered robust means of choosing these.

Chain lines detection: This research project has the ability to detect and extract chain lines, which appear as vertical lines in paper. This process can provide us with various measures, such as the distance between sequential lines, chain line orientation, thickness of lines and the number of chain lines in the paper.

Enhancing watermark details: Similar watermarks existing in different documents can be combined together to provide better detailed features. This is possible since watermarks can be distinguished and retrieved with their exact location in documents. This operation is important as it can reveal subtle details in designs that are difficult to observe in single watermark designs.

Distinguish 'identical' from 'twin' watermarks: This project can be used to differentiate between similar watermarks and classify them as 'identical' or 'twin', since it may be easy to identify the differences between designs when aggregated together.

Interactive interfaces: The project is also built with easy-to-use interactive tools, which allow different users to use the approaches without any difficulty or need of programming skills.

We believe that this research displays advantages in paper and watermark studies over many existing approaches due to its simplicity and usability. It will help in studying and understanding the materials and the structure of valuable historical manuscripts.

7.2 Capabilities and possible improvements

The work presented in this thesis has its weaknesses, but these can be improved in many ways. We summarise limitations of the research approaches, and provide possible improvements:

Adaptive parameter selection: We presented a number of algorithms to determine optimal parameter selection in various operations used in both approaches. However, some processes, e.g. edge detection in the bottom-up approach, still need manual parameter adjusting. This may be improved by providing more assumptions when selecting these parameters. For example, when selecting parameters for edge detection, we already know that the watermark feature is among the brightest (highest intensities) features in the image: in this case it is wise to choose high parameter values.

In the approach of modelling back-lighting, the choice of the parameter λ in Equation 5.10 (used to recognise different scale of distance and angle) was not explored deeply. $\lambda = 1$ gave satisfactory results in our datasets. Perhaps testing with more datasets and a deep analysis and understanding of this parameter selection will provide better results.

Characteristics of manuscripts: The bottom-up approach is limited to datasets characterised by non-uniform background and thin pen stroke used in writing. Datasets used are thin paper, with the watermark design clearly visible. This approach did not succeed in processing more difficult datasets, such as the Qur'ān manuscripts. However, this could be improved by enhancing the image processing operations used. Adding more assumptions to recognise and remove noise features could be effective.

Automatic watermark location: The modelling approach succeeded in retrieving watermark designs by selecting a *part* of a watermark, but still this requires foreknowledge of the watermark design – or at least *part* of it – in order to proceed. It is possible to propose an automatic approach to locate these designs without any previous knowledge of their structure.

Automatic location is possible if 'hidden' watermark materials can be completely separated from recto and verso materials. A better understanding of the exact structure of these designs is also useful, such as their feature width (in pixels), the change of intensity value between the watermark pixels and their surrounding neighbours, or knowledge of the watermark shape itself. Since it is built from wires, which form lines and curves, all of these characteristics will be helpful in identifying watermarks in paper automatically.

Perfect shape extraction: The extracted patterns using the bottom-up approach, which are further exported to vector form, show good results. The project is equipped with the necessary tools that aid users to complete these designs interactively. However, these vector patterns are still far from perfect in their resemblance to original shapes, formed by twisted wires. This may be improved by establishing a known lexicon of these designs: with help from pattern matching techniques, it will be possible to recognise and complete the messing design parts using that lexicon. Related literature in this field can be found in [49, 90, 155].

Linearity of modelling back-lighting: The model of back-lighting assumes a linear relationship (seen in Equation 5.2). Lighting effects are often subtle and it is most unlikely that the effect we observe will indeed be linear, but we proceeded with this simplification on the understanding that it is applied only to pixels that are 'similar', and in the ideal case identical. It is possible that trying the same approach with quadratic or cubic approximations may provide better models of back-lighting.

Evaluation: The bottom-up approach was evaluated quantitatively (by devising a similarity measure) and qualitatively (by judging by eye). Results of similarity comparisons show that our extracted results are comparable and sometimes better than traced designs, which proves the potential applicability of the approach. However, the standard used in evaluation may not be optimal because it is manually drawn. This can be improved by using the original patterns with no interference as a standard for evaluation. These may be found in the special collections located in libraries, or museums. They may also be located in watermark collections traced by popular historians, such as 'Les filigranes' by Briquet [21], more collections are in [29, 66, 71, 106, 118].

7.3 Future directions

Suggested future directions for this research study include improving the proposed approaches to avoid the limitations presented in Section 7.2. These improvements will provide more usability and simplicity for the study of paper and watermarks. Working on extended datasets may explore various enhancements.

Watermarks used in this thesis were line (wire) watermarks – we did not have the chance to study shadow (light and shade) watermarks which appear as dark and light areas in paper. Our approaches may locate and extract these patterns. However, some of the operations we used assume that features are (relatively) bright, which is the case of wire watermarks. In this case, these operations need to be improved to provide good localisation and extraction of this type. We believe that exploring shadow watermarks, or even better, the combined type (line and shadow watermarks combined in one paper sheet), is an encouraging way forward, and an important and under-explored area of study of paper watermarks.

This thesis presented a retrieval system for watermarks located in the Qur'ānic and Prayer manuscripts. Another future direction is to develop an approach to extract the patterns that exist in these manuscripts without any foreknowledge of their design.

This thesis used back-lighting acquisition to capture paper watermarks. Another direction is to explore other reproduction techniques, and investigate their usability in locating and extracting watermarks compared to our approaches. A thorough comparison would be essential in this case.

Bibliography

[1] Adobe Corporation. Adobe Photoshop software. `http://www.adobe.com/photoshop/`. Last accessed: 25th July 2008.

[2] Adobe Corporation. Adobe SVG Viewer software. `http://www.adobe.com/svg/`. Last accessed: 25th July 2008.

[3] AHDS (Arts and Humanities Data Service) project. `http://www.ahds.ac.uk/`. Last accessed: 25th July 2008.

[4] Robert W. Allison. An automated WWW search tool for papers and watermarks: The archive of papers and watermarks in greek manuscripts. In *Puzzles in Paper: Concepts in Historical Watermarks*, pages 201–210. Oak Knoll Books and the British Library, New Castle, UK, 2000. `http://abacus.bates.edu/Faculty/wmarchive/`, `http://www.watermarkarchive.org/`. Last accessed: 25th July 2008.

[5] Ross J. Anderson and Fabien A. P. Petitcolas. On the limits of steganography. *IEEE Journal of Selected Areas in Communications*, 16(4):474–481, 1998.

[6] Nancy Ash. Recording watermarks by beta-radiography and other means. In *The Book and Paper Group Annual*, volume 1. American Institute for Conservation of Historic and Artistic Works, 1982.

[7] Nancy E. Ash. Watermark research: Rembrandt prints and the development of a watermark archive. *The Paper Conservator*, 10:64–69, 1986.

[8] Nancy E. Ash and Shelley Fletcher. Watermarks in Rembrandt's prints: The use of watermarks to study the prints of an artist. In *Puzzles in Paper: Concepts in Historical Watermarks*, pages 57–66. Oak Knoll Books and the British Library, New Castle, UK, 2000.

[9] Vlad Atanasiu. Ad751 for laid lines density measurement. http://mywebpage.netscape.com/atanasiuvlad/ad751/. Last accessed: 25th July 2008.

[10] Vlad Atanasiu. BlueNile for laid lines suppression and extraction. http://mywebpage.netscape.com/atanasiuvlad/bluenile/. Last accessed: 25th July 2008.

[11] Vlad Atanasiu. Assessing paper origin and quality through large-scale laid lines density measurements. In *XXVI Congress of the International Paper Historians Association*, Rome and Verona, Italy, August 26-September 6 2002.

[12] V. V. Belov, V. A. Esipova, V. T. Kalaida, and V. M. Klimkin. Physical and mathematical methods for the visualisation and identification of watermarks. *Solanus*, 13:80–92, 1999.

[13] Bernstein - The Memory of Papers project. http://www.bernstein.oeaw.ac.at/. Last accessed: 25th July 2008.

[14] Ted M. Blair and Thomas Cooper. Litolff, Henry. *Grove Music Online*, http://www.grovemusic.com/shared/views/article.html?section=music.16780. Last accessed: 25th July 2008.

[15] Jonathan M. Bloom. *Paper before print: the history and impact of paper in the Islamic world*. Yale University Press, New Haven, London, 2001.

[16] Don Francisco de Bofarull Y Sans. *ANIMALS IN WATERMARKS*. Paper Publication Society, Hilversum, 1959.

[17] Peter Bower. *Turner's papers : a study of the manufacture, selection and use of his drawing papers 1787-1820*. Tate Gallery Publishing, London, UK, 1990.

[18] Peter Bower. *Turner's later papers : a study of the manufacture, selection, and use of his drawing papers 1820-1851*. Tate Gallery Publishing, London, UK, 1999.

[19] Charles F. Bridgman. Radiography of paper. *Studies in Conservation*, 10(1):8–17, February 1965.

[20] Charles F. Bridgman, Sheldon Keck, and Harold F. Sherwood. The radiography of panel paintings by electron emission. *Studies in Conservation*, 3(4):175–182, October 1958.

[21] Charles Moïse Briquet. *Les filigranes : dictionnaire historique des marques du papier dès leur apparition vers 1282 jusqu'en 1600*. Hiersemann, Leipzig, 1923.

[22] Charles Moïse Briquet. Notice sur le recueil de filigranes ou marques des papiers, presented at the Paris Exposition in 1900. In *Briquet's opuscula : the complete works of C.M. Briquet without Les filigranes*, pages 281–288. Paper Publication Society, Hilversum, 1955.

[23] von Carl Brockelmann. *Geschichte der arabischen litteratur*, volume I/II. E.J. Brill, Leiden, 1949.

[24] Adrian Brockett. Aspects of the physical transmission of the Qur'ān in 19th-century Sudan: Script, binding, decoration and paper. *Manuscripts of the Middle East*, 2:45–67, 1987.

[25] Cambridge University Press. The Cambridge edition of the works of Ben Jonson. `http://www.cambridge.org/uk/literature/features/cwbj/`. Last accessed: 25th July 2008.

[26] John Canny. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8(6):679–698, 1986.

[27] Ian Christie-Miller. Early paper project. `http://www.earlypaper.com/`. Last accessed: 25th July 2008.

[28] Ian Christie-Miller. Paperprint: Trapping the tome raiders. *International Journal of Micrographics and Optical Technology*, 22(4):2–4, 2004.

[29] William Algernon Churchill. *Watermarks in paper in Holland, England, France, etc. in the XVII and XVIII centuries and their interconnection*. M. Hertzberger, Amsterdam, 1935.

[30] CLIR (Council on Libraries and Information Resources). `http://www.clir.org/`. Last accessed: 25th July 2008.

[31] Confederation of Paper Industries (CPI). The papermaking machine. `http://www.paper.org.uk/info/process/machine.htm`. Last accessed: 25th July 2008.

[32] Corel Corporation. Paint Shop Pro software. `http://www.corel.com/`. Last accessed: 25th July 2008.

[33] Ingemar J. Cox, Matthew L. Miller, and Jeffrey A. Bloom. *Digital watermarking*. Morgan Kaufmann, San Francisco, London, 2002.

[34] Brigitte de La Passardière and Claire Bustarret. Profil: An iconographic database for modern watermarked papers. *Computers and the Humanities*, 36(2):143–169, 2002.

[35] Rolf Dessauer. Dylux, Thomas L. Gravell, and watermarks of stamps and papers. In *Puzzles in Paper: Concepts in Historical Watermarks*, pages 183–185. Oak Knoll Books and the British Library, New Castle, UK, 2000.

[36] DFG (Deutsche Forschungsgemeinschaft, German Research Foundation). `http://www.dfg.de/`. Last accessed: 25th July 2008.

[37] Jin Di. Paper and watermark digitization and analysis. Master's thesis, School of Computing, University of Leeds, 2004.

[38] DI.MU.SE project (Ministero per i Beni e le Attività Culturali and Palatina Library of Parma). `http://www.bibpal.unipr.it/bibliotecaPalatina/index.html`. Last accessed: 25th July 2008.

[39] DLF (Digital Library Federation), Digital library standards and practices. `http://www.diglib.org/standards.htm`. Last accessed: 25th July 2008.

[40] David H. Douglas and Thomas K. Peucker. Algorithms for the reduction of the number of points required to represent a digitized line or its caricature. *The Canadian Cartographer*, 10(2):112–122, 1973.

[41] DuPont Corporation. `http://www.dupont.com`. Last accessed: 25th July 2008.

[42] Dutch University Institute for Art History, Florence - International database of watermarks and paper used for prints and drawings c. 1450-1800. `http://www.iuoart.org/wmdb.htm`. Last accessed: 25th July 2008.

[43] John P. Eakins, A. Jean E. Brown, K. Jonathan Riley, and Richard Mulholland. Evaluating a shape retrieval system for watermark images. In *Proceedings of CHArt 17th Annual conference (CHArt2001), Digital Art History - A Subject in Transition: Opportunities and Problems*, volume 4, British Academy, November 27-28 2001.

[44] David Eberly. Polyline reduction. `http://www.geometrictools.com/Documentation/PolylineReduction.pdf`. Last accessed: 25th July 2008.

[45] Rifaat Y. Ebied and Michael J. L. Young. Some Maghribi manuscripts in the Leeds university collection. *Semitic Studies*, 21(1-2):109–119, 1976.

[46] Dexter Edge. The digital imaging of watermarks. *Computing in Musicology*, 12:261–274, 2001.

[47] Brian S. Everitt, Sabine Landau, and Morven Leese. *Cluster Analysis*. Hodder Arnold, London, UK, 4th edition, 2001.

[48] EVTEK Paper Identification Database. `http://conservation.evtek.fi/`. Last accessed: 25th July 2008.

[49] Rob Fergus, Pietro Perona, and Andrew Zisserman. Object class recognition by unsupervised scale-invariant learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '03)*, pages 264–271, Madison, Wisconsin, USA, June 18-20 2003. IEEE Computer Society Press.

[50] Fraunhofer-Institute for Wood Research – Wilhelm-Klauditz-Institut (WKI). `http://www.wki.fraunhofer.de/english/index.html`. Last accessed: 25th July 2008.

[51] FUJIFILM FinePix S1 Pro, camera and shooting software manuals. `https://www.fujifilm.ca/documents/Fuji_S1Pro_CM.pdf`, `http://www.fujifilmusa.com/shared/bin/FX-S1_PSS_Manual.pdf`. Last accessed: 25th July 2008.

[52] David L. Gants. A digital catalogue of watermarks and type ornaments used by William Stansby in the printing of *The Workes of Beniamin Jonson* (London: 1616). `http://www.iath.virginia.edu/gants/`. Last accessed: 25th July 2008.

[53] David L. Gants. The early English booktrade database (EEBD). `http://www.lib.unb.ca/Texts/Gants/EEBD/`. Last accessed: 25th July 2008.

[54] David L. Gants. Pictures for the page: Techniques in watermark reproduction, enhancement and analysis, April 23 1994. Delivered at the Annual meeting of the

Bibliographical Society of the University of Virginia, McGregor Room, Alderman Library.

