Algorithms for extracting text from degraded document images

Chen, Yan

2007

Chen, Y. (2007). Algorithms for extracting text from degraded document images. Doctoral thesis, Nanyang Technological University, Singapore.

https://hdl.handle.net/10356/2545

https://doi.org/10.32657/10356/2545

Nanyang Technological University

Algorithms for Separating Text from the Background in Scanned Document Images

Chen Yan

School of Computer Engineering

A thesis submitted to the Nanyang Technological University in fulfillment of the requirement for the degree of Doctor of Philosophy

2007

Table Of Contents

Acknowledgement

I would like to acknowledge and express my deepest gratitude to my supervisor, Associate Professor Graham Leedham, for his intellectual guidance, advice and support rendered during the research and development of this report. Without his patience, understanding and encouragement, I cannot stick to and dip into my project. He has also constantly helped me to proof read all my papers and reports as well as updating me on the latest technological advances relating to the areas of my work.

Many thanks also Professor T. Srikanthan for providing me a very good and harmonic research environment and advanced facilities at the Centre for High Performance Embedded Systems (CHiPES).

I would also like to extend my thankful to the staff of CHiPES, especially Ms Boh-Nah Kiat Joo and Mr Chua Chiew Song, they kindly assisted me in many aspects, and provided me with all the convenience I need.

Lastly but not the least to my husband, Zheng Xuebin for his understanding and encouragement during my research studies here in NTU.

Abstract

Documents usually contain a large amount of information and have been the primary information medium in our society. From handwritten or printed letters and words, signed cheques and contracts in our normal life to the covenants between countries, documents are an important medium of record and their importance in law still cannot be replaced by any other medium. For newly created electronic documents, searching based on keywords or phrases is relatively straightforward as the documents are created using appropriate software which makes them easily compatible with other software enabling keyword searching to be readily performed. However, many older documents only exist in paper form and are usually converted to computer form by scanning the documents and storing them as images in appropriate formats. Popular image formats are Adobe pdf, Postscript and TIFF. Images scanned in these formats can only be displayed or printed using computer tools. It is not possible to search them by keywords unless sophisticated image-processing (such as image segmentation, image layout analysis, image understanding and image classification) tools are applied. This makes document image analysis an important research area.

The main objective of this research is to automatically separate text from the background in degraded scanned document images and locate individual words in the text. In this thesis, techniques are presented to extract the handwritten text from the noisy or degraded background. This is accomplished through a multi-stage technique, which analyses the feature vectors in a local block and then chooses the most appropriate threshold method in a database for each block. The multi-stage algorithm is suitable for any document and is demonstrated on four different types: historical documents, form documents, newspaper images and cheque images.

Qualitative and quantitative comparison of several thresholding algorithms is reported. Quantitative comparison of thresholding and separation techniques is achieved through the calculation of ‘recall’, the proportion of complete correct words retained in the document after thresholding and ‘precision’, the proportion of correctly detected words to apparently detected words retained in the document. Independent Component Analysis is investigated as a means of separating touching and overlapping of descenders and ascenders on adjacent lines of the extracted text.

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

Documents usually contain a large amount of information and have been the primary information medium in our society for many years. From handwritten or printed letters, signed cheques, legal contracts, application forms to the treaties between countries, documents are an important recording media. And their acceptance in law still cannot be replaced by any other recording medium.

Whilst paper documents continue to perform an important role in the world, the storage space and safety equipment required to keep important papers has become a major problem. Many companies and organizations systematically convert their records from hardcopy paper versions to electronic versions. This reduces the large volume of paper storage to a few Giga-bytes of data on magnetic or optical disc and occupies a tiny fraction of the original storage space. These digital data documents can then be made available on-line and accessible to many people over a wide geographical area for electronic searching and reviewing.

One of the major difficulties with these electronic document archives is the difficulty of searching for keywords or phrases in the documents.

For newly created electronic documents, searching is relatively straightforward as the documents are usually created using appropriate software which makes them easily compatible with other software (e.g. Adobe Acrobat), thus enabling keyword searching to be readily carried out.

However, many older documents were created using handwriting or typewriter/printer and only exist in paper form. These are usually converted to computer compatible form by scanning the documents and storing them in the appropriate image format. The more popular image formats include Adobe pdf, Postscript and TIFF. Images scanned and stored in these formats can only be displayed or printed using computer tools. It is not possible to search them for keywords unless sophisticated image-processing (such as image segmentation, image layout analysis, image understanding and image classification) tools are applied. This makes document image processing an important research area. Esposito et al. (1993) [27] and Suen et al. (1993) [84] described the main components of a document management system, which included the main document image processing tools. A block diagram of a typical document management system is shown in Figure 1.1.

In the ‘Scanned Document’ stage, the scanner is the main component. Scanned documents are usually grey-scale or colour images. Normally, to minimisc digital storage space, grey-scale representations are chosen. As such, this research is constrained to grey-scale scanned images. Scanners usually come with software, such as Adobe's Photoshop product, that allows a captured image to be resized and modified. The software allows the images to be saved in different image formats such as .bmp, .tiff, .gif or .JPEG.

The recognition system and analysis requires access to all information of the original image, and requires uncompressed images that do not take up large storage space. Bitmap (BMP) format requires large storage and thus is not usually used. GIF and JPEG formats are mainly used for Internet graphics. The TIFF format describes image data that comes from scanners and allows images to be saved as compact digital files without any loss of data. As a result, the image scanned in the 'Scanned Document' of Figure 1.1 is frequently saved in TIFF format.

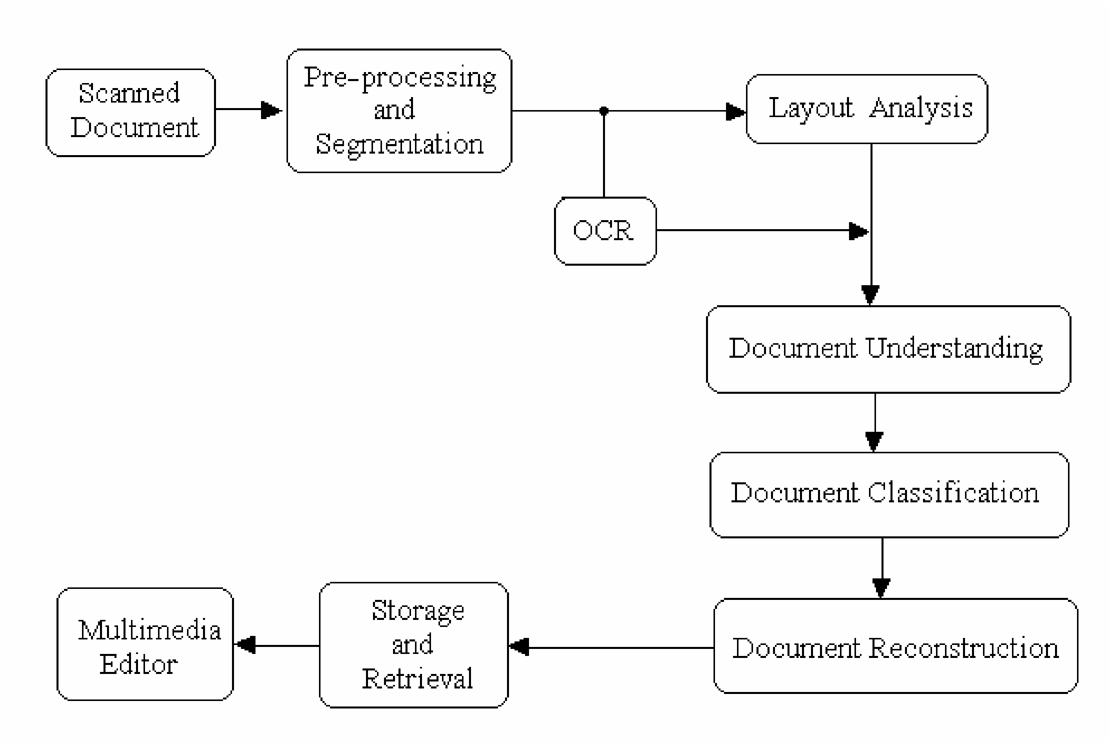


Figure 1.1 The Block Diagram of a Document Management System

In the 'Pre-processing and Segmentation' stage of Figure 1.1, an important step or process is binarization, which should ideally separate the text from the background by setting the text to 0 (black) and setting the background to 1 (white). The quality of this binary image can directly affect the subsequent processing steps. All black pixels are assumed to be foreground or useful pixels while all white pixels are assumed to contain no information and are background pixels. The black foreground pixels may subsequently be processed as binary pixels (to save computation time) or may be restored to their original (non-white) grey scale value to allow the extraction of detailed features. In many applications all significant details of the handwriting or printing, including faint skate-on and skate-off pen strokes at the beginning and end of strokes must be retained as these features often contain essential features necessary for recognition or validation. Binarization is performed using an appropriate thresholding technique. A number of researchers have obtained interesting results in this area. However, the results are not ideal, and most of the reported techniques normally deal with constrained image types, such as printed images, handwritten images, or engineering drawing maps with clear text on a homogeneous white background.

The resulting binary image is then segmented into independent words and subsequently used in the 'Layout analysis' stage as well as the 'Optical Character Recognition (OCR)' stage. Layout analysis is the process of constructing a layout hierarchy of document components whereby the document is constructed in terms of paragraphs, text columns and graphic sections. The document understanding step seeks to detect the logical properties of the layout components, such as title, author, abstract, etc. This step can be considered as the classification of layout components. Ejiri (1989) [26] showed that document understanding can be summarized in three steps:

1. Extracting features and analyzing structures;
2. Matching these results with models;
3. Controlling the process using the models or the matched results.

The 'Optical Character Recognition (OCR)' stage recognizes characters in the binary image. The OCR results are sent as textual information to the 'Document Understanding' stage for capturing the layout structure of a document, geometric information and textual feature.

The 'Document Classification' stage aims to categorize documents into user-defined classes. For example, in a library, a librarian may want to classify the documents according to different authors, or different genre, and in a postal sorting office, an officer may want to classify letters according to different destination countries or cities, etc.

After extracting the information about the document during the 'Document Classification' stage, the document can be reconstructed.

Finally, since textual, graphic and layout information is managed, the document can be efficiently stored and subsequently retrieved. The 'Multimedia Editor' is used to edit a new digital version of the document for keyword searching on the website or in the digital library.

Layout analysis and OCR modules require clean noise-free binary images which retain all useful information. The qualities of the binary image and word separation accuracy are significant to the subsequent processing. Binarization and segmentation are arguably the two most important steps in a document processing system.

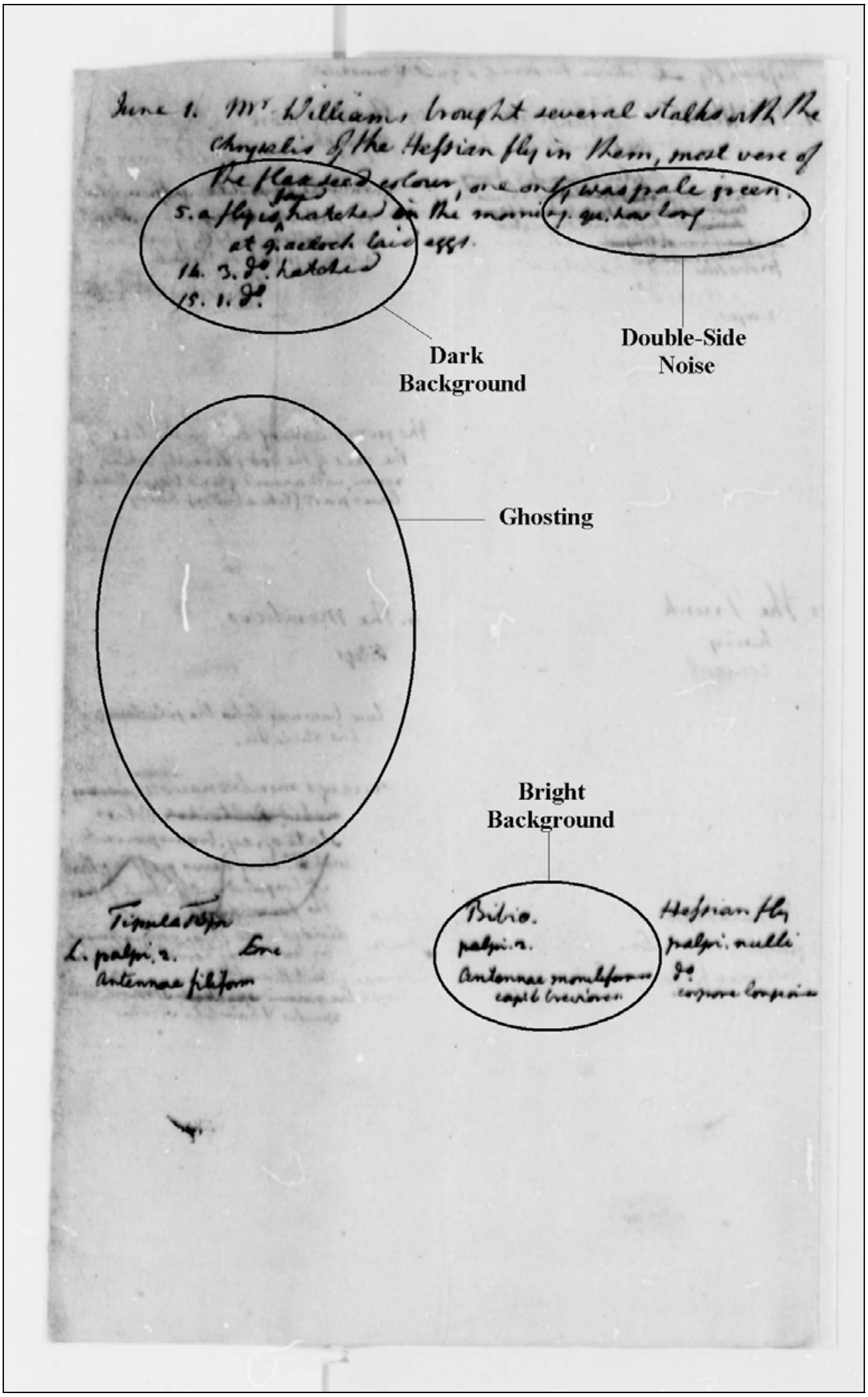
There are currently millions of historical documents stored in museums, libraries and government record offices all over the world. These documents contain important and interesting information that was written and recorded in handwritten letters and notes, during the past few hundreds years before the invention of typewriters and even printing. In order to ensure the preservation of these delicate documents, whilst also providing wider access to scholars and researchers, the documents are frequently scanned and made available as high-resolution images. Given the current state-of-the-art in computer recognition and processing of script, most of these historical documents are currently impossible to read automatically. To facilitate future automatic searching and analysis of the words and content in the documents, it is necessary to separate the useful pixels containing document content such as handwriting, drawings, pictures and other information representing useful artifacts, from the background pixels representing the paper on which the useful information is written.