[55] David L. Gants. The application of digital image processing to the analysis of watermarked paper and printers' ornament usage in early printed books. In *New ways of looking at old texts, II : papers of the Renaissance English Text Society, 1992-1996*, pages 133–147. Medieval & Renaissance Texts & Studies, in conjunction with Renaissance English Text Society, Tempe, Arizona, USA, 1998.

[56] David L. Gants. The CUP Ben Jonson: Ruminations on the electronic edition. *Ben Jonson Journal*, 5:271–281, 1998.

[57] David L. Gants. Patterns of paper use in the *Workes* of Beniamin Jonson (William Stansby, 1616). *Studies in Bibliography*, 51:127–153, 1998. `http://etext.lib.virginia.edu/bsuva/sb/toc/sib51toc.htm`. Last accessed: 25th July 2008.

[58] David L. Gants. The printing, proofing and press-correction of Ben Jonson's folio *Workes*. In *Re-presenting Ben Jonson : text, history, performance*, pages 39–58. Macmillan, London, UK, 1999.

[59] David L. Gants. Identifying and tracking paper stocks in early modern London. *Papers of the Bibliographical Society of America*, 94:531–540, 2000.

[60] David L. Gants. A quantitative analysis of the London book trade 1614-1618. *Studies in Bibliography*, 55:185–213, 2002. `http://etext.lib.virginia.edu/bsuva/sb/toc/sib55toc.htm`. Last accessed: 25th July 2008.

[61] Gabriel García. Collection of paper watermarks. `http://www.watermarks.info/indexi.htm`. Last accessed: 25th July 2008.

[62] Gene H. Golub and Charles F. van Loan. *Matrix computations*. Johns Hopkins University Press, Baltimore, London, 3rd edition, 1996.

[63] Rafael C. Gonzalez and Richard E. Woods. *Digital Image Processing*. Prentice Hall, Upper Saddle River, N.J., USA, 2nd edition, 2002.

[64] Thomas L. Gravell. A new method of reproducing watermarks for study. *Restaurator*, 2:95–104, 1975.

[65] Thomas L. Gravell. The wizard of watermarks. *Du Pont Magazine*, 84(1):4–6, 1990.

[66] Thomas L. Gravell and George Miller. *A catalogue of foreign watermarks found on paper used in America, 1700-1835*. Garland Publishing, New York, USA, 1983.

[67] Robert M. Haralick and Linda G. Shapiro. *Computer and Robot Vision*, volume 2. Addison-Wesley, Boston, MA, USA, 1992.

[68] Wolfgang Haupt. Wasserzeichenwiedergabe in schwierigen Fällen. *Restauro*, 87:38–43, 1981.

[69] Hauptstaatsarchiv Stuttgart. Wasserzeichenkartei Piccard - Piccard watermark collection. `http://www.piccard-online.de/`. Last accessed: 25th July 2008.

[70] Gabriel Heaton. LIMA: Literary Manuscript Analysis. `http://www2.warwick.ac.uk/fac/arts/ren/projects/publications/lima/`. Last accessed: 25th July 2008.

[71] Edward Heawood. *Watermarks mainly of the 17th and 18th centuries*. Paper Publications Society, Hilversum, 1950.

[72] María Herrero, Jóse García, Arturo Saez, and Rafael Lucia. Visualisation: Using SVG for the representation of interactive maps. In *Proceedings of International Conference on Multimedia, Image Processing, and Computer Vision (IADAT-micv2005)*, pages 156–160, Madrid, Spain, March 30-April 1 2005. International Association for the Development of Advances in Technology (IADAT).

[73] John Hershberger and Jack Snoeyink. Speeding up the Douglas-Peucker line-simplification algorithm. In *Proceedings of 5th International Symposium on Spatial Data Handling*, volume 1, pages 134–143, Charleston, SC, USA, 1992.

[74] Herzog Anton Ulrich-Museum. `http://www.museum-braunschweig.de/`. Last accessed: 25th July 2008.

[75] Hazem Hiary. Using fractal coding techniques in data hiding. Master's thesis, King Abdullah II School for Information Technology, University of Jordan, November 2003.

[76] Hazem Hiary, Roger Boyle, and Kia Ng. Digitised Arabic texts from the University of Leeds. `http://www.comp.leeds.ac.uk/arabictexts/`. Last accessed: 25th July 2008.

[77] Hazem Hiary and Kia Ng. Paper-based watermark extraction with image processing. `http://www.icsrim.org.uk/watermark`. Last accessed: 25th July 2008.

[78] Dard Hunter. *Papermaking : the history and technique of an ancient craft*. Dover Publications, New York, USA, 1978.

[79] INTAS project 00-0081. A distributed database and processing system for watermarks. `http://www.viskom.oeaw.ac.at/~weng/intas0081/intas0081.htm`. Last accessed: 25th July 2008.

[80] International Association of Paper Historians (IPH). International Standard for the Registration of Paper with or without Watermarks, English Version 2.0, 1997. `http://www.paperhistory.org/standard.htm`. Last accessed: 25th July 2008.

[81] Jean Irigoin. La datation par les filigranes du papier. *Codicologica, Les matriaux du livre manuscrit*, 5:9–36, 1980.

[82] Carlo James, Caroline Corrigan, Marie Christine Enshaian, and Marie Rose Greca. *Old master prints and drawings : a guide to preservation and conservation*. Amsterdam University Press, Amsterdam, 1997. translated and edited by Marjorie B.Cohn.

[83] Neil F. Johnson. *Information hiding : steganography and watermarking, attacks and countermeasures*. Kluwer Academic, Dordrecht, London, 2002.

[84] Neil F. Johnson and Sushil Jajodia. Exploring steganography: Seeing the unseen. *IEEE Computer*, 31(2):26–34, 1998.

[85] Kaiser Fototechnik. `http://www.kaiser-fototechnik.de`. Last accessed: 25th July 2008.

[86] Andrey Karnaukhov, Igor Aizenberg, Alois Haidinger, Victor Karnaukhov, Nikolai Merzlyakov, Olga Milyukova, and Emanuel Wenger. Digital restoration of watermark images. In *Proceedings of EVA'01 Moscow*, pages 196–199, Centre PIC of Ministry Culture of Russia, STG, Moscow, 2001.

[87] Stefan Katzenbeisser and Fabien A. P. Petitcolas. *Information Hiding Techniques for Steganography and Digital Watermarking*. Artech House, Boston, USA, 2000.

[88] Koninklijke Bibliotheek – National library of the Netherlands. Watermarks in incunabula printed in the low countries (WILC). `http://watermark.kb.nl/`. Last accessed: 25th July 2008.

[89] Theo Laurentius, Harry van Hugten, Erik Hinterding, and Jan Piet Kok. Het Amsterdamse onderzoek naar Rembrandts papier: radiografie van de watermerken in de etsen van Rembrandt. *Bulletin van het Rijksmuseum*, 40:353–384, 1992. English summary is in pages 417-420.

[90] Bastian Leibe, Ales Leonardis, and Bernt Schiele. Combined object categorization and segmentation with an implicit shape model. In *Proceedings of the ECCV '04 Workshop on Statistical Learning in Computer Vision*, pages 17–32, Prague, Czech Republic, May 15 2004. Springer-Verlag.

[91] Jae S. Lim. *Two-dimensional signal and image processing*. Prentice Hall, Englewood Cliffs, N.J., USA, 1990.

[92] Edo G. Loeber. *Paper mould and mouldmaker*. Paper Publications Society, Amsterdam, 1982.

[93] Peter Meinlschmidt. Original or fake? *Fraunhofer – Research news*, July 2007.

[94] Irwin and Marylees Miller. *John E. Freund's mathematical statistics with applications*. Pearson Prentice Hall, Upper Saddle River, NJ, 7th edition, 2004.

[95] MINERVA project. `http://www.minervaeurope.org/`. Last accessed: 25th July 2008.

[96] Daniela Moschini. La marca d'acqua: A system for the digital recording of watermarks. In *Puzzles in Paper: Concepts in Historical Watermarks*, pages 187–192. Oak Knoll Books and the British Library, New Castle, UK, 2000.

[97] Daniel W. Mosser, Michael Saffle, and Ernest W. Sullivan, II. *Puzzles in Paper: Concepts in Historical Watermarks (Papers from the 1996 International Conference on Watermarks at Roanoke, Virginia)*. Oak Knoll Books and the British Library, New Castle, UK, 2000.

[98] Daniel W. Mosser and Ernest W. Sullivan, II. The Thomas L. Gravell watermark archive on the Internet. In *Puzzles in Paper: Concepts in Historical Watermarks*, pages 211–228. Oak Knoll Books and the British Library, New Castle, UK, 2000. `http://www.gravell.org/`. Last accessed: 25th July 2008.

[99] MUSICNETWORK EC IST project. `http://www.interactivemusicnetwork.org/wg_imaging/`. Last accessed: 25th July 2008.

[100] Hanns Peter Neuheuser, Volker Märgner, and Peter Meinlschmidt. Wasserzeichendarstellung mit Hilfe der Thermographie. *ABI TECHNIK*, 25(4):266–278, 2005.

[101] Kia Ng. Interdisciplinary Centre for Scientific Research in Music (ICSRiM). `http://www.icsrim.org.uk/`. Last accessed: 25th July 2008.

[102] Kia Ng, Di Jin, Richard Sage, and Bee Ong. Paper digitization and analysis. In *Fourth MUSICNETWORK Open Workshop, Integration of Music in Multimedia Applications*, Universitat Pompeu Fabra, Barcelona, Spain, 15-16 September 2004.

[103] Kitty Nicholson. Making watermarks meaningful: Significant details in recording and identifying watermarks. In *The Book and Paper Group Annual*, volume 1. American Institute for Conservation of Historic and Artistic Works, 1982.

[104] Courtney R. H. Owen. Watermarks and watermarks lesson at the American Museum of Papermaking (AMP), of the Institute of Paper Science and Technology (IPST). `http://www.ipst.gatech.edu/amp/education/watermark/index.htm`. Last accessed: 25th July 2008.

[105] Fabien A. P. Petitcolas, Ross J. Anderson, and Markus G. Kuhn. Information hiding, a survey. *IEEE Proceedings - Special Issue on Protection of Multimedia Content*, 87(7):1062–1078, July 1999.

[106] Gerhard Piccard. *Die Wasserzeichenkartei Piccard im Hauptstaatsarchiv Stuttgart: Findbuch*. Kohlhammer, Stuttgart, 1961-96.

[107] PULMAN project. `http://www.pulmanweb.org/`. Last accessed: 25th July 2008.

[108] Christian Rauber, Joe Ó Ruanaidh, and Thierry Pun. Secure distribution of watermarked images for a digital library of ancient papers. In *Proceedings of the Second ACM International Conference on Digital Libraries (DL'97)*, pages 123–130, Philadelphia, Pennsylvania, USA, July 23-26 1997.

[109] Christian Rauber, Peter Tschudin, and Thierry Pun. Retrieval of images from a library of watermarks for ancient paper identification. In *Proceedings of EVA'97 -*

Elektronische Bildverarbeitung und Kunst, Kultur, Historie, volume vol. 14, Berlin, Germany, November 12-14 1997.

[110] Christian Rauber, Peter Tschudin, Serguei Startchik, and Thierry Pun. Archival and retrieval of historical watermark images. In *Proceedings of the International Conference on Image Processing (ICIP)*, pages 773–776, Lausanne, Switzerland, September 16-19 1996. IEEE Computer Society Press.

[111] K. Jonathan Riley and John P. Eakins. Content-based retrieval of historical watermark images: I-tracings. In *Proceedings of the International Conference on Image and Video Retrieval, Lecture Notes in Computer Science*, volume 2383, pages 253–261, London, UK, July 18-19 2002. Springer-Verlag.

[112] K. Jonathan Riley, Jonathan D. Edwards, and John P. Eakins. Content-based retrieval of historical watermark images: Ii - electron radiographs. In *Proceedings of the International Conference on Image and Video Retrieval, Lecture Notes in Computer Science*, volume 2728, pages 131–140, Illinois, USA, July 24-25 2003. Springer-Verlag.

[113] John C. Russ. *The image processing handbook*. CRC Press, Boca Raton, USA, 2nd edition, 1995.

[114] Stan Salvador and Philip Chan. Determining the number of clusters/segments in hierarchical clustering/segmentation algorithms. In *Proceedings of 16th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'04)*, pages 576–584. IEEE Computer Society Press, 2004.

[115] Dierk Schnitger and Eberhard Mundry. Elektronenradiographie, ein Hilfsmittel für die Analyse von Wasserzeichen und Miniaturmalereien. *Restaurator*, 5(1/2):156–164, 1981/82.

[116] Dierk Schnitger, Eva Ziesche, and Eberhard Mundry. Elektronenradiographie als Hilfsmittel für die Identifizierung schwer oder nicht erkennbarer Wasserzeichen. *Gutenberg-Jahrbuch*, 58:49–67, 1983.

[117] David Schoonover. Techniques of reproducing watermarks: A practical introduction. In *Essays in Paper Analysis*, pages 154–167. Folger Shakespeare Library, Washington, USA, 1987.

[118] Alfred Henry Shorter. *Paper mills and paper makers in England, 1495-1800*. Paper Publications Society, Hilversum, 1957.

[119] Carol Ann Small. Phosphorescence watermark imaging. In *Puzzles in Paper: Concepts in Historical Watermarks*, pages 169–181. Oak Knoll Books and the British Library, New Castle, UK, 2000.

[120] Larry D. Smith. Watermarks. `http://www.motherbedford.com/watermarks/WatermarksMain.htm`. Last accessed: 25th July 2008.

[121] Solar Imaging Ltd. APIS (Advanced Paper Imaging System). `http://www.solar-imaging.com/digital-apis.html`. Last accessed: 25th July 2008.

[122] Milan Šonka, Vaclav Hlavač, and Roger Boyle. *Image Processing, Analysis, and Machine Vision*. Thomson, 3rd edition, 2008.

[123] Special Collections at Leeds University Library. `http://www.leeds.ac.uk/library/spcoll/index.htm`. Last accessed: 25th July 2008.

[124] Stephen Spector, editor. *Essays in Paper Analysis*. Folger Shakespeare Library, Washington, USA, 1987.

[125] Allan Stevenson. Shakespearian dated watermarks. *Studies in Bibliography*, 4:159–164, 1951-2. `http://etext.lib.virginia.edu/bsuva/sb/toc/sib04toc.htm`. Last accessed: 25th July 2008.

[126] Allan Stevenson. Watermarks are twins. *Studies in Bibliography*, 4:57–91, 1951-2. `http://etext.lib.virginia.edu/bsuva/sb/toc/sib04toc.htm`. Last accessed: 25th July 2008.

[127] Allan Stevenson. Chain-indentations in paper as evidence. *Studies in Bibliography*, 6:181–195, 1954. `http://etext.lib.virginia.edu/bsuva/sb/toc/sib06toc.htm`. Last accessed: 25th July 2008.

[128] Allan Stevenson. Paper as bibliographical evidence. *The Library, 5th series*, 17(3):197–212, September 1962.

[129] Allan Stevenson. *The problem of the Missale speciale*. The Bibliographical Society, London, UK, 1967.

[130] David Stewart, Robert A. Scharf, and Jonathan S. Arney. Techniques for digital image capture of watermarks. *Imaging Science and Technology*, 39(3):261–267, 1995.

[131] Dan Sunday. Polyline simplification. `http://softsurfer.com/Archive/algorithm_0205/algorithm_0205.htm`. Last accessed: 25th July 2008.

[132] Technische Universität Braunschweig, Institut für Nachrichtentechnik. `http://www.ifn.ing.tu-bs.de/`. Last accessed: 25th July 2008.

[133] The Bibliographical Society of the University of Virginia. Studies in Bibliography. `http://etext.virginia.edu/bsuva/sb/`. Last accessed: 25th July 2008.

[134] The MathWorks. Matlab application. `http://www.mathworks.com`. Last accessed: 25th July 2008.

[135] The World Wide Web Consortium (W3C). Scalable Vector Graphics (SVG) 1.1 specification. `http://www.w3.org/TR/SVG11/`. Last accessed: 25th July 2008.

[136] Peter Toft. *The Radon Transform - Theory and Implementation*. PhD thesis, Department of Mathematical Modelling, Technical University of Denmark, July 1996.

[137] Hiroshi Tomimasu, Daijin Kim, Minsoo Suk, and Philip Luner. Comparison of four paper imaging techniques: beta-radiography, electrography, light transmission, and soft X-radiography. *Tappi*, 74(7):165–176, July 1991.

[138] Touch and Turn Corporation. `http://www.touchandturn.com/`. Last accessed: 25th July 2008.

[139] UNESCO/ICA/IFLA, Guidelines for digitization projects for collection and holdings in the public domain, particularly those held by libraries and archives. `http://www.ifla.org/VII/s19/pubs/digit-guide.pdf`. Last accessed: 25th July 2008.

[140] Jan van Aken. An improvement in Grenz radiography of paper to record watermarks, chain and laid lines. *Studies in conservation*, 48(2):103–110, 2003.

[141] Jan van der Lubbe, Eugene P. van Someren, and Marcel J.T. Reinders. Dating and authentication of rembrandt's etchings with the help of computational intelligence. In *Proceedings of International Cultural Heritage Informatics Meeting: Cultural Heritage and Technologies in the Third Millennium (ichim01)*, volume 2, pages 485–492, Milan, Italy, September 3-7 2001.

[142] Harry van Hugten. Weichstrahlradiographie Z. B. Bei Papier. In *Proceedings of Vorträge des Symposiums: Zerstörungsfreie Prüfung von Kunstwerken*, pages 43–49, Berlin, Germany, November 19-20 1987.

[143] Mark van Staalduinen. Enhancement of paper reproductions generated by X-ray or backlight imaging for several applications. Technical report, Delft University of Technology, October 2005.

[144] Mark van Staalduinen, Jan van der Lubbe, Eric Backer, and Pavel Paclík. Paper retrieval based on specific paper features: Chain and laid lines. In *Proceedings of International Workshop on Multimedia Content Representation, Classifictaion and Security (MRCS '06)*, pages 346–353, Istanbul, Turkey, September 11-13 2006.

[145] Mark van Staalduinen, Jan van der Lubbe, Dietz Georg, and Frans and Theo Laurentius. Comparing X-ray and backlight imaging for paper structure visualization. In *Proceedings of EVA, Electronic Imaging and Visual Arts*, pages 108–113, Florence, Italy, April 3-7 2006.

[146] David L. Vander Meulen. The identification of paper without watermarks. *Studies in Bibliography*, 37:58–81, 1984. `http://etext.lib.virginia.edu/bsuva/sb/toc/sib37toc.htm`. Last accessed: 25th July 2008.

[147] Terence Walz. The paper trade of Egypt and the Sudan in the eighteenth and nineteenth centuries. In M. W. Daly, editor, *Modernization in the Sudan. Essays in Honor of Richard Hill*, pages 29–48. Lilian Barber Press, New York, USA, 1985.

[148] Peter Wayner. *Disappearing cryptography, Information hiding: Steganography and watermarking*. Morgan Kaufmann, San Francisco, USA, 2nd edition, 2003.

[149] Emanuel Wenger and Victor Karnaukhov. A distributed database and processing system for watermarks. In *Proceedings of EVA'04 Moscow*, The State Tretyakov Gallery, Moscow, 2004. CD-ROM, 5 pages.

[150] Emanuel Wenger, Victor Karnaukhov, Alois Haidinger, Nikolai Merzlyakov, Gerard van Thienen, Elena Oukhanova, and Dmitry Erastov. A distributed database and processing system for watermarks: an INTAS project. In *Proceedings of EVA'01 Moscow*, pages 200–206, Centre PIC of Ministry Culture of Russia, STG, Moscow, 2001.

[151] Emanuel Wenger, Victor Karnaukhov, Alois Haidinger, and Maria Stieglecker. A digital image processing and database system for watermarks in medieval

manuscripts. In *Proceedings of International Cultural Heritage Informatics Meeting: Cultural Heritage and Technologies in the Third Millennium (ichim01)*, volume 2, pages 259–264, Milan, Italy, September 3-7 2001.

[152] Paul F. Whelan, Pierre Soille, and Alexandru Drimbarean. Real-time registration of paper watermarks. *Real-Time Imaging*, 7(4):367–380, 2001.

[153] WIES - Watermarks in Incunabula printed in España. `http://www.ksbm.oeaw.ac.at/wies/`. Last accessed: 25th July 2008.

[154] Wikipedia. Top-down and bottom-up design. `http://en.wikipedia.org/wiki/Top-down_and_bottom-up_design`. Last accessed: 25th July 2008.

[155] John Winn and Jamie Shotton. The layout consistent random field for recognizing and segmenting partially occluded objects. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '06)*, pages 37–44, New York, NY, USA, June 17-22 2006. IEEE Computer Society Press.

[156] WZMA - Wasserzeichen des Mittelalters. `http://www.oeaw.ac.at/ksbm/wz/wzma2.htm`. Last accessed: 25th July 2008.

[157] Piero Zamperoni. Wasserzeichenextraktion aus digitalisierten Bildern mit Methoden der digitalen Bildsignalverarbeitung. *Das Papier*, 43(4):133–143, 1989.

[158] Eva Ziesche and Dierk Schnitger. Elektronenradiographische Untersuchungen von Wasserzeichen in Inkunabeln. In *Proceedings of the 19th International Congress of Paper Historians (IPH Yearbook)*, volume 7, pages 209–223, Durham and Hertford, UK, September 4-10 1988.

Appendix A

Mean and variance of a match measure on two binary vectors of known 'tally'

Suppose we have two binary vectors of dimension N:

$$\mathbf{v_1} = (v_1^1, v_1^2, \ldots, v_1^N)\ ,\ \mathbf{v_2} = (v_2^1, v_2^2, \ldots, v_2^N)\ ,\ v_i^j \in \{0,1\}$$

We are told that there are I 1's in $\mathbf{v_1}$ and J in $\mathbf{v_2}$:

$$\sum_{k=1}^{N} v_1^k = I\ ,\ \sum_{k=1}^{N} v_2^k = J$$

Count $w(\mathbf{v_1}, \mathbf{v_2})$ as the number of times corresponding vector components are both 1 or 0; then $0 \le w(\mathbf{v_1}, \mathbf{v_2}) \le N$:

$$w(\mathbf{v_1}, \mathbf{v_2}) = \sum_{k=1}^{N} (1 - XOR(v_1^k, v_2^k))$$

Given $\mathbf{v_1}$, suppose $\mathbf{v_2}$ is chosen randomly– we seek the mean and variance of w. Suppose

v_1^k	v_2^k	Occurrences
1	1	a
1	0	b
0	1	c
0	0	d

where then

$$\begin{aligned} I &= a+b \\ J &= a+c \\ N-I &= c+d \\ N-J &= b+d \\ N &= a+b+c+d \end{aligned}$$

Then we seek

$$\begin{aligned} w &= a+d \\ &= a+(N-a-b-c) \\ &= a+(N-a-(I-a)-(J-a)) \\ &= 2a+N-I-J \end{aligned}$$

Now the distribution of a is hyper-geometric (see, e.g., [94]) giving

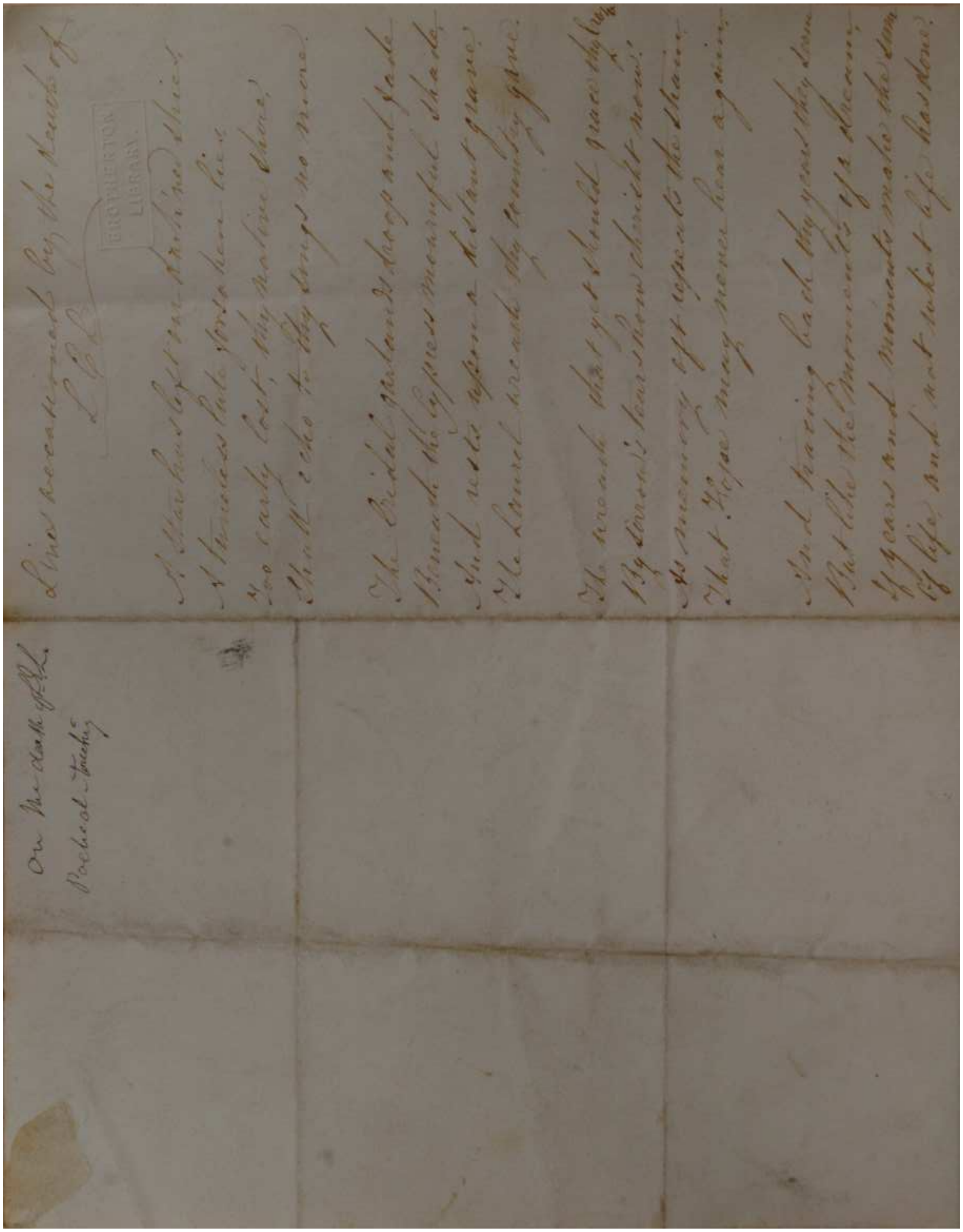
$$\begin{aligned} \mu(a) &= \frac{IJ}{N} \\ \sigma^2(a) &= \frac{IJ(N-I)(N-J)}{N^2(N-1)} \end{aligned}$$

So

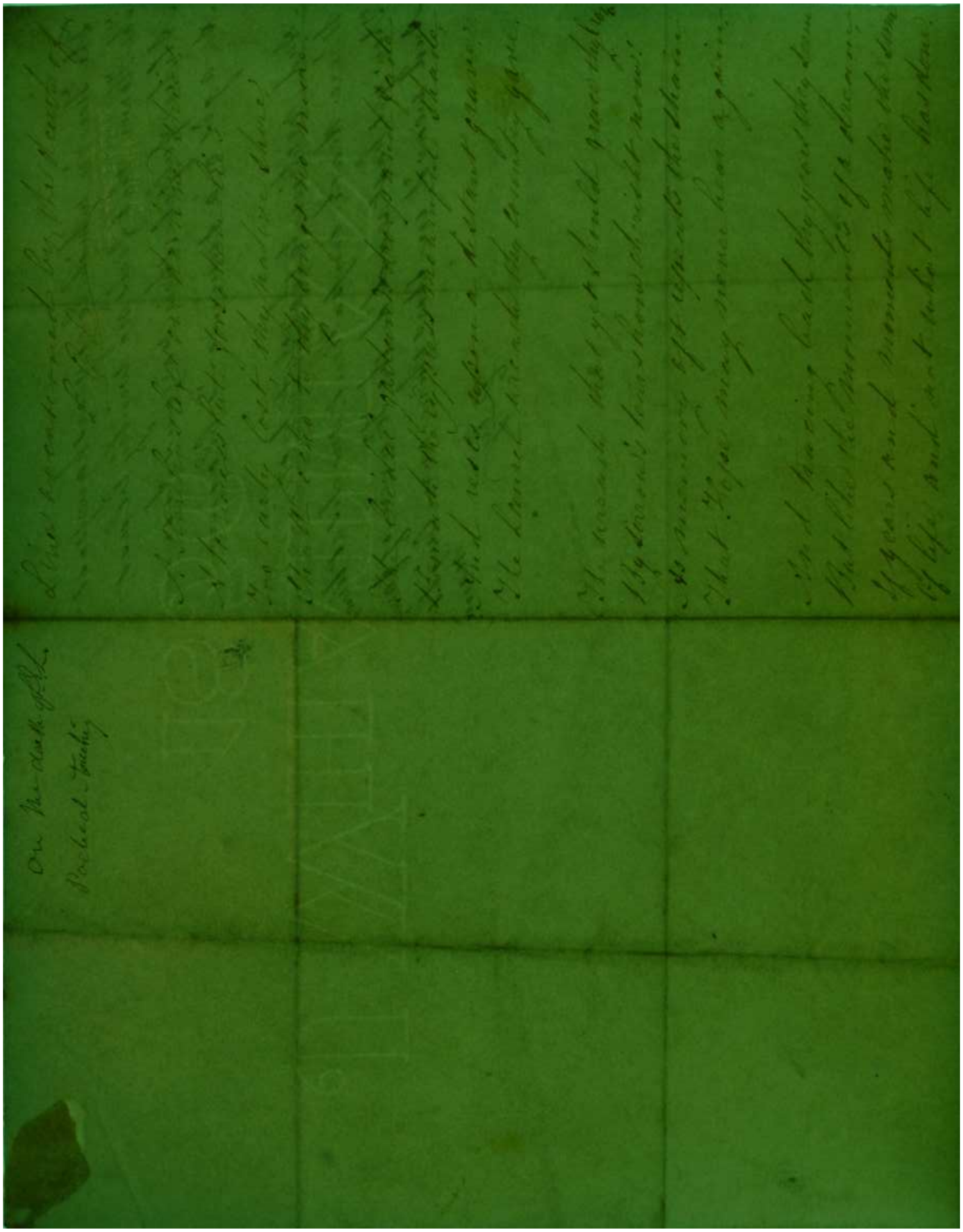
$$\begin{aligned} \mu(w) &= 2\mu(a)+N-I-J \\ &= 2\frac{IJ}{N}+N-I-J \\ \sigma^2(w) &= 4\sigma^2(a) \\ &= \frac{4IJ(N-I)(N-J)}{N^2(N-1)} \end{aligned}$$