This is a non-trivial task as the documents are frequently degraded due to poor storage and have been damaged over time resulting in content which is often difficult for a human to decipher. Historical handwritten document images frequently contain handwriting that was written by ink pen hundreds of years ago. The long storage time and adverse storage environments at some time in their past make the pen strokes fade or run and the quality of the document paper degrade, even producing some spots or darkened areas due to mould or bacterial growth.

As shown in Figure 1.2, a typical historical document image may include double-sided noise where writing on the reverse side of the paper has soaked through the paper and merged with writing on the front of the paper, ghosting noise where writing on the reverse side has soaked through giving the appearance of writing on the front of the paper, and varying background contrast due to varying changes in the paper colouration over time. These are the typical problems encountered in historical document images.

In order to automatically process handwriting in the wide range of historical document images, it is first necessary to separate the handwriting from the background and then separate the individual handwritten lines and words. Many historical document images contain florid handwriting, which frequently exhibits extravagant loops in ascenders, descenders and upper case letters as shown in Figure 1.3(a)~(d). These often result in touching or overlapping of words on adjacent lines. Separating the lines and words is difficult as the overlapping words on adjacent lines are often degraded to such as extent by poor storage environment and other damage inflicted over several hundred years that they are difficult for a human to decipher.

The segmentation of touching or overlapping words on adjacent lines is an important stage in the processing of historical cursively written documents.



June 1. Mr. Williams brought several stalks with the Chrysalis of the Hessian fly in them, most were of the flaxseed colour, one only was pale green.
5. a fly hatched in the morning. qu. how long
at 9 o'clock laid eggs.
14. 3. do. hatched
15. 1. do.

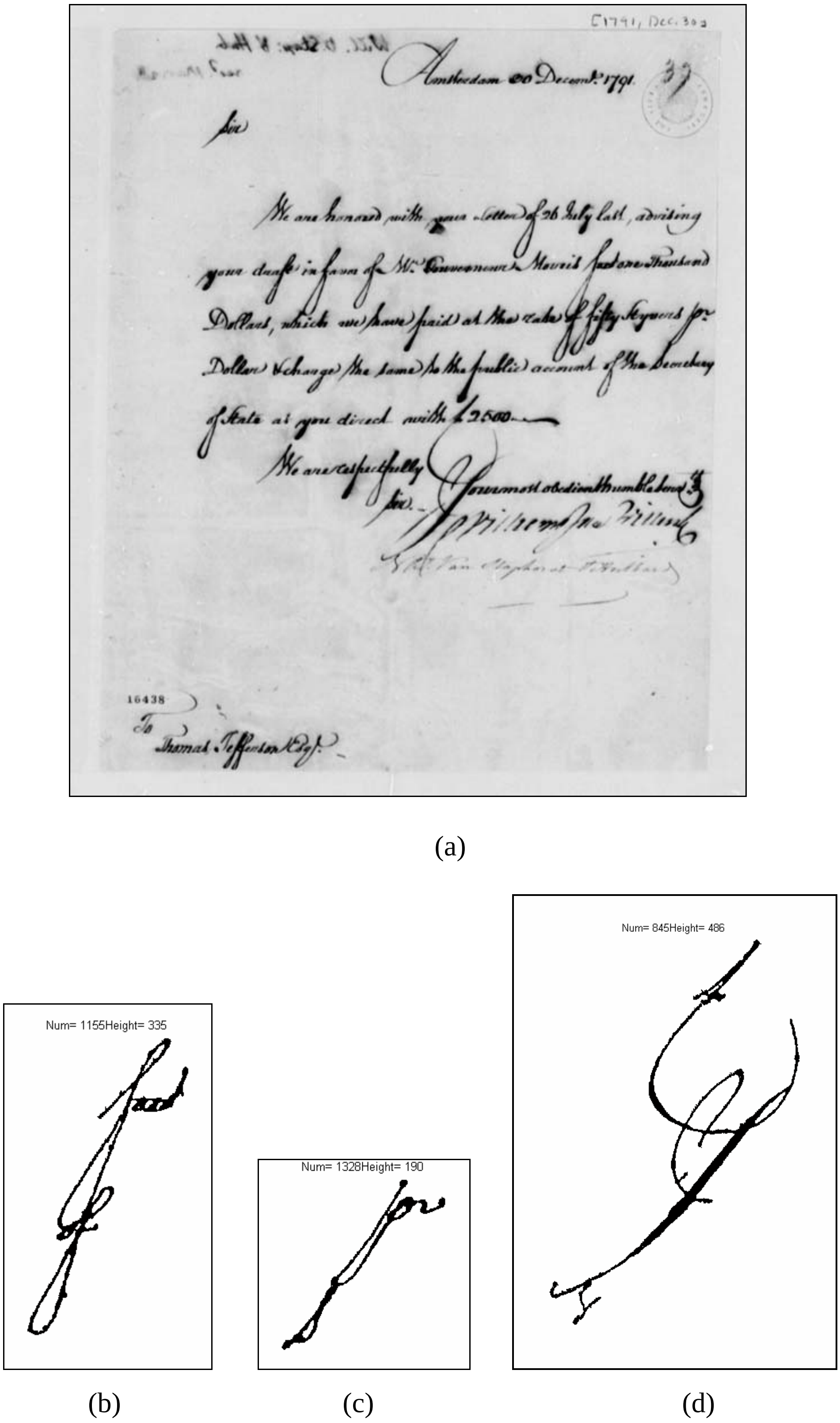
Tipula
L. palpi 2.
Antennae filiform

Bibio.
palpi 2.
Antennae moniliform
capite breviores

Hessian fly
palpi nulli
do
corpore longiora

Dark Background
Double-Side Noise
Ghosting
Bright Background

Figure 1.2 Example of a Typical Historical Document Image



[1791, Dec. 30]

Amsterdam 30 Decemr. 1791.

37

Sir

We are honored with your letter of 26 July last, advising your draft in favor of Mr. Gouverneur Morris for one thousand Dollars, which we have paid at the rate of fifty Stuyvers per Dollar & charge the same to the public account of the Secretary of State as you direct with f 2500 —

We are respectfully Sir

Your most obedient humble Servts

Wilhem & Jan Willink

N. & J. Van Staphorst & Hubbard

16438

To Thomas Jefferson Esqr.

(a)

(b) (c) (d)

Figure 1.3 (a) Typical Degraded Historical Document Image; (b)~(d) Connected character components extracted from image (a)

1.2 Objectives

Numerous techniques have previously been proposed for single-stage thresholding of document images to separate the written or printed information from the background. Whilst these global or local thresholding techniques have proven effective on particular sub-classes of documents, none is able to produce consistently good results on the wide range of documents and image qualities that exist in general or the image qualities encountered in degraded historical documents.

Also, whilst a number of separation techniques have proven effective at segmenting words correctly if the handwritten text lines are not overlapping or touching, none has been shown able to produce consistently good results on the wide range of document images containing touching or overlapping handwritten strokes.

The objectives of this thesis are:

1. To investigate, compare and evaluate thresholding algorithms for the separation of text from background in scanned document images. The documents to be studied are poor quality grey-scale images from several document types, including historical documents, forms, newspapers and cheques. The primary goal is to separate text for subsequent processing (for example OCR or forensic analysis). The study is not concerned with the location or separation of diagrams or pictures embedded within the text.
2. To propose and evaluate thresholding algorithms in order to separate text from background in scanned historical document images. The historical document images are ones which have become degraded due to age, handling or paper quality making the task difficult.