Appendix B

Sample test data

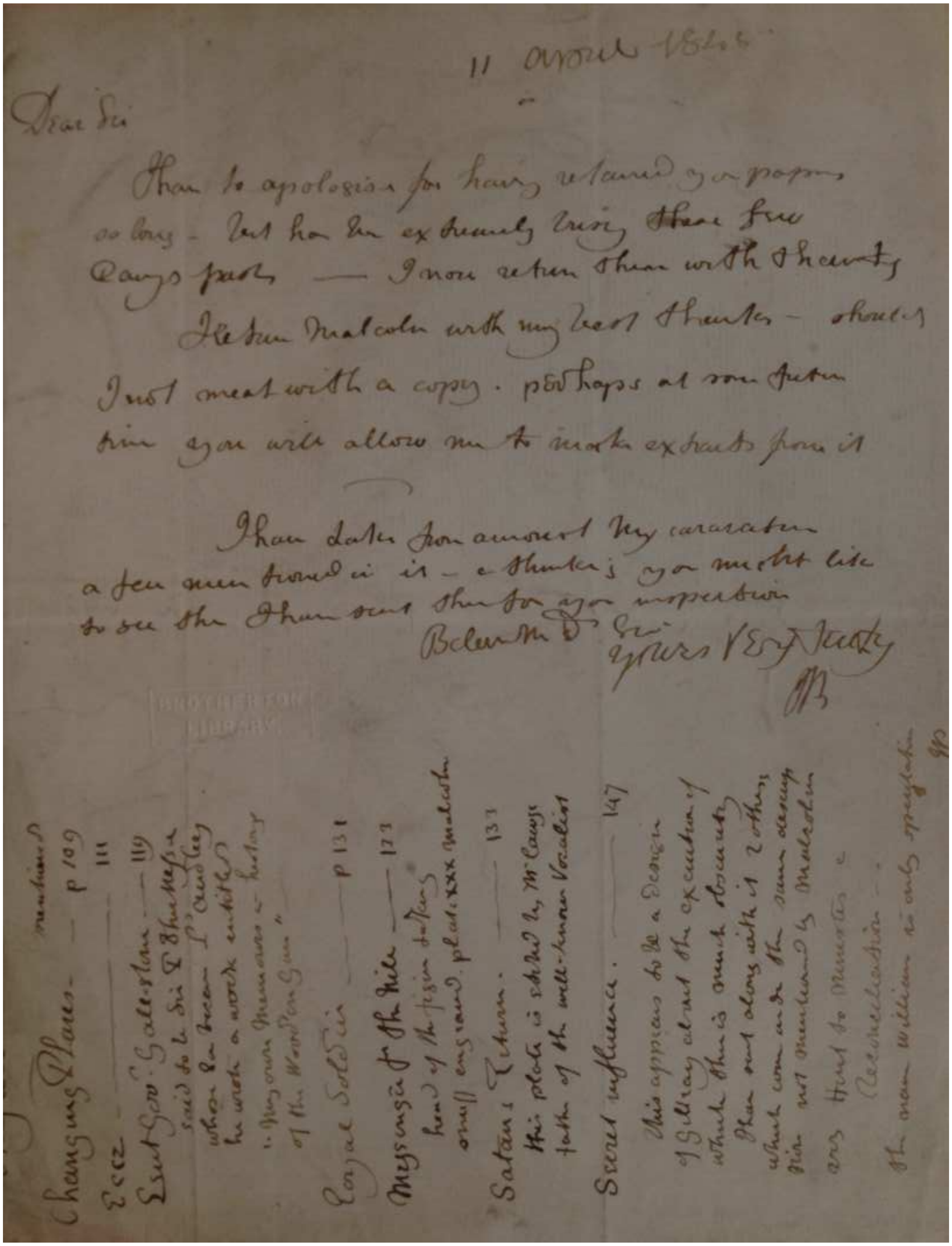


(a) Reflected



(b) Transmitted

Figure B.1: Reflected and transmitted images of a historical wove paper


11 April 18[illegible]

Dear Sir

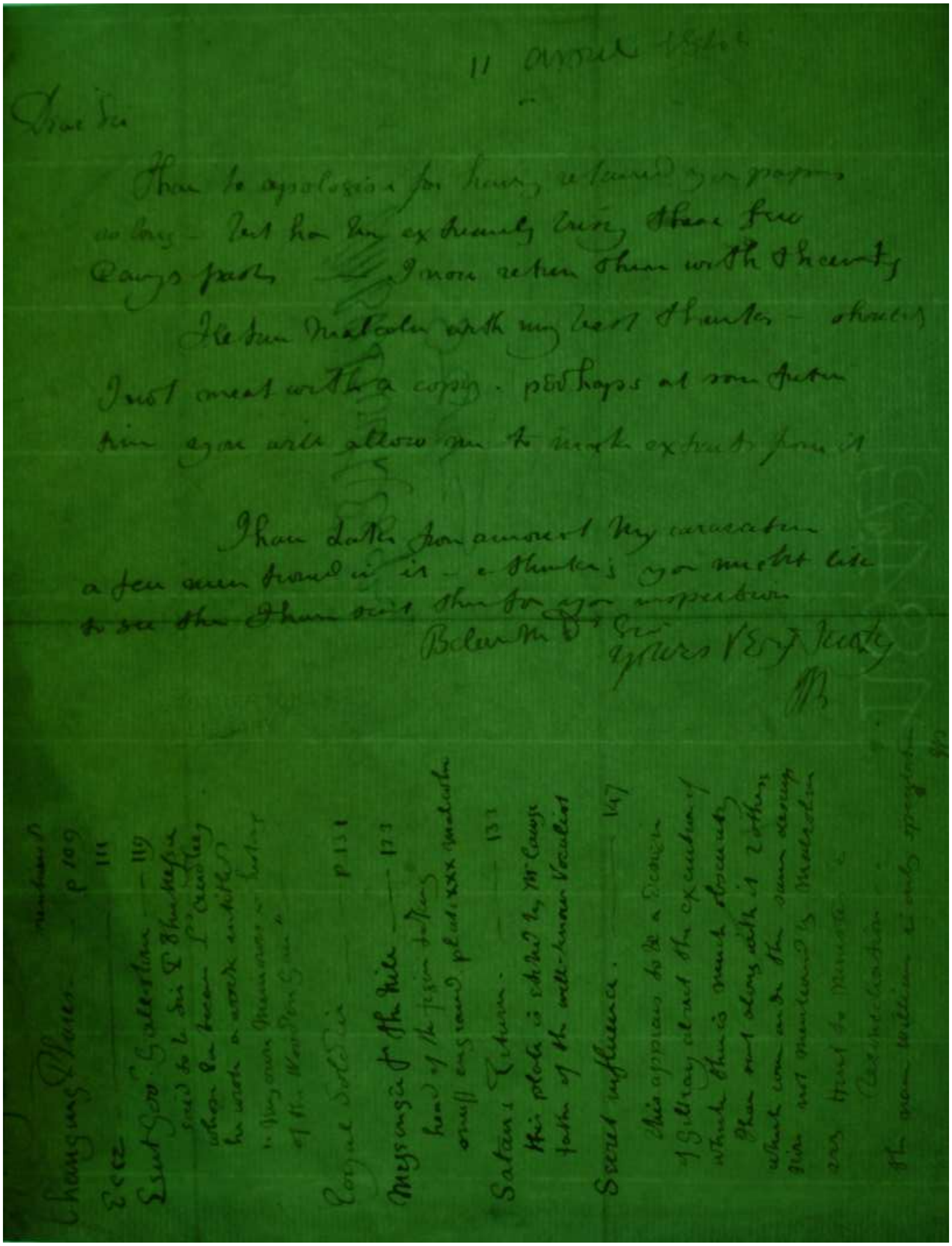
I have to apologise for having retained your papers so long - but have been extremely busy these few days past — I now return them with thanks

I return Malcolm with my best thanks - should I just meet with a copy. perhaps at some future time you will allow me to make extracts from it

I have taken from amongst my caricatures a few mentioned in it - & thinking you might like to see them I have sent them for your inspection

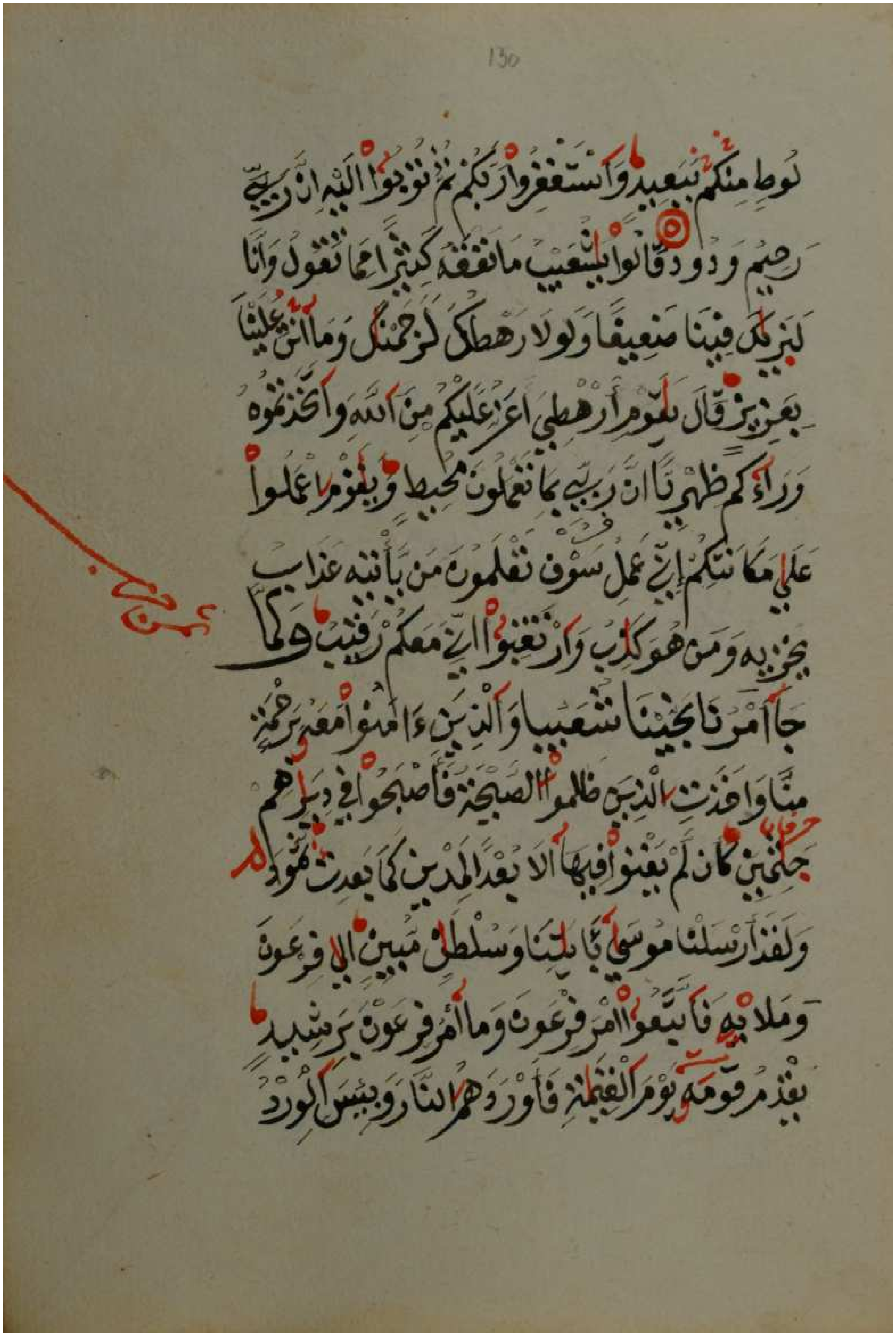
Believe me Dear Sir
Yours very truly

(a) Reflected



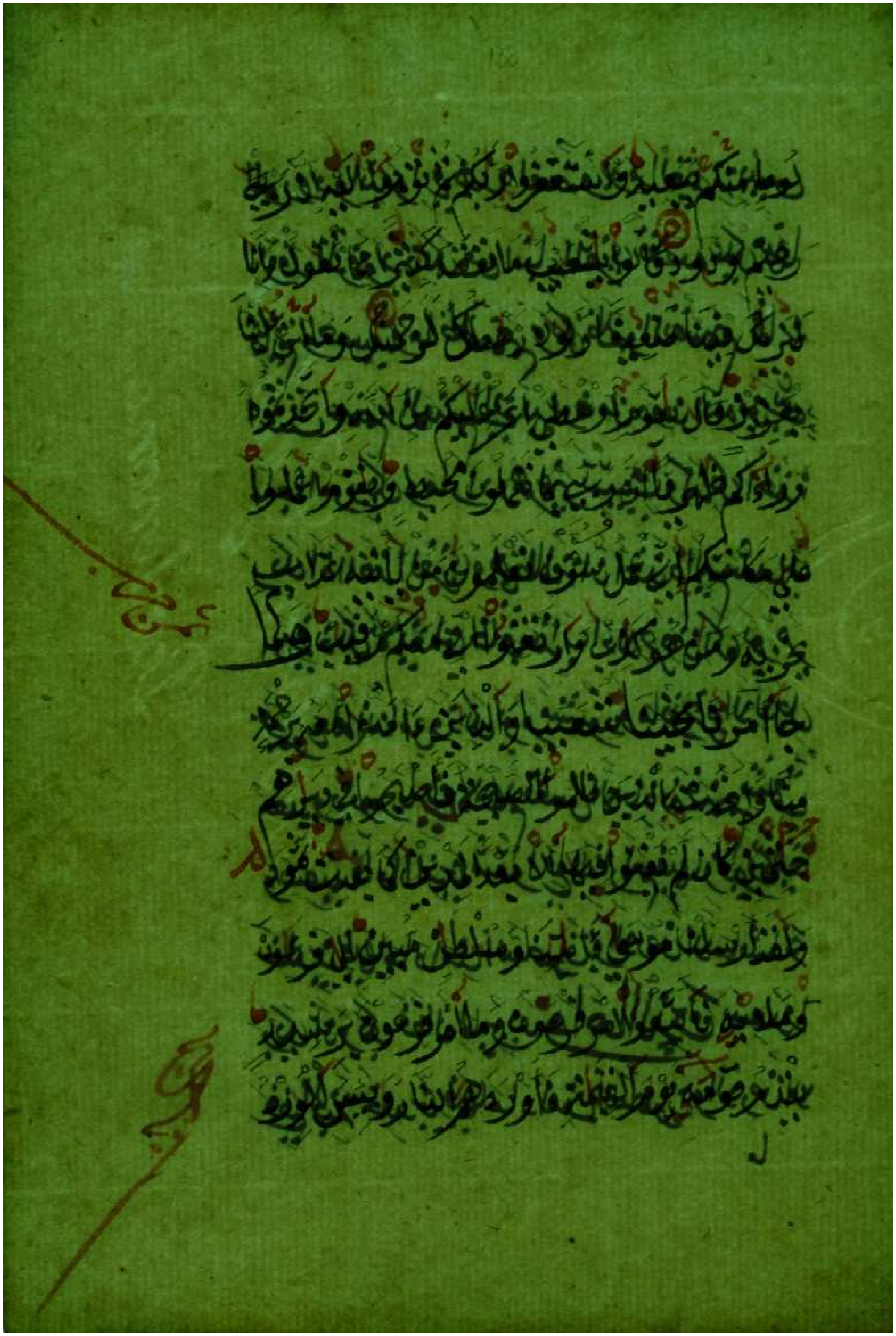
(b) Transmitted

Figure B.2: Reflected and transmitted images of a historical laid paper


130

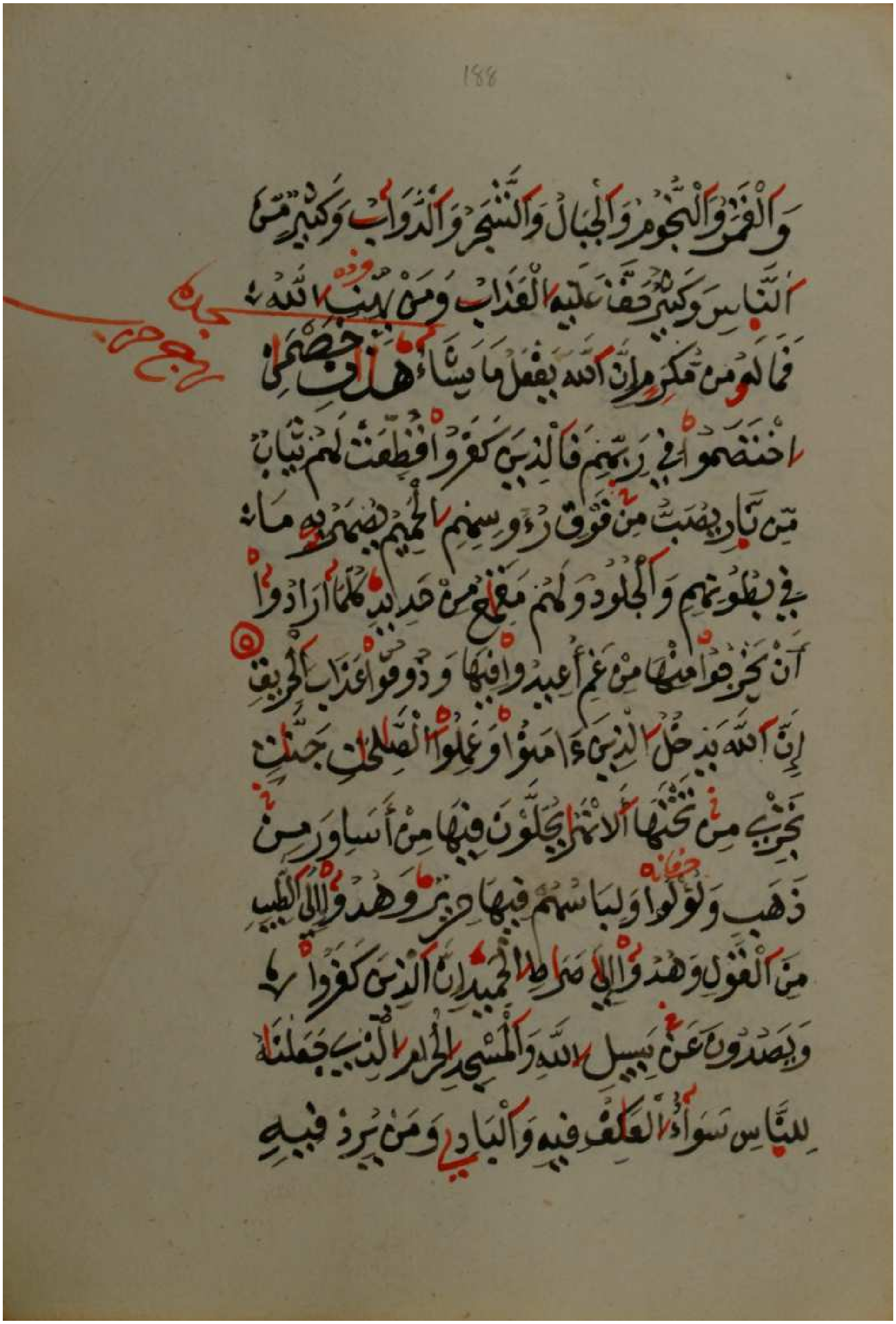
لوط منكم ببعيد واستغفروا ربكم ثم توبوا اليه ان ربي
رحيم ودود قالوا يا شعيب ما نفقه كثيرا مما تقول وانا
لنريك فينا ضعيفا ولولا رهطك لرجمناك وما انت علينا
بعزيز قال يا قوم ارهطي اعز عليكم من الله واتخذتموه
وراءكم ظهريا ان ربي بما تعملون محيط ويا قوم اعملوا
على مكانتكم اني عامل سوف تعلمون من ياتيه عذاب
يخزيه ومن هو كاذب وارتقبوا اني معكم رقيب ولما
جاء امرنا نجينا شعيبا والذين امنوا معه برحمة
منا واخذت الذين ظلموا الصيحة فاصبحوا في ديارهم
جاثمين كان لم يغنوا فيها الا بعدا لمدين كما بعدت ثمود
ولقد ارسلنا موسى باياتنا وسلطان مبين الى فرعون
وملايه فاتبعوا امر فرعون وما امر فرعون برشيد
يقدم قومه يوم القيامة فاوردهم النار وبئس الورد

(a) Reflected

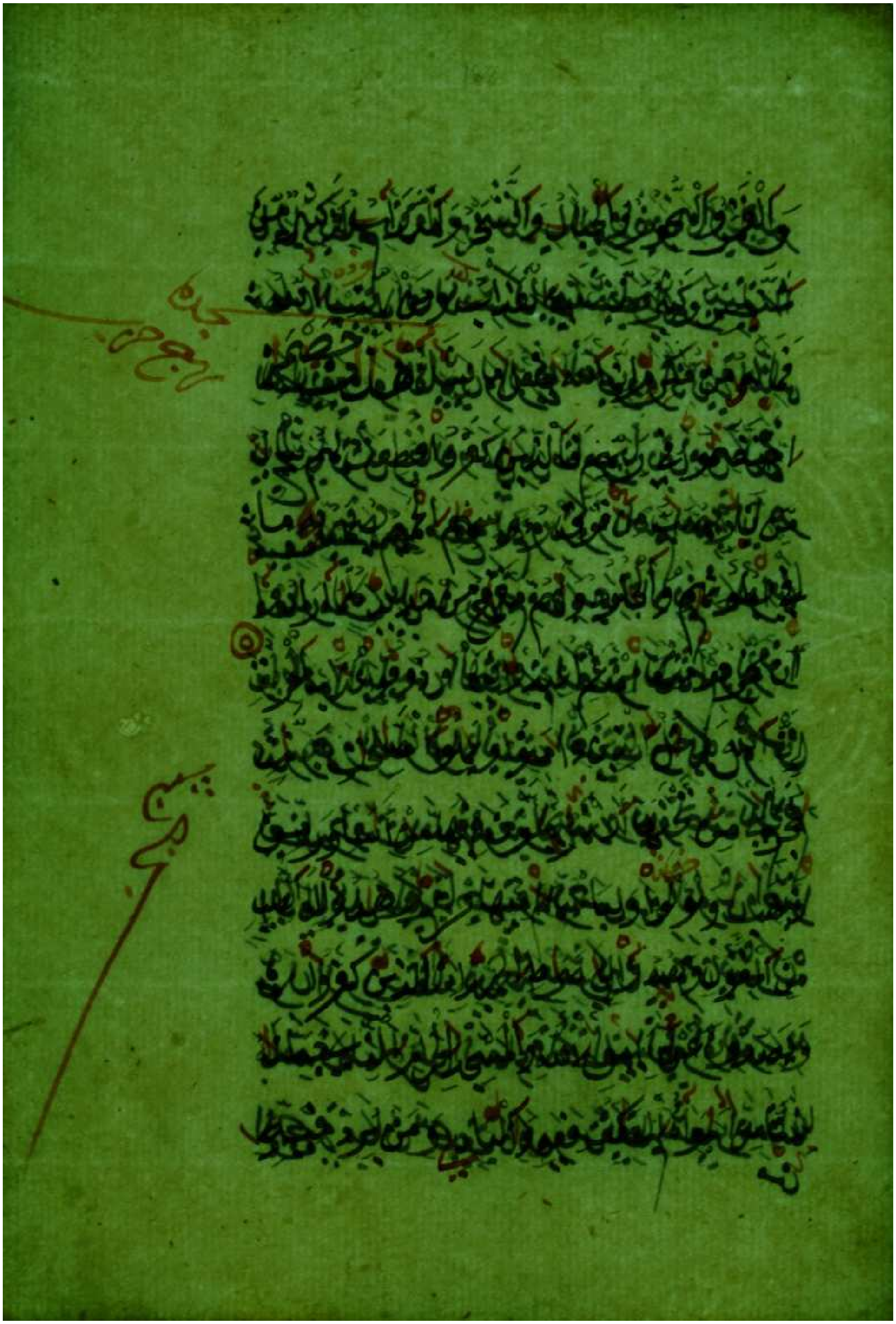


(b) Transmitted

Figure B.3: Reflected and transmitted images of a sample of the ‘Mahdiyya’ copy of the Qur’ān

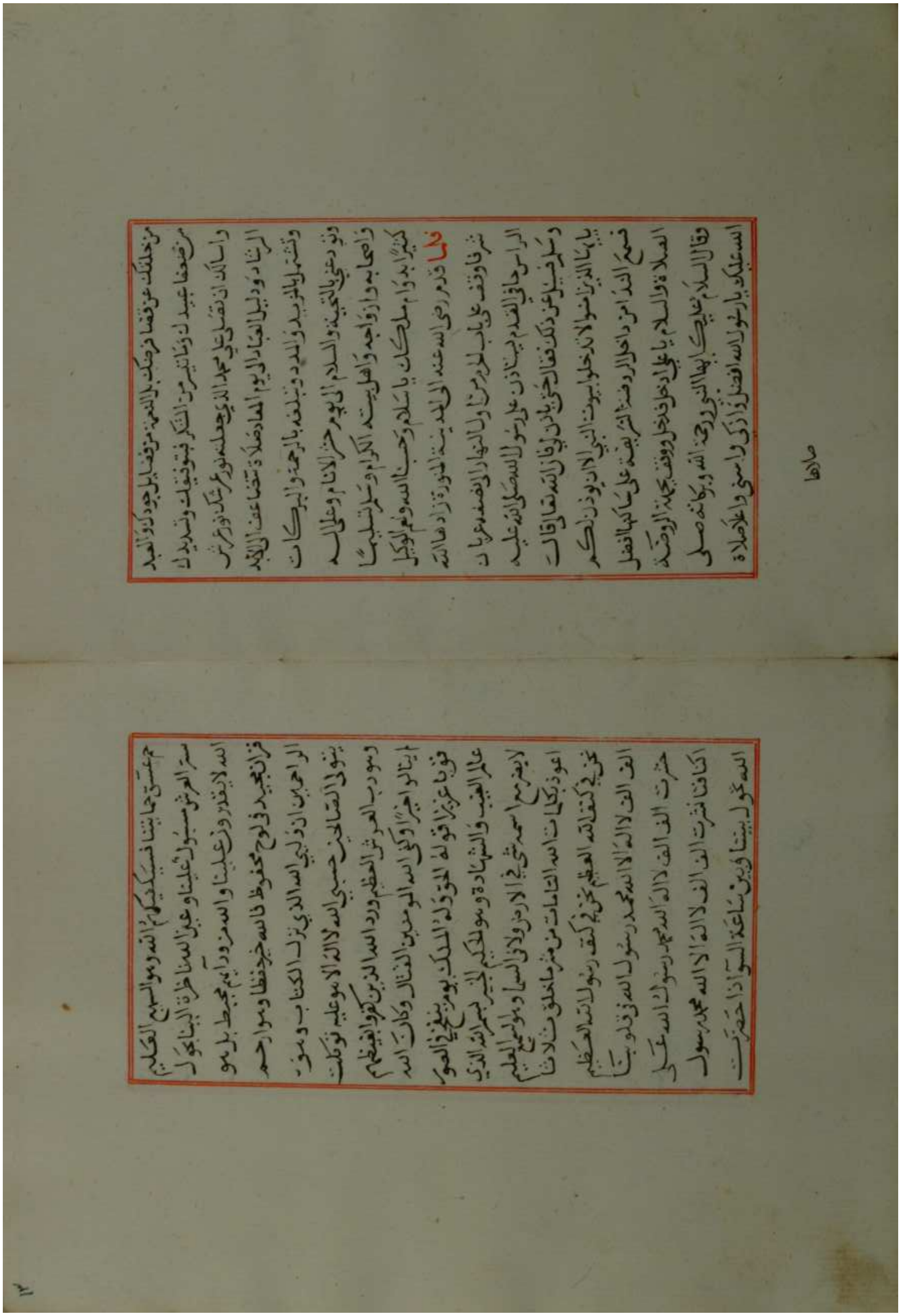

188
والقمر والنجوم والجبال والشجر والدواب وكثير من
الناس وكثير حق عليه العذاب ومن يهن الله
فما له من مكرم إن الله يفعل ما يشاء هذان خصمان
اختصموا في ربهم فالذين كفروا قطعت لهم ثياب
من نار يصب من فوق رءوسهم الحميم يصهر به ما
في بطونهم والجلود ولهم مقامع من حديد كلما أرادوا
أن يخرجوا منها من غم أعيدوا فيها وذوقوا عذاب الحريق
إن الله يدخل الذين آمنوا وعملوا الصالحات جنات
تجري من تحتها الأنهار يحلون فيها من أساور من
ذهب ولؤلؤا ولباسهم فيها حرير وهدوا إلى الطيب
من القول وهدوا إلى صراط الحميد إن الذين كفروا
ويصدون عن سبيل الله والمسجد الحرام الذي جعلناه
للناس سواء العاكف فيه والباد ومن يرد فيه
سجدة