3. To investigate techniques to separate the overlapping words on adjacent lines of historical document images resulting from florid handwriting in ascenders, descenders and upper case letters.

1.3 Contributions of this Dissertation

There are many avenues for research in document processing. The study reported in this thesis has developed new algorithms to separate and segment foreground words in degraded document images and especially historical document images. This research has led to the following contributions and original results:

1. An improved QIR (Quadratic Integral Ratio) technique for extracting the handwritten text from noisy backgrounds. It is an effective global thresholding method for aged and poor quality grey-scale image.
2. A mean-gradient technique, which analyses the mean-gradient in local regions for different types of document images. The mean-gradient thresholding method was published in *Proc. 7th Int. Conf. on Document Analysis and Recognition*, Edinburgh, Scotland, Vol. 2, pp 859-865, 2003.
3. A multi-stage technique, which analyses the block information in local areas after which the most appropriate threshold method for that area is determined. Some thresholding techniques can only be effectively applied to one type of image. The multi-stage technique was published in *Proc. 17th Int. Conf. on Pattern Recognition*, Cambridge, United Kingdom, Vol.1, pp 445-448, 2004.

4. A decomposition technique depending on features extraction for degraded historical document image thresholding. These features can be usefully used in knowledge-based segmentation/separation. The technique was published in *Proc. 9th Int. Workshop on Frontiers in Handwriting Recognition*, Tokyo, Japan, pp 239-244, 2004; and in *IEE Proceedings on Vision, Image and Signal Processing,* Vol. 152, No. 6, pp 702–714, 2005.

5. New segmentation techniques, separating overlapping or touching words on adjacent lines in handwritten documents. The segmentation technique was published in the *Proc. 8th Int. Conf. on Document Analysis and Recognition*, Seoul, Korea, 2005; and is under review by the journal *Pattern Recognition Letters*.

1.4 Organization of this Thesis

Chapter 1 has provided an introduction to the project and describes the motivation, the need and the importance of the project. Objectives and the research contributions are also presented.

Chapter 2 presents a review of binarization and segmentation techniques reported in the literature. The general definition of thresholding is discussed, followed by its development. Subsequently, a review of existing techniques in this field is presented. The review includes pre-processing, global thresholding, local thresholding, post-processing and evaluation methods for each binarization technique. The last section of Chapter 2 presents a review of segmentation techniques.

Chapter 3 considers a global thresholding technique called QIR (Quadratic Integral Ratio), and an **improved QIR method** is described. The last part of the chapter presents the experimental results of the improved QIR algorithm.

Chapter 4 proposes a **mean-gradient based local adaptive algorithm**. A background subtraction method is used to remove noisy or patterned backgrounds. Experiments on four types of images: historical document images, form images, newspaper images and cheque images are illustrated and compared with existing techniques.

Chapter 5 proposes a **multistage structure** for degraded historical document images. A **decompose multistage algorithm** is described to separate degraded text from degraded background. An **improved decompose multistage algorithm** is also described. Experiments on historical document images and comparison with existing techniques are presented.

Chapter 6 proposes an **ICA (Independent Component Analysis) based segmentation algorithm** to separate touching and overlapping lines in degraded document images as encountered in many historical documents. Experiments on historical document images are presented.

Chapter 7 summarises the results and presents the conclusions of the study. Avenues for future research are also included.

Chapter 2

Review of Document Thresholding and Segmentation Methods

This chapter first provides a literature review of pre-processing methods for document images, and then moves on to review thresholding methods for document images. Finally, segmentation methods for handwritten images are reviewed.

2.1 Review of Pre-processing Methods

Everyday, many of us spend a considerable amount of our time processing paper documents. This document processing involves human visual processes, which have become highly adapted to extracting information in the presence of noise. Currently, no automatic visual system is able to compete with human vision in complex images where noise and degradation is present. However, there are a number of shortcomings with human vision when applied extensively: it is time consuming, prone to errors when applied for long periods, and costly. Because of the increasing competition in the business world, a quicker and more convenient computerized processing algorithm for documents is in demand. Before high level automatic document processing can be achieved, pre-processing methods need to be applied to the original scanned document images.

To achieve automatic processing of documents, an original paper document is usually scanned or imaged to produce a grey-scale image which is subsequently binarized. There are, as noted in Chapter 1, numerous features, which contribute towards corrupting an image with various kinds of noise, including characteristics of the scanner or camera, non-uniform illumination, etc. In addition, the document may be degraded due to inappropriate storage and handling.

To remove the unwanted noise from the image, a pre-processing step can be applied before it is presented to the binarization step. In the pre-processing step, a noise pre-filter will normally be applied to the scanned image.

Fan et al. (2001) [28] described three desirable properties of a pre-filter:

- It should completely remove the impulsive noise which could cause misclassification due to directly using the thresholding algorithm;
- It should make the selection of threshold more accurate and robust;
- It should retain edges and corners of objects for subsequent processing.

Among numerous methods of pre-filtering, Gaussian and median filtering are the most commonly used filters. The Gaussian filter can be used to filter out spurious points (noise) in an image, and to soften edges. In general, the median filter allows high spatial frequency detail (edges and other sharp details) to pass while removing noise on images. It is good at removing impulsive noise, but its output is ragged and not smoothed. One of the major problems with the median filter is that it is relatively expensive, in terms of computation time, and complex to compute.

Fontanot & Ramponi (1993) [31] proposed a simple quadratic filtering technique for pre-processing. They selected a filter with fixed coefficients, but the spatially invariant filter was incapable of adapting to the images that had spatially varying statistics. From the experimental result presented, the quadratic filter may not work well for images, which have different characteristics in different locations.

Fan et al. (2001) [28] presented a coplanar pre-filter as a pre-processing method. The method exploited the co-planarity of the grey-level distribution of neighbouring pixels, which can remove impulsive noise, apply piecewise smoothing and achieve sharp edge preservation. The output of the coplanar filter is sharper than those of Gaussian and median filters. The coplanar filter outperforms Gaussian, median and the quadratic filters, it allows piecewise smoothing and can better retain the edges and corners.

Using a pre-filter before thresholding can significantly improve the performance of the subsequent document image binarization result.

In Section 3.2.1, a new Window-based Enhancement Method is proposed for enhancing contrast of the document to obtain a sharper histogram. Gonzalez & Woods (1993) [33] proposed a closing method, which is an effective method to remove unwanted image background. This method will be described particular in Section 4.2, as a pre-processing step of the proposed local adaptive mean-gradient technique, which will be proposed in Chapter 4.

2.2 Review of Thresholding Methods

The primary task in any document processing system is to extract the information (usually text) from the background. This is commonly referred to as thresholding.

In the published literature, thresholding methods can be divided into two types: those that use static (global) thresholds and those that use dynamic (local) thresholds.

Neural network approaches to image thresholding have also been investigated by researchers in recent years.

❖ **Definition 1: Thresholding** - a technique, which transforms a grey-scale image into its binary version representing objects and background, respectively. It can be categorized into two methods: Global & Local.

Sezgin & Sankur (2004) [78] categorized the thresholding techniques into six groups according to the extracted information on which they were based.

1. Histogram shape information
2. Histogram entropy information
3. Clustering of grey-level information
4. Image attribute information
5. Spatial context information
6. Local adaptation

2.2.1 Global Thresholding Algorithm

Global thresholding is the simplest binarization method. Only one threshold value is selected for the entire image according to globally extracted information.