(a) Reflected

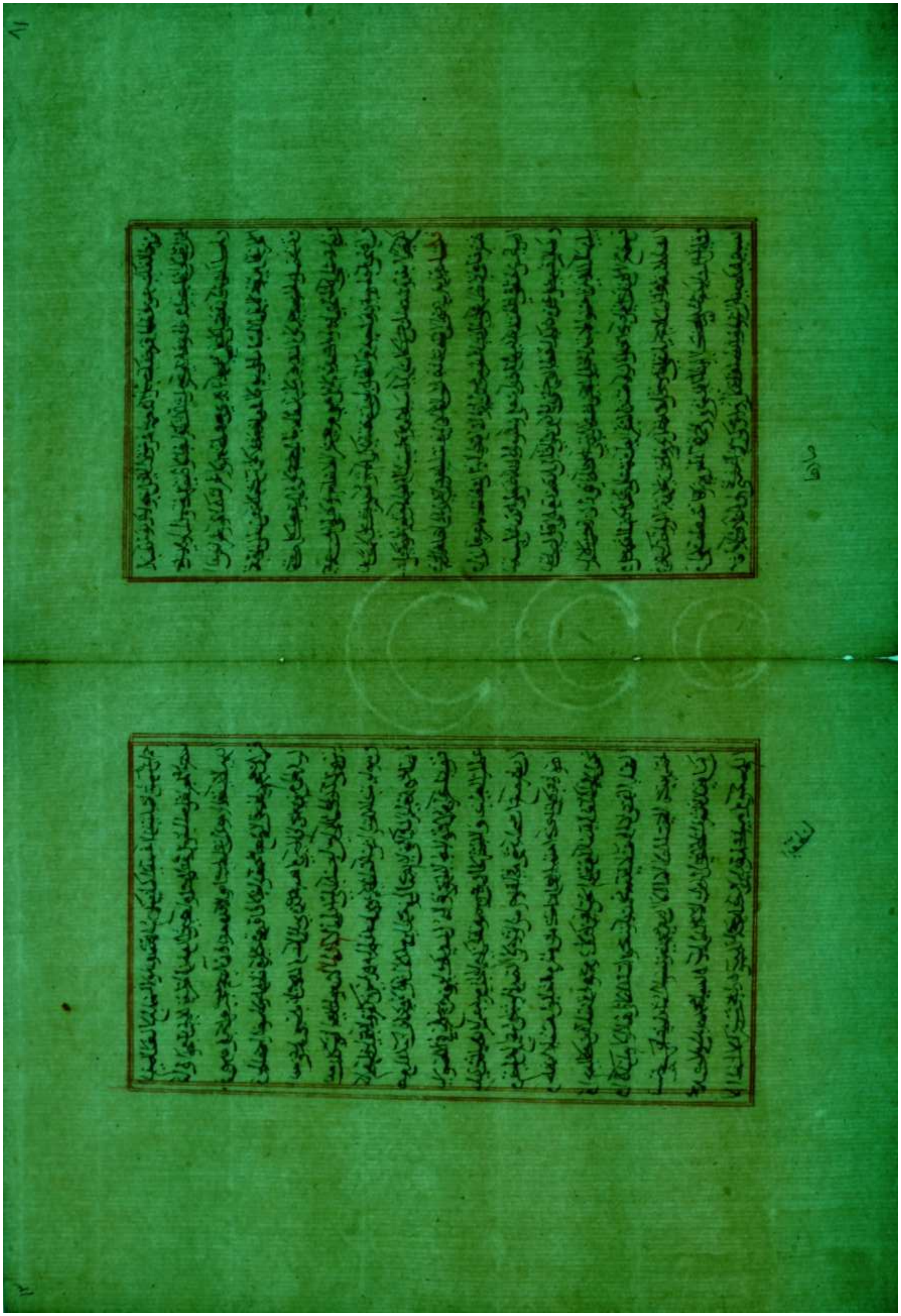


(b) Transmitted

Figure B.4: Reflected and transmitted images of a sample of the ‘Mahdiyya’ copy of the Qur’$\bar{a}$n

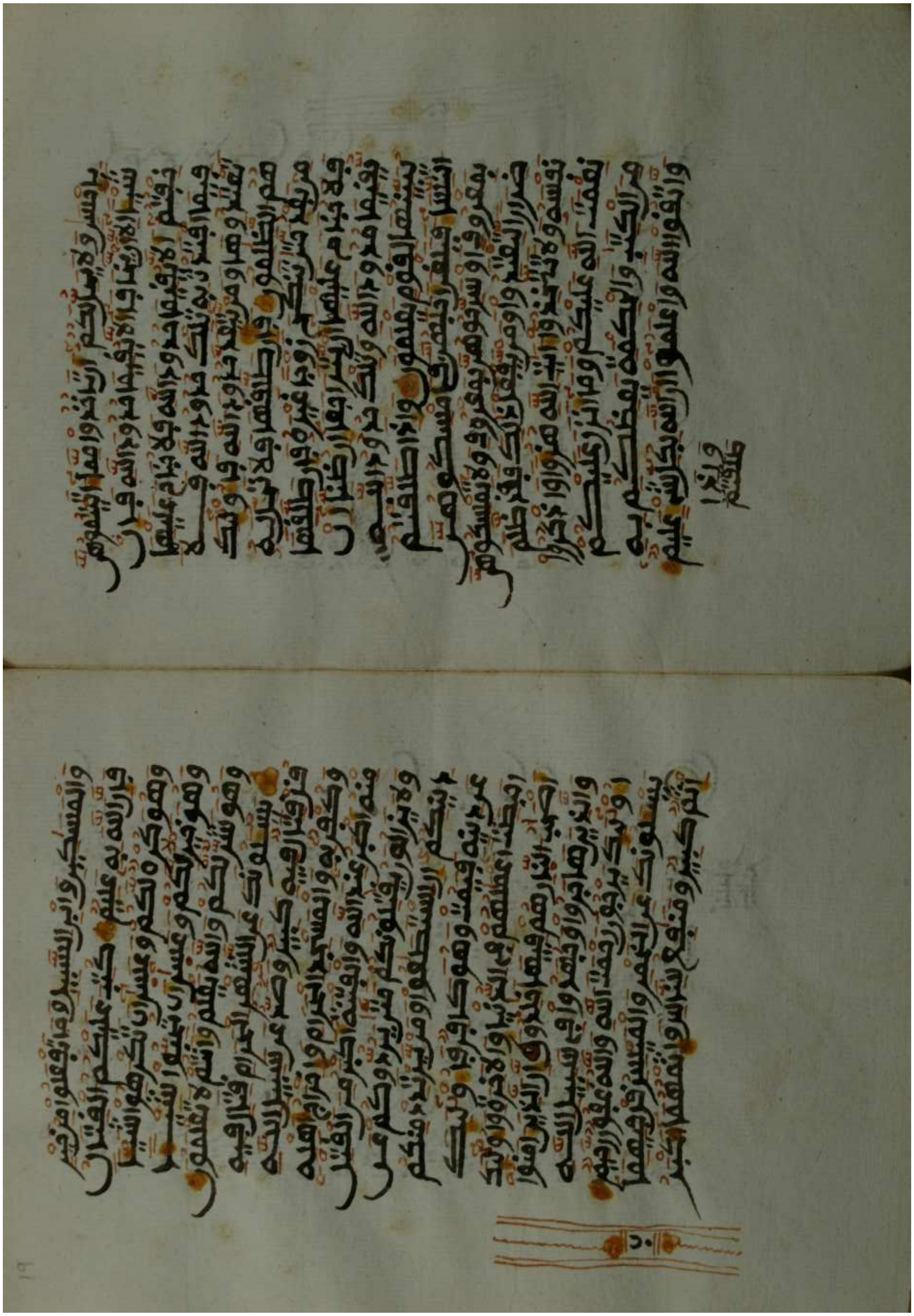


(a) Reflected

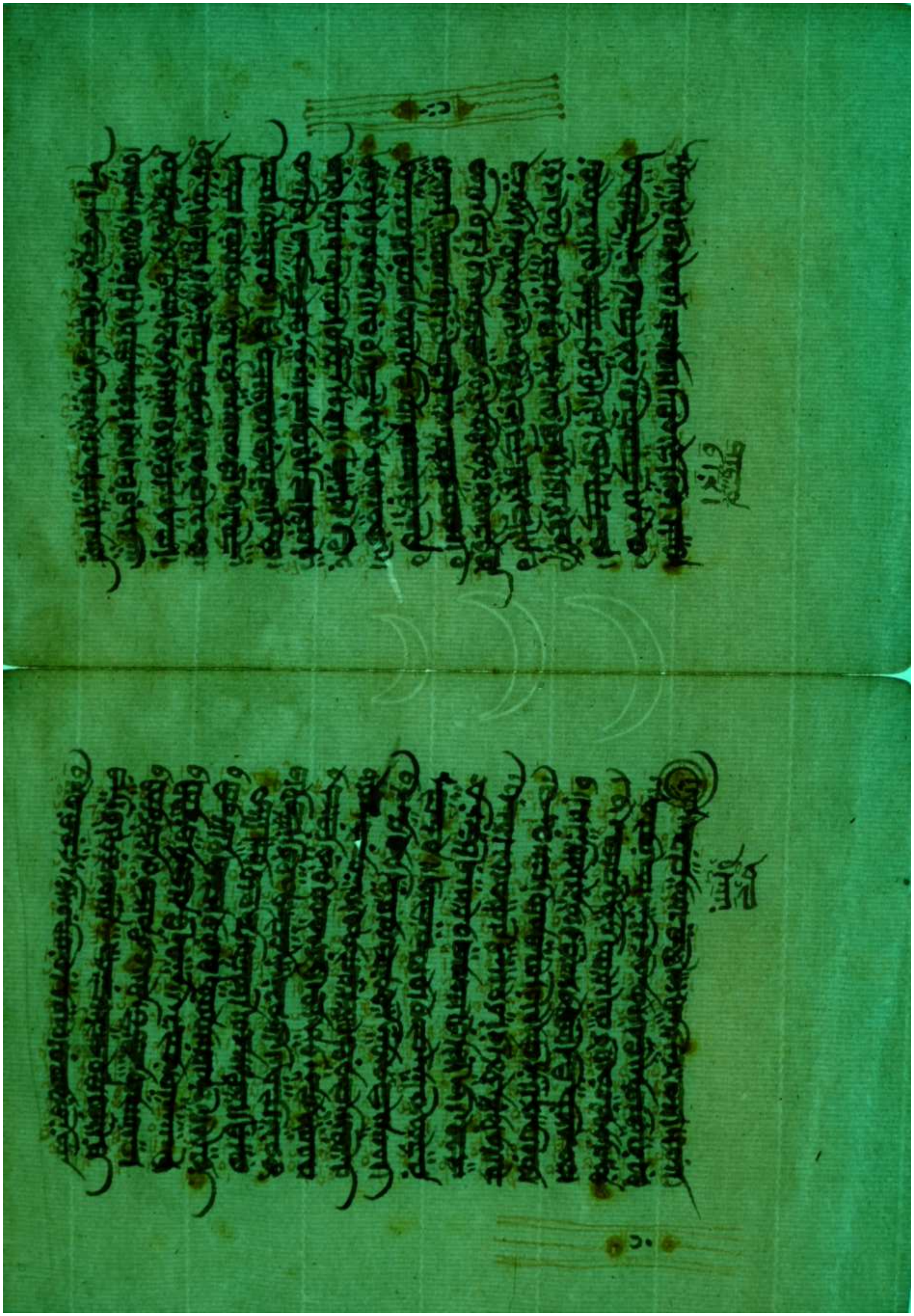


(b) Transmitted

Figure B.5: Reflected and transmitted images of a sample of the Prayer manuscript

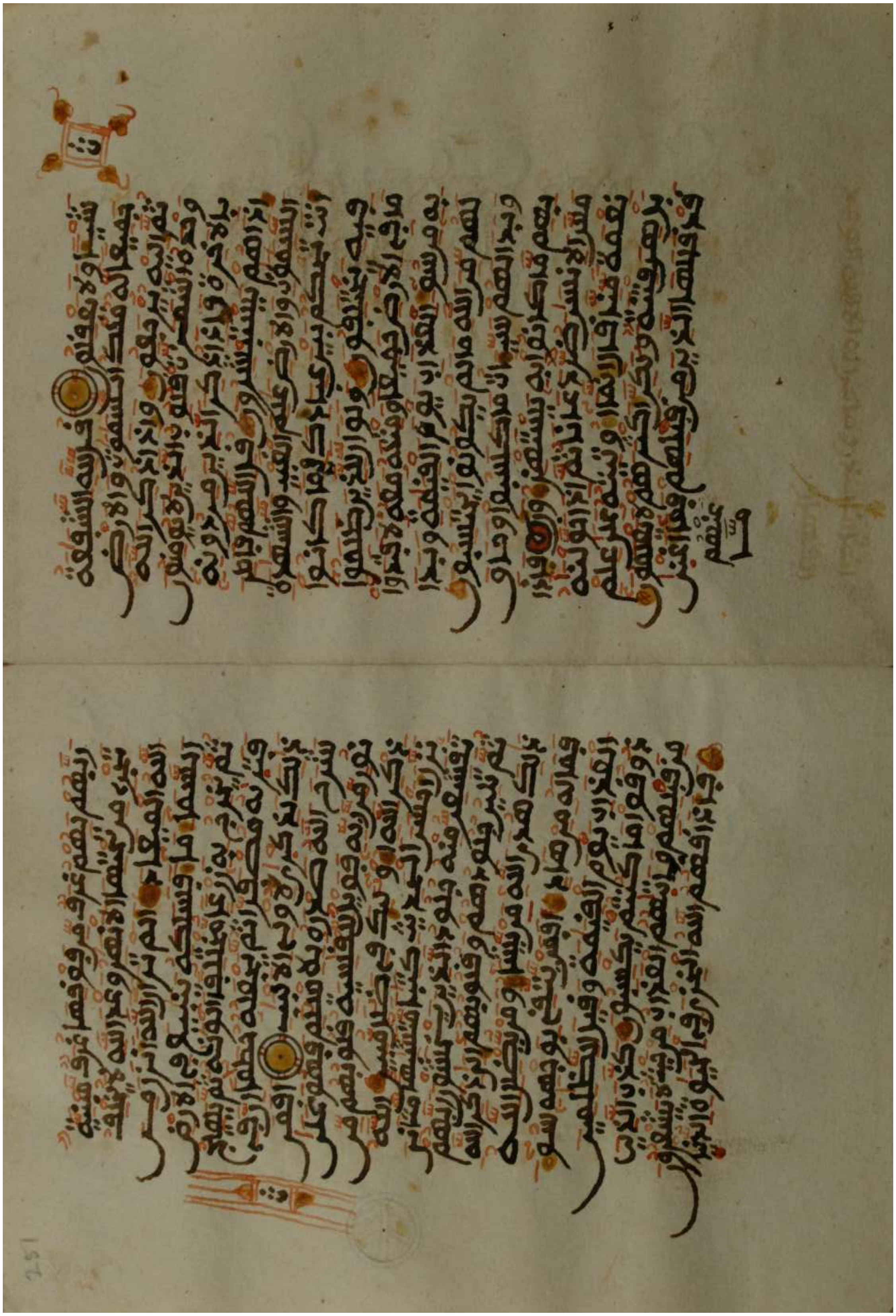


(a) Reflected

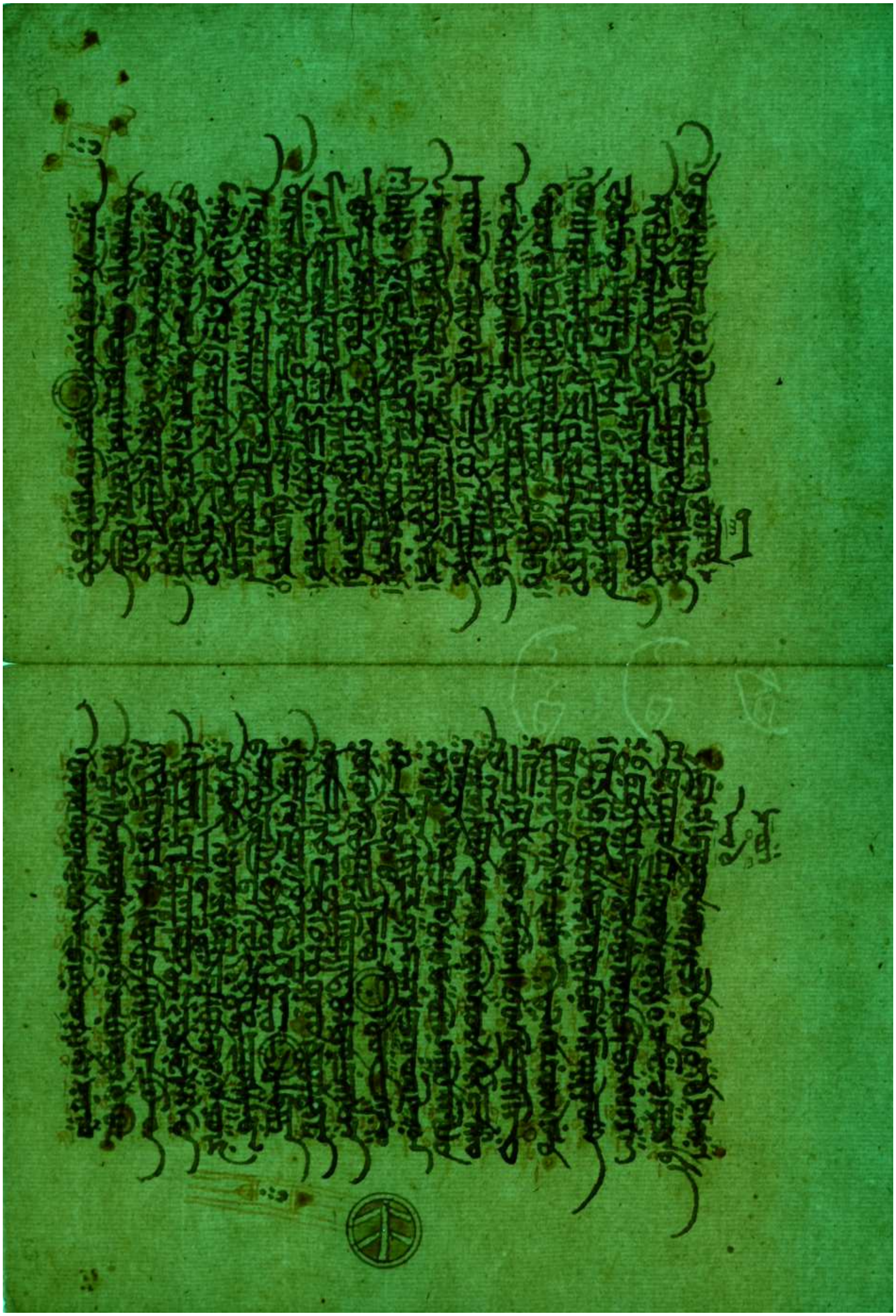


(b) Transmitted

Figure B.6: Reflected and transmitted images of a sample of the ‘West African’ copy of the Qur’ān



(a) Reflected



(b) Transmitted

Figure B.7: Reflected and transmitted images of a sample of the ‘West African’ copy of the Qur’ān