❖ **Definition 2: Global Threshold** - One grey scale thresholding value is chosen for the whole image.

Histogram and entropy based global techniques are two of the mature global methods.

The global thresholding methods can be classified as:

1. Histogram shape-based techniques

2. Histogram entropy-based techniques
3. Others

2.2.1.1 Histogram Shape-Based Techniques

The global information of an image can be obtained from the histogram of the image. The features of the peaks, valleys, and curvatures in the smoothed histogram are analyzed using histogram shape based global thresholding techniques to determine the final global threshold value.

Thresholding algorithms based on histogram shape information are summarized in Table 2.1. The histogram-based techniques produce good results on bimodal histogram images, even when the image has strong noisy background. However, these histogram-based techniques cannot solve the problem when the histogram of the objects overlaps with that of the background.

The popular global thresholding algorithms seek to find the best single cutting point of the histogram to separate the object pixels from the background, Gonzalez &Woods (1993) [33].

Otsu's method (1979) [64], is an early, but still popular histogram-based global threshold algorithm. It proposed a criterion for maximizing the variance of the between-class of pixel intensity to perform thresholding. It is a class separability method. It can achieve good performance with simple documents where the background and foreground are clearly distinct in the histogram. However, Otsu's algorithm is very time-consuming for image binarization because of its inefficient formulation of the between-class variance, and the performance varies with data sets.

Cheriet et al. (1998) [12] applied a recursive Otsu's algorithm for cheque image segmentation. This technique is more flexible than Otsu's thresholding. At each recursion, the technique segments the object with the lowest intensity from the input image. This process continues until there is the darkest object left in the image.

Table 2.1 Histogram Shaped Based Global Thresholding Techniques

	Technique Author	**Year**	**Main Features**	**Major Field**
1	Otsu	1979	Class Separability Method	Image
2	Boukharouba et al.	1985	Distribution Function Based	Image
3	Sezan	1985	Histogram Shape Information Analysis	Image
4	Papamarkos et al.	1994	Distribution Function Based	Image
5	Don	1995	Noise Attribute Feature-Based	Printed & Mail Image
6	Liu & Srihari	1997	Stroke-Based	Printed Image
7	Cheriet et al.	1998	Recursive Otsu Algorithm	Cheque Image
8	Solihin & Leedham	1999	IR Class Based	Handwritten Image
9	Negishi et al.	1999	Automatic Reference System	Old Handwritten Literature Image
10	Liao et al.	2001	A Fast Version of Multilevel Otsu	Picture Image

Don (1995) [22] proposed a histogram-based global technique that utilizes noise attribute features from the image. It is based on a noise model to overcome the difficulty created when some objects do not form prominent peaks in the histogram. The experimental results show that this method is very effective for printed and mail document images.

Liu & Srihari (1997) [54] proposed a local thresholding method based on global histogram feature extraction. It was a two-level threshold selection. The technique was based on texture features (stroke-width based) to extract characters from the run-length featured texture (histogram based) background. It is a flexible method in global techniques for printed document images.

Negishi (1999) [61] presented an automatic reference system based on Otsu's histogram thresholding method to extract text and then the connected components were extracted using a labelling method. This method can process very large size images because of its advantage of saving memory and reducing processing time, but it cannot provide good results for degraded historical handwritten documents with different types of noise in different areas of the image because of its global process characteristics.

Solihin & Leedham (1999) [83] described two global techniques: Native Integral Ratio (NIR) and Quadratic Integral Ratio (QIR). These two histogram shape based methods developed from Integral Ratio, which is a new class of global thresholding techniques.

The Integral Ratio Class [83] is a collection of all two-stage thresholding techniques, it classifies the pixels of an image into three classes: foreground, background, and a class between them, which is defined as the fuzzy class. The new class can be used for all two-stage thresholding methods. In the first stage, the fuzzy area of the histogram is found by using an Integral Ratio Function. In the second stage, a final threshold value T is found in the range of the fuzzy area. The NIR and QIR use different Integral Ratio functions to find the range of the fuzzy class in the first stage. In the second stage, features of the writing instrument were used to determine the final threshold value. However, in general a global threshold cannot be selected before the fuzzy class was found.

QIR achieved better results than NIR from the experiments carried out. But both of them use a simple method to determine the final threshold value. An improvement of the QIR method in the second stage is carried out as part of the research reported in this thesis and described in Chapter 3. Because of the limitations of histogram-based global thresholding, NIR and QIR do not work well in some images, which do not have obvious bimodal peaks in the histogram (A bimodal histogram indicates that the two peaks correspondingly refer to the object pixels and the background pixels).

The QIR algorithm was applied to four different kinds of image database:

1) Historical Document Images; 2) Form Document Images; 3) Cheque Document Images; 4) Newspaper Document Images. The experimental results are presented in Chapter 3.

The original grey-scale scanned handwritten image ‘Image_1’ in Figure 2.1 was processed by the QIR technique and is shown in Figure 2.2:

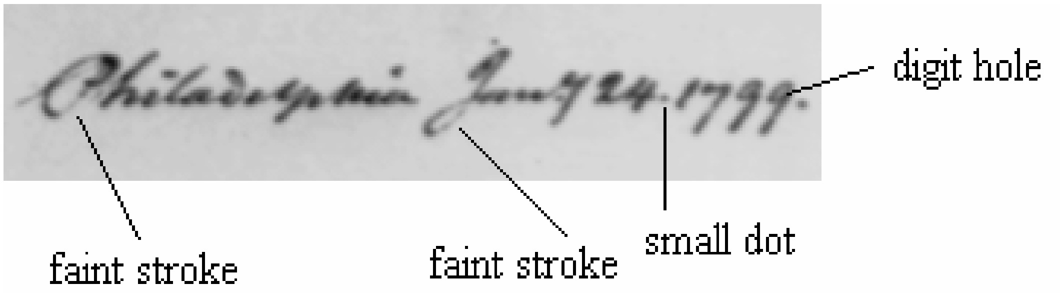


Figure 2.1 Original handwritten ‘Image_1’

There are several difficult aspects of thresholding in Figure 2.1: faint stroke in 'C' and 'g', small dot between words, and faint hole in characters and digit numbers.

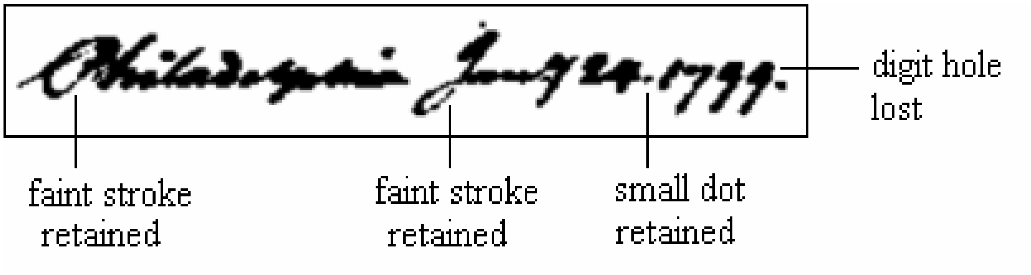


Figure 2.2 Binarization of 'Image_1' by using QIR Algorithm

The QIR global algorithm retains the faint strokes and small dots in Figure 2.2. However, the faint holes in words are lost. Further analysis of the QIR algorithm is described in Chapter 3.

Boukharouba et al. (1985) [6] and Papamarkous et al. (1994) [67] proposed histogram distribution function based techniques; Sezan (1985) [79] presented a global technique based on histogram shape information analysis. Liao (2001) [53] proposed a criterion for maximizing the between-class variance of pixel intensity to perform picture thresholding.

Histogram shape based thresholding methods, as its name suggests, are based on shape information of the grey-scale histogram, which are efficiency of computation for thresholds of an image. However, they cannot provide good results for degraded historical document images which include different complex characteristics in different image areas.

2.2.1.2 Histogram Entropy-Based Techniques Review

Entropy can be used to separate the global thresholding classes. For example, the optimal threshold value can be calculated by maximizing the sum of the foreground and background entropies. i.e., maximally separate region intensities of the foreground and background. Shannon's entropy information theory has been used in image segmentation for many years.

Histogram entropy information based global thresholding techniques are summarised in Table 2.2. The techniques use entropy information of the foreground and the background to find the global thresholding value. These techniques have good performance when applied to high contrast images.

Pun (1981) [70] presented a maximum entropy based method. It used Shannon's concept to define the entropy of an image, which assumed that an image is presented by its grey level histogram. Pun used this concept to derive an expression for an upper bound of a posteriori entropy. The expression was finally used to threshold an image.

Kapur et al. (1985) [43] achieved an improvement of Pun's (1981) [70] method. It is a histogram analysis and maximum entropy based global technique, which uses the maximum of the sum of the entropy of the grey-level distribution of the foreground and background. Kapur et al. (1985) [43] showed a technique, which demonstrated good performance for picture images.

Johannsen & Bille (1982) [36] and Sahoo (1988) [75] also presented two global thresholding methods based on maximum Shannon Entropy. Leung & Lam (1996) [51] presented a spatial information analysis based technique that maximized the segmented scene spatial entropy to determine the final thresholding value.

Abutaleb (1989) [1] used an entropy-based global technique for picture images, and Beghdadi (1995) [3] proposed an entropy-based global technique by using a block source model.

Table 2.2 Histogram Entropy Based Global Techniques

	Technique Author	**Year**	**Main Features**	**Major Field**
1	Pun	1981	Maximum Shannon's entropy method;	Image
2	Johannsen & Bille	1982	Maximum entropy method	Image
3	Kapur at el.	1985	Improved of no. 1.	Picture Image
4	Pal & Pal	1988	Higher order entropy based	Picture Image
5	Sahoo	1988	Maximum Shannon's Entropy based	Image
6	Abutaleb	1989	Entropy-based	Picture Image
7	Chang et al.	1994	Minimize Relative Entropy based;	Image
8	Beghdadi et al.	1995	Entropy based using a block source model	Image
9	Brink	1995	Spatial information analysis; Minimum spatial entropy based	Image
10	Leung & Lam	1996	Spatial information analysis; Maximum segmented scene spatial entropy	Image
11	Wang	2002	Relative entropy based; Extend from no. 7	Picture Image

Pal & Pal (1988) [66] presented a higher order entropy method for object extraction and summarized several entropy methods used in image processing.

Chang et al. (1994) [9] presented a spatial context thresholding algorithm that minimizes relative entropy. It used spatial dependence (co-occurrence probability) of the pixels. Wang (2002) [98] extended this relative entropy based technique to picture images.

Brink (1995) [8] proposed a minimum spatial cross-entropy based global thresholding algorithm. The cross-entropy was interpreted as a measure of data consistency between the original and the binarized images. This is a spatial information analysis technique.

Histogram entropy-based techniques utilize spatial probability information like maximum entropy of foreground-background regions (e.g. [70] and [43]), or minimizing the cross-entropy (e.g. [8]) between the original and binarized image to produce optimal global thresholding.

2.2.1.3 Other Global Techniques

In addition to histogram and entropy informatics theory, other global information can also be used in global thresholding.

Global thresholding techniques based on other information are summarised in Table 2.3. The first three methods in Table 2.3 are edge information based global thresholding techniques. These techniques are sensitive to noise. The other methods in Table 2.3 are based on different information.

Weszka & Rosenfeld (1978) [100] used a co-occurrence matrix to compute the sum of transitions between the object and background.

Kohler (1981) [48] and Wang & Haralick (1984) [96] proposed an edge analysis based global multi-threshold technique.

Kittler & Illingworth (1986) [47] presented a popular method that uses a minimum error criterion for threshold determination. This technique models the histogram as two distributed classes that form bimodal histogram peaks. The threshold value is found by minimizing the classification error between the two classes.

Gorman (1994) [34] presented a global method based on local features, which used connectivity-preserving measure. The method determined threshold values at intensities between region levels where region break-up is least likely.

Table 2.3 Other Global Thresholding Techniques

	Technique Author	**Year**	**Main Features**	**Major Field**
1	Weszka & Rosenfeld	1978	Co-occurrence matrix is used to compute the sum of transition between the object and the background.	Image
2	Kohler	1981	Edge Analysis	Image
3	Wang & Haralick	1984	Edge Analysis	Image
4	Kittler & Illingworth	1986	Minimum Error Based	Image
5	Gorman	1994	Connectivity-Preserving Measure;	Document Image
6	Said et al.	1996	Extended Formal Model	Business Forms & Check Images
7	Gu et al.	1998	Differential Top-Hat Transform Based	Scene Image

Said et al. (1996) [76] used an extended formal model.

Gu et al. (1998) [35] used the differential top-hat transform to extract characters from scene images. The main disadvantage is that the parameters used in the technique are fixed, so problems arise when the input images have variable contrast.

In summary, the main advantages of global thresholding techniques are: they are convenient to use and generally involve fast computation. However, using global thresholding techniques may lose some local information and may not achieve good results in the presence of noise, high illumination, low or variable contrast images, but its computation is relatively fast. If an image has non-uniform background, then a local thresholding technique may need to be applied.

2.2.2 Local Thresholding Algorithms

Local thresholding determines the threshold values for a single or group of pixels locally and adjusts the threshold dynamically based on the neighbourhood information of each pixel.

❖ **Definition 3: Local (Adaptive or Dynamic) Threshold –** Compute a separate threshold for each pixel based on the neighborhoods of the pixel.

Local techniques work well for poor contrast and noisy background images. However, the techniques are not always good when applied to clean documents. They work slowly and will often enhance and retain useless background information.

In a grey-scale image, the neighbourhood information of each pixel can represent the variation of the local area. The local area can be a window, which is centred on the pixel, or can be a line of the input image. There are various descriptions for variation of the local area in different local thresholding algorithms.

Local thresholding can be classified into five classes according to the local area analysed:

1. Local Adaptive-Based
2. Clustering-Based
3. Object Attribute-Based
4. Spatial Information-Based
5. Training-Based

2.2.2.1 Local Adaptative-Based Techniques

Local thresholding methods, which adapt the threshold value depending upon the local image characteristics, can be classified as local adaptative-based techniques. Locally adaptive based local thresholding techniques are summarised in Table 2.4.