Appendix C

Sample output

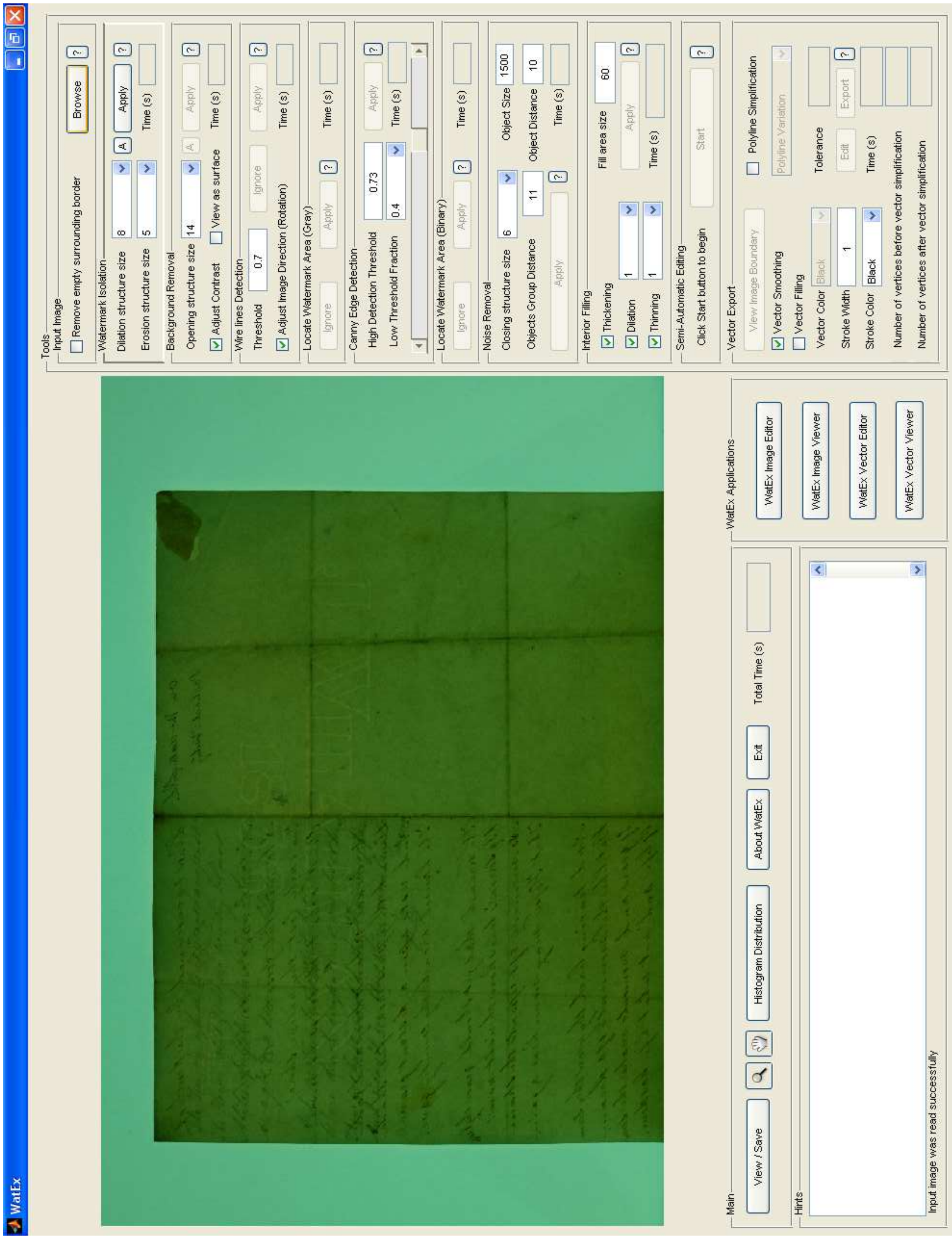


Figure C.1: Main system graphical interface of bottom-up approach

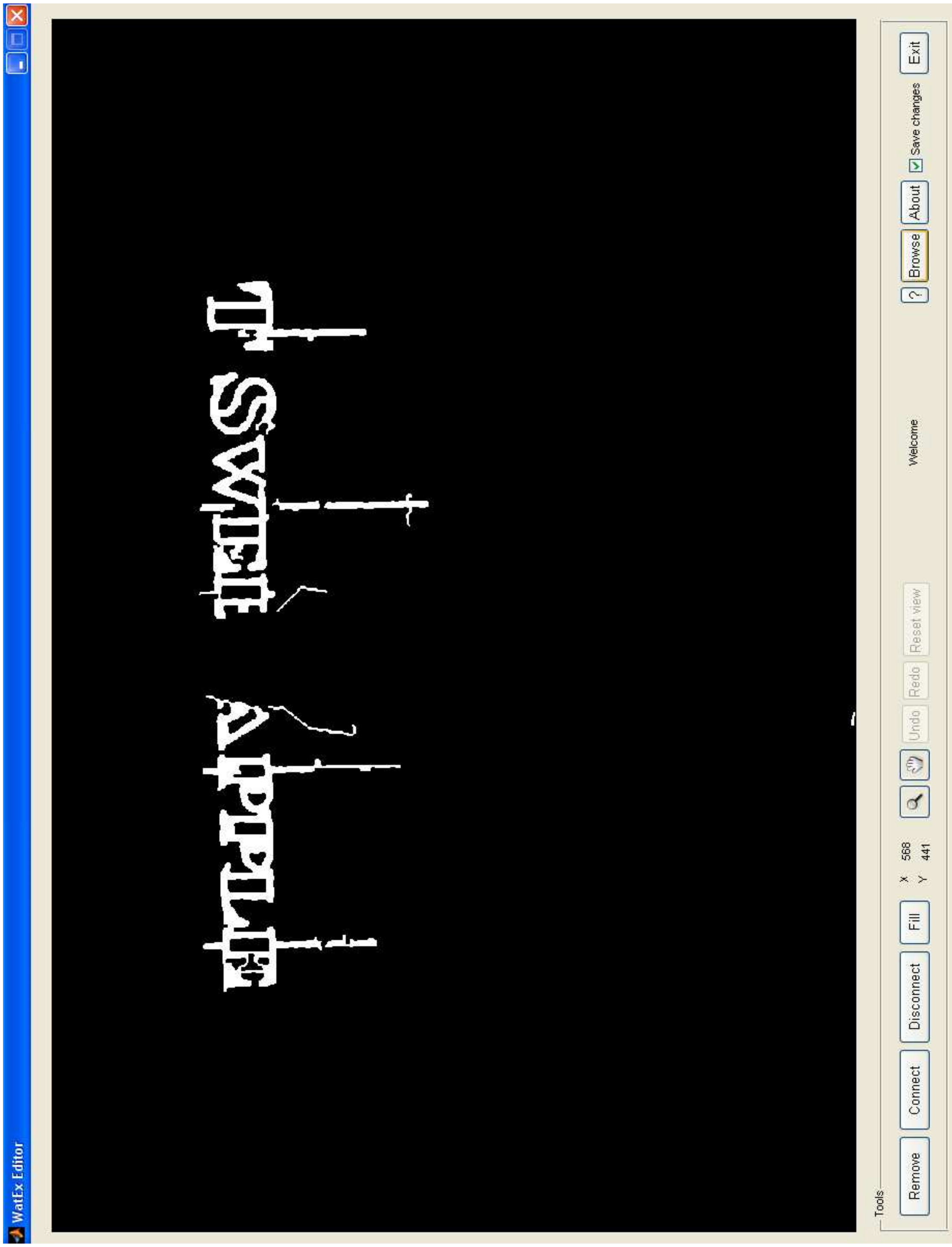


Figure C.2: Image editor graphical interface

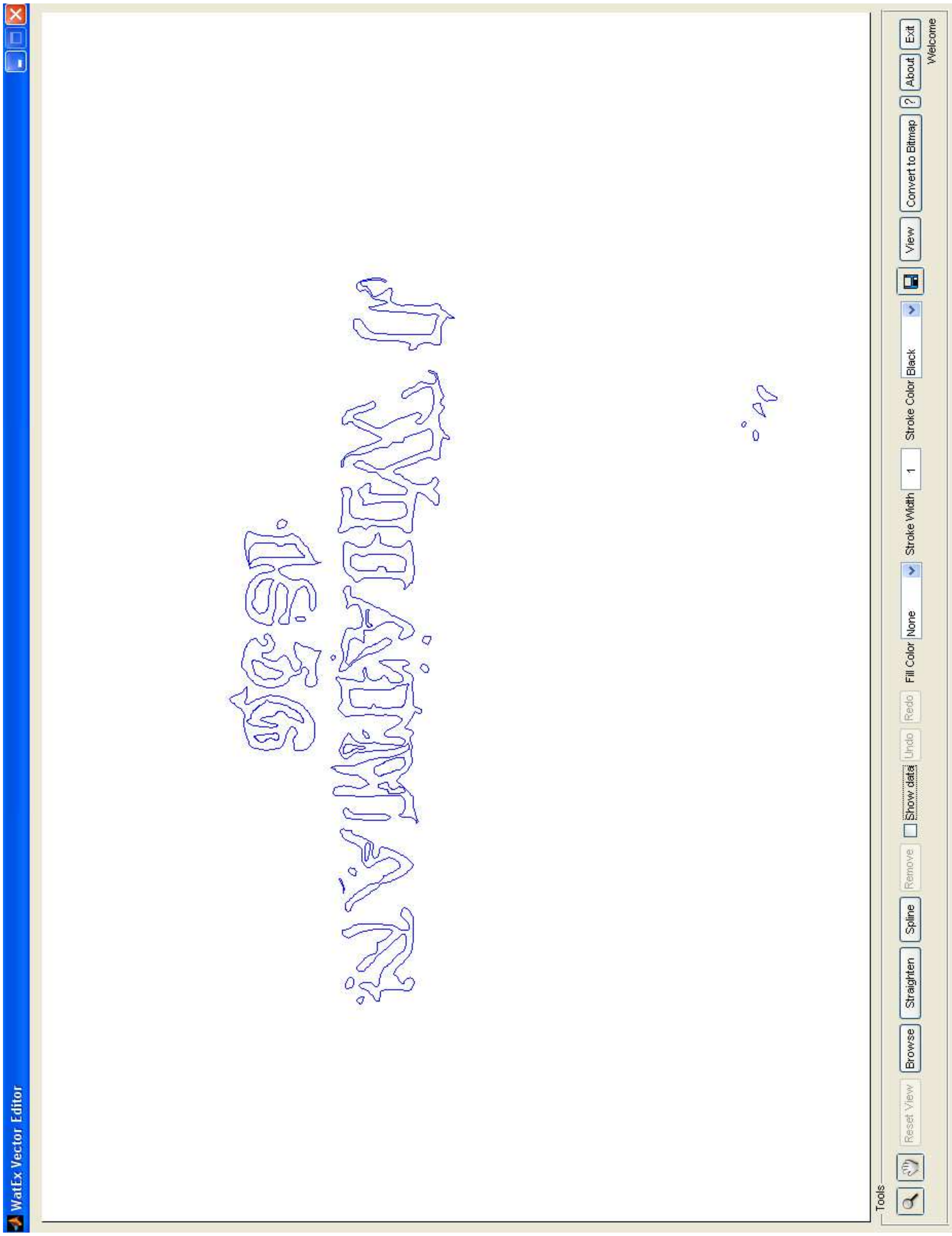


Figure C.3: Vector editor graphical interface

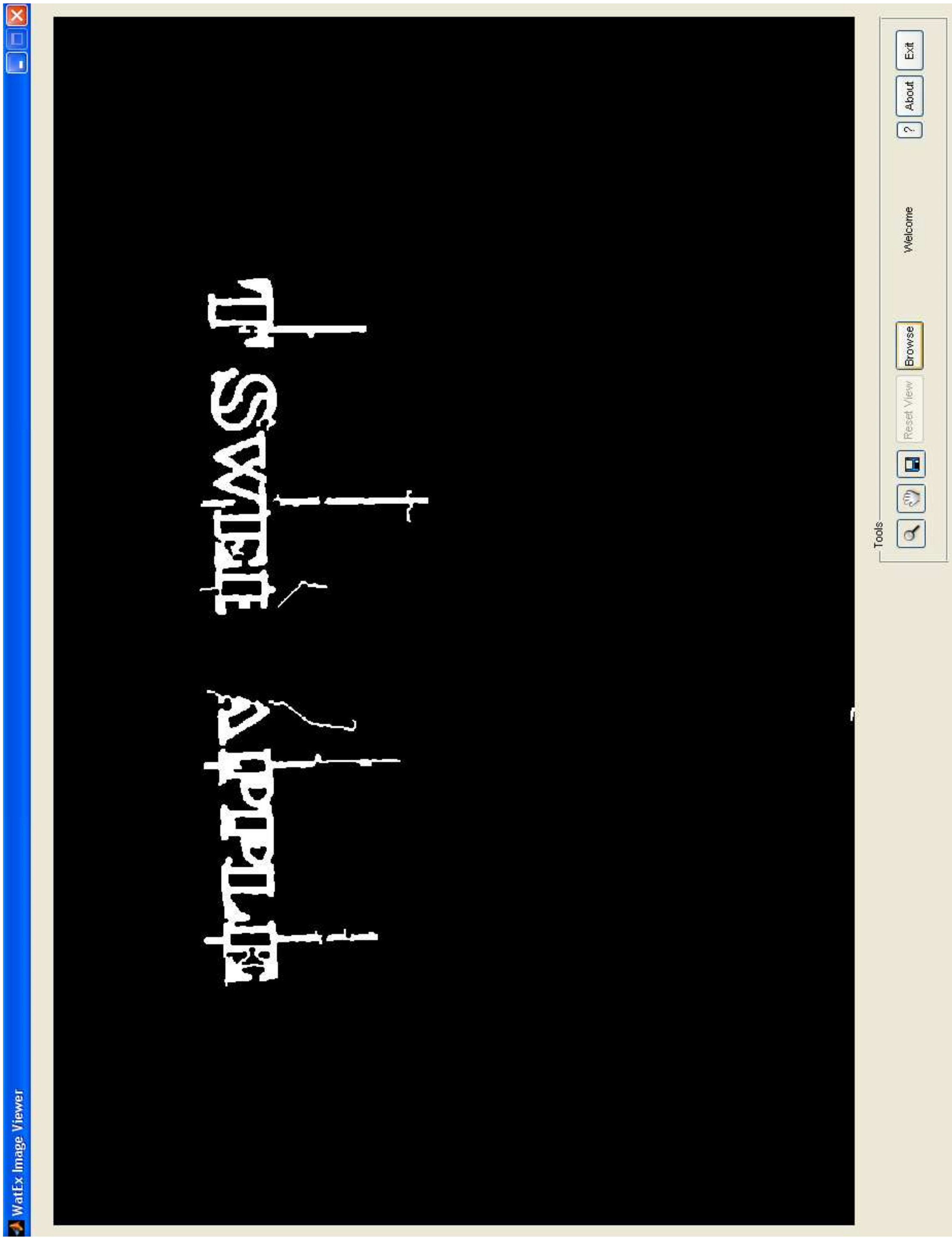


Figure C.4: Image viewer graphical interface

Figure C.5: Vector viewer graphical interface

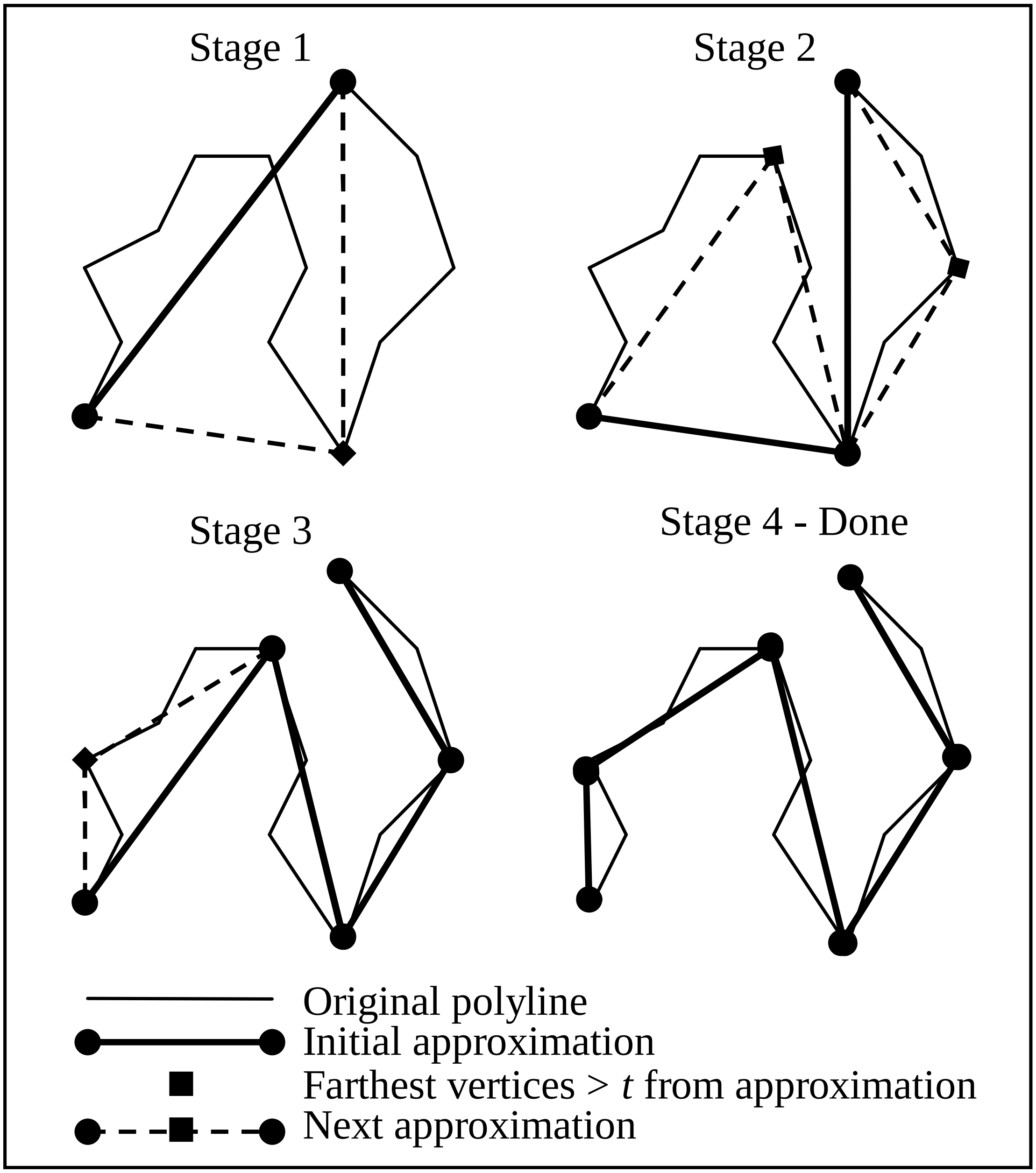


Figure C.6: An example shows Douglas-Peucker algorithm stages in details [131]

Figure C.7: Complete design of moonface-within-shield countermark

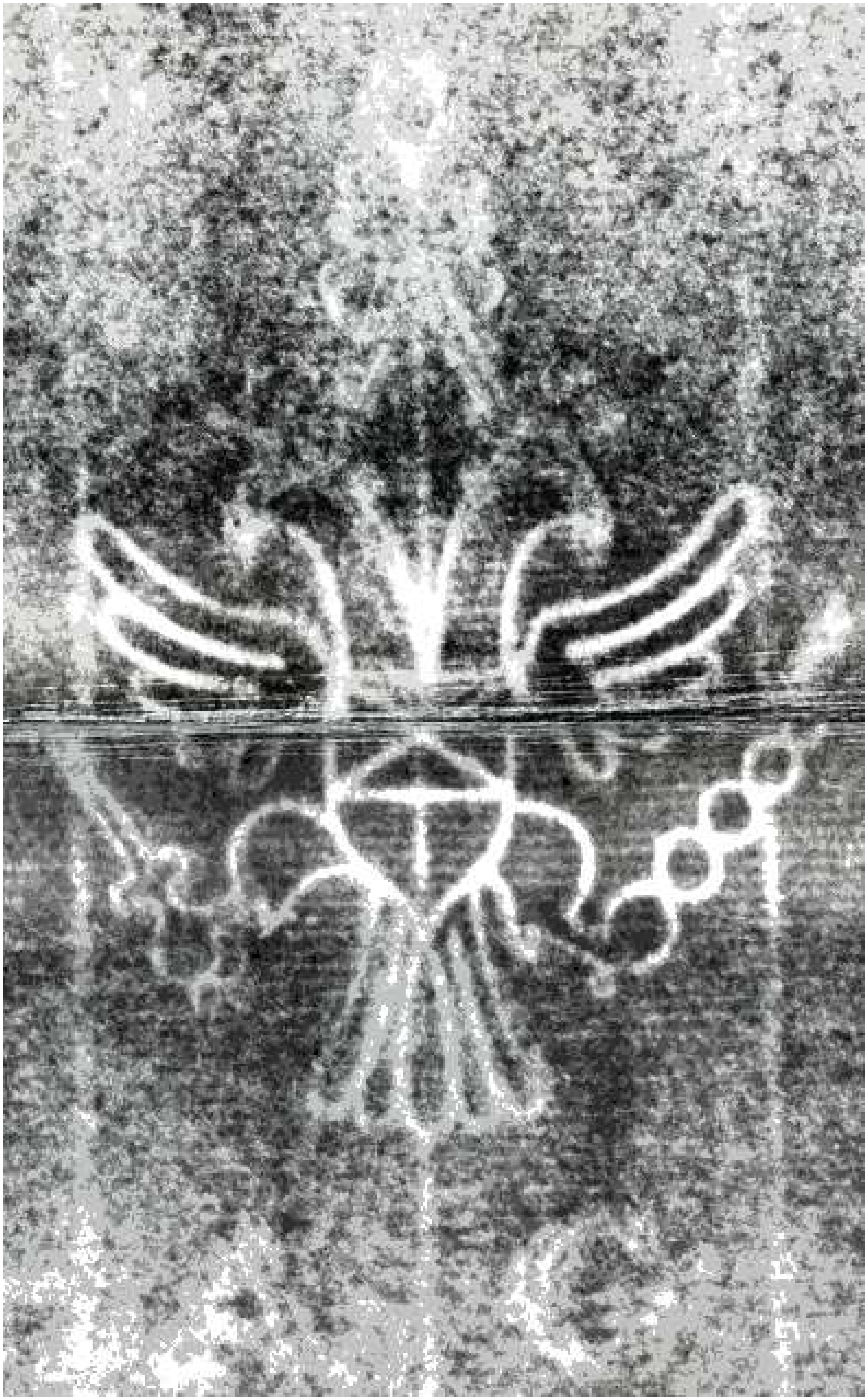


Figure C.8: Complete design of double-headed eagle watermark

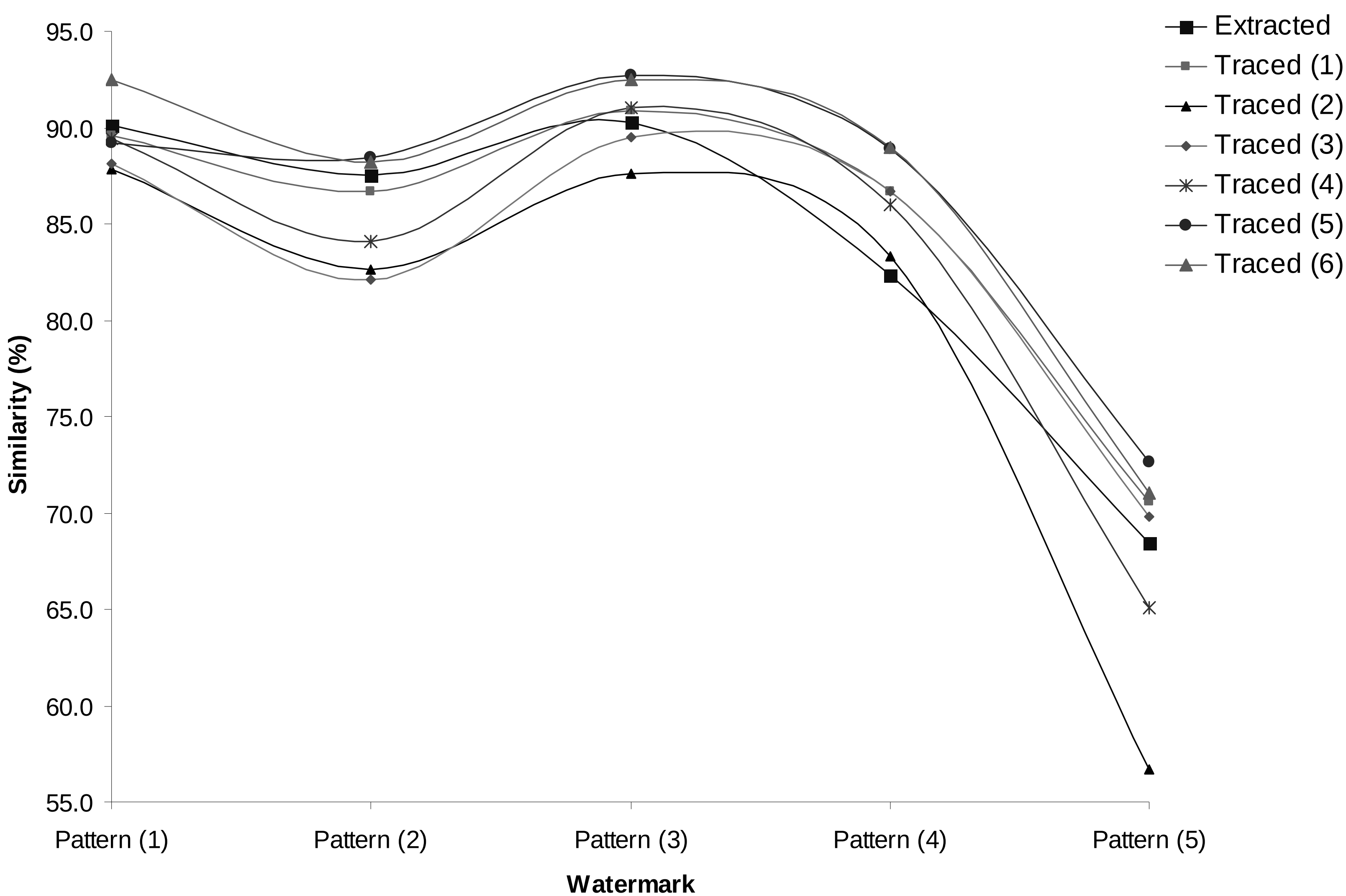


Figure C.9: Plot of similarity comparisons of extracted and traced watermarks