Table 2.4 Locally Adaptive Based Local Thresholding Techniques

	Technique Author	Year	Main Features	Major Field
1	Giuliano et al.	1977	Contrast measure method used five 9*9 windows	Printed Document Image
2	Yasuda et al.	1980	Base on Local intensity change (max/min and contrast)	Check Image
3	White & Rohrer (Integrated Function Algorithm)	1983	Gradient-Based method	Printed & Handwritten Image
4	White & Rohrer (Dynamic Threshold Algorithm)	1983	Running average is used as a threshold at each pixel location.	Printed & Handwritten Image
5	Bernsen	1986	Locally based on neighbors	Image
6	Palumbo	1986	A second deviation contrast measure based method (Improvement of No.1)	Document Image
7	Niblack	1986	Local Mean and Local Standard Deviation;	Image
8	Yanowitz & Bruckstein	1986	Binarized using the Threshold Surface;	Image
9	Eikvil et al.	1991	The pixels inside a small window S are threshold on the basis of clustering of the pixels inside a large window *L*.	Document Image
10	Manjunath & Chellappa	1991	Feature extraction based (Boundary Detection)	Images
11	Parker	1991	Gradient-Based	Hand printed Image

12	Chen & Takagi	1993	Run length coding based	Image
13	Kamel & Zhao	1993	Stroke-Model-Based	Printed / Handwritten Document Image
14	Wang & Pavlidis	1993	Adaptive Thresholding	Handwritten Image
15	Liang & Ahmadi	1994	Morphological Operation	Printed Image
16	Ohya et al.	1994	Connected-component Analysis	Image
17	Oh	1995	Bright-Average Scheme	Document
18	Trier & Taxt	1995	Edge Information-Based; Improvement of “ Integrated Function Algorithm”	Circuit Image
19	Liu et al.	1997	Edge Detection Based	Check Image
20	Djeziri et al.	1998	Use *Filiformity* as the closet description of the local feature	Check Image
21	Ye et al.	1999	Local thresholding based on Bernsen (1986) method	Check Image
22	Zhao & Yan	1999	Utilize geometrical features combined with grey level analysis	Blueprint Image
23	Yang & Yan	2000	A modified logical method	Printed Image
24	Zhang & Tan	2001	Improvement of Niblak’s Algortihm	Handwritten Image
25	Rangsanseri & Rodtook	2001	Update locally the threshold value whenever the Laplacian sign of the input image changes along the raster-scan line.	Printed Image
26	Sharma	2001	Adaptive Linear Filtering Scheme	Printed Image
27	Ye et al.	2001	Stroke-Model-Based	Check Image

Giuliano et al. (1977) [32] used a contrast measure local thresholding technique for document images.

Yasuda et al. (1980) [105] proposed a local intensity change based local technique. It was determined by image contrast.

White & Rohrer (1983) [101] presented two thresholding algorithms: Dynamic Threshold Algorithm and Integrated Function Algorithm. Dynamic Threshold Algorithm uses a running average as a threshold at each pixel location.

An Integrated Function Algorithm is a gradient-based local thresholding technique. Firstly, the differential image of the original image is obtained. Secondly, an integrated function is used to extract the sequence of +, -, and 0 in the differential image, and take the pixels between sequences “+ -” and “- +” as objects. The Integrated Function Algorithm is based on boundary characteristics to reject unwanted background patterns. The algorithm is sensitive to noise and edges.

Trier & Taxt (1995) [92] presented an improvement of the Integrated Function Algorithm. It is a local technique based on edge information.

Bernsen (1986) [4] proposed a local thresholding technique based on neighbours. It has been proven to be a fast algorithm. The disadvantage is that it does not work on background regions with varying grey-level and ‘ghost’ images. The contour pixels, which distinguish dark backgrounds from brighter ones, are classified as object pixels, called a ‘ghost’.

Ye et al. (1999) [106] developed a local thresholding technique based on the method proposed by Bernsen (1986).

Niblack (1986) [60] presented his method based on the calculation of the local mean and local standard deviation. The method obtains an adaptive threshold by varying the slicing level according to the local variance of the image. The size of the neighbourhood should be small enough to preserve local details, but at the same time large enough to suppress noise. There is a weighted value for the local standard deviation, which is used to adjust how much of the total printed object boundary is taken as a part of the given object. In Trier & Taxt (1995) [91], the Niblack algorithm was reported to perform well and fast. The threshold is decided by:

$$T(x,y) = m(x,y) + K \times s(x,y) \qquad \text{Eq (2-1)}$$

where *m(x, y)* and *s(x, y)* are the average of a local area and standard deviation values, respectively.

Zhang & Tan (2001) [110] proposed another improved version of Niblack's method. In comparison with the Niblack (1986) [60] algorithm, the improved version is less sensitive to noise and edges; and it is good at shadow boundary detection. The threshold is determined by:

$$T(x,y) = m(x,y) \times \left[1 + K \times \left(1 - \frac{s(x,y)}{R}\right)\right] \qquad \text{Eq (2-2)}$$

where *K* and *R* are empirical constants.

Zhang & Tan's improved Niblack method is used in Chapters 4 and 5 for comparing performance with other thresholding methods.

A local gradient-based thresholding algorithm was proposed by Yanowitz & Bruckstein (1989) [103]. In this algorithm, the thresholding was determined by interpolating the image grey level at points where the gradient is high, indicating probable object edges. The gradient map of the image was used to point at well-defined portions of object boundaries. It demonstrated that both the location and grey levels at these boundary points are a good choice for determining local thresholds. In Trier & Taxt (1995) [91], Yanowitz's algorithm was found to be the best of eleven thresholding techniques they investigated.

Eikvil et al. (1991) [25] proposed an adaptive thresholding technique. According to the technique, the pixels inside a small window *S* were thresholded on the basis of clustering of the pixels inside a large window *L*. The principle of the technique is that a large window *L* with a small window *S* in the centre is moved across the image in a zig-zag fashion, in steps equal to the size of *S*. Window *S* is labelled as print (foreground) or background on the basis of the clustering of the pixels inside *L*.

Parker (1991) [68] used a local intensity gradient to propose a local thresholding technique. The local intensity gradient was based on the local contrast estimated by the grey level difference between a pixel and its neighbours. This technique produced good performance for handwritten and printed document images.

Kamel & Zhao (1993) [44] used a stroke-model-based technique, which emphasized the local linearity of character strokes. According to the experimental results, the technique works better for printed or handwritten document images than the methods proposed by Palumbo et al. (1986) [65], which is an improvement of a method in Giuliano et al. [32] and Liu et al. (1997) [55]. This method extracts baselines from bank checks by searching for a pair of opposite edges in a predefined distance to threshold the image. Yang & Yan (2000) [104] proposed an improved logical method for printed images.

Wang & Pavlidis (1993) [97] assumed the grey-scale image as a surface with topological features corresponding to the shape features of the original image. Each pixel of the image was classified as: peak, pit, ridge, ravine, saddle, flat, or hillside. Rules can be built based on the estimated first and second directional derivatives of the underlying image intensity surface. The characters can be extracted according to the rules and the features.

Liang & Ahmadi (1994) [52] used morphological operation for text extraction and background pattern removal. It worked effectively for printed images.

Ohya et al. (1994) [62] used a connected-component analysis based local thresholding algorithm. The algorithm assumed the grey level within character objects is homogeneous. The disadvantage is that fixed parameters are used, so that problems will arise when the input images have variable contrasts.

Oh (1995) [63] used a bright-average scheme, which replaced the average value in local windows by the average of the maximum in the columns and rows.

Djeziri et al. (1998) [20] proposed a local thresholding technique based on extracting handwritten information by means of an intuitive approach that is close to human visual perception, defining a topological criterion specific to handwritten lines, which is called filiformity. The experimental results showed that it has better performance than Palumbo (1986) et al. [65], Liu et al. (1997) and White & Rohrer (1983) (Dynamic Threshold Algorithm) [101] for check images.

Zhao & Yan (1999) [109] utilized geometrical features combined with grey level information analysis to determine the local threshold values for blueprint images.

Rangsanseri & Rodtook (2001) [72] proposed a local thresholding technique, which used Laplacian sign to describe the variation of the local area. This local technique is based on updating the threshold value whenever the Laplacian sign of the input image changes along the raster-scan line. In an image, where the sign of a pixel has changed, we find the edges of the image's components, so the Laplacian sign of the image presents the physical alteration of the image. The Laplacian sign image is derived from Differential of the Gaussian (DoG) algorithm. The technique is sensitive to edges and noise, because it is based on edge information. It works well when applied to noiseless low contrast images. However, the binary output will retain some useless noise if the technique is applied to an image with noisy background.

Ye et al. (1999) [107] also used a stroke-model-based local thresholding algorithm. The stroke-based model needs to account for both grey level features and local geometric features.

In all these techniques, the threshold values are determined locally or adjusted dynamically based on the criterion function of each pixel. The criterion functions include the local intensity change [60], stroke width of the characters [44], surface fitting parameters [97], gradient and edge information [103]. The locally adaptive techniques show better performance in variable background intensity images due to non-uniform illumination. In existing locally adaptive-based techniques, they applied the same process to a whole image even through there were always different characteristics at different areas in the image.

2.2.2.2 Clustering-Based Techniques

Clustering-based methods cluster the grey level samples into two parts as background and foreground (objects) or alternately model the grey level as two Gaussian distributions.

Cluster-based techniques are summarised in Table 2.5.

Table 2.5 Clustering Based Local Thresholding Techniques

	Technique Author	**Year**	**Main Features**	**Major Field**
1	Taxt et al.	1989	Clustering Measure;	Document Image
2	Qian & Chew	2001	Mapping of Double-Sided	Handwritten Image

Taxt et al. (1989) [88] presented a clustering measure thresholding technique. The image was divided into non-overlapping windows; a mixture of two Gaussian distributions approximated the histogram in each window, where the pixels were classified using the quadratic Bayes' classifier.

Qian & Chew (2001) [71] calculated the difference between both sides of an image to estimate the threshold value. The experimental results showed that the technique had relatively better performance for double-sided noisy background in handwritten images. However this method cannot work for other noise encountered in degraded historical documents, like 'ghosting' and 'dark background'.

Clustering-based techniques perform better when removing signal-dependent noise. The images binarized using these methods show no obvious loss of useful information.

2.2.2.3 Object Attribute-Based Techniques

Object attribute-based methods search for a measure of similarity such as fuzzy similarity, shape, edges, and number of objects between the grey-level and the binarized images.

Object attribute based techniques are summarised in Table 2.6.

Table 2.6 Object Attribute Based Local Thresholding Techniques

	Technique Author	**Year**	**Main Features**	**Major Field**
1	Tsai	1985	Moment-Preserving Thresholding	Picture Image
2	Hertz & Schafer	1988	Attribute Information Analysis	Picture Image

Tsai (1985) [94] investigated moment-preserving local thresholding. The threshold value at which the original and threshold images have the closest moments was determined to be the optimal threshold.

Hertz & Schafer (1988) [38] proposed an image attribute information analysis based technique. Matching of edge fields of the original grey level image was used to determine the binarized image.

The techniques search for a measure of similarity between the grey level image and the binarized image. They perform better with variable background intensity and very low contrast images but changeless background texture.

2.2.2.4 Spatial Information-Based Techniques

The spatial methods use the probability mass function models taking into account the correlation between pixels on a global scale.

Spatial information based techniques are summarised in Table 2.7.

Table 2.7 Spatial Information Based Local Thresholding Techniques

	Technique Author	**Year**	**Main Features**	**Major Field**
1	Deravi & Pal	1983	Second-order statistic	Image
2	Kittler & Illingworth	1985	Image statistic based	Image

Deravi & Pal (1983) [21] proposed a second-order statistics based technique, which is an earlier threshold method depending upon local statistics.

Kittler & Illingworth (1985) [46] presented an image statistic based technique. This technique produces good results in some cases but the performance varies with data sets.

Spatial information based techniques use probability models to account the correlation between pixels. The techniques show good results with variable background intensity and very low contrast images.

2.2.2.5 Training-Based Techniques

Sample training techniques are an active research topic widely used in OCR areas. They can also be used in binarization systems.

Two techniques are summarised in Table 2.8.

Table 2.8 Training Based Local Thresholding Techniques

	Technique Author	Year	Main Features	Major Field
1	Chigusa et al.	1992	Hopfield Neural Network Based	Image
2	Guo & Ma	2001	Hidden Markov Model (HMM) Based	MachinePrinted Image

Chigusa et al. (1992) [16] introduced a new binarization method for image segmentation based on a Hopfield Neural Network. The threshold of each neuron can be decided adaptively to be proportional to the grey level value of the corresponding input pixel, and then the stable state of the network will be assigned as the output image.

Guo & Ma (2001) [23] separated handwritten material from printed text using a Hidden Markov model (HMM). It uses small set of machine printed texts for classifier training.

Training based local thresholding techniques work better on variable background with complex patterns. However, the training based techniques are difficult to implement in hardware because of their complex computation structure and limitation of the training database selection.

The training based thresholding method is more natural than some simple binarization techniques, and should have a significant future in image segmentation.

Sezgin & Sankur (2004) [78] summarises 44 binarization methods. Before that, Sahoo et al. (1988) [75] did a survey of more than 20 thresholding methods. Weszka (1978) [99], Weszka & Rosenfield (1985) [100], Palumbo et al. (1986) [65] and Zhang (1996) [108] summarise most of the binarization methods published.

There are several published papers that have summarized and compared thresholding methods. Trier & Jain (1995) [90] described details of eleven local thresholding techniques and presented a goal directed evaluation of binarization methods. In their review the methods reported by Eikvil et al. (1991) [25] and Bersen (1986) [4] were the most promising techniques.

2.2.3 Post-processing Methods

After thresholding, because of the limitations of some techniques, the binarization result may retain some unwanted ‘ghost’ noise, which will affect the visual result. The aim of the post-processing for binary images is to remove binary noise and false information to improve the final results.

Trier & Taxt (1994) [89] introduced a post-processing method in their paper. Firstly, the average gradient value was calculated at the edge of each printed object. Secondly, if objects had an average gradient below a threshold, then the objects can be labeled as misclassified and will be removed. The threshold was chosen by experimentation.

Yang & Yan (2000) [104] used run-length information to detect the false information.

Post-processing is quite important for improving the binary result. After this step, the binarization procedure in Figure 1.1 is completed. The evaluation of the binarization system is presented in the Section 2.2.4.

2.2.4 Evaluation Methods for Binarization

A human expert evaluates the binarized images according to his or her visual criteria. However, to formulate a binarization evaluation, some quantitative effectiveness measures are needed. In this section, some definitions for binarization effectiveness are described.

Salton (1989) [77] described the well knows measures of *Recall* and *Precision* to show the effectiveness of a text retrieval system.

❖ **Definition 4 –** ***Recall*** is the ratio of the number of relevant words in document returned to the total number of relevant words in the document for the user query in the collection.

Recall = Correctly Detected Words / Total Actual Words;

❖ **Definition 5 –** ***Precision*** is the ratio of the number of relevant words in document returned to the total numbers of words in the document for a given user query.

Precision = Correctly Detected Words / Total Detected Words.

This standard measure is used to compute the effectiveness of document algorithm analysis. The criterion of what denotes correct detection is an important point, which needs to be defined.

Solihin & Leedham (1999) [83] suggested that the binary result should have the following characteristics:

1. Retain all details of the handwriting, including faint skate-on and skate-off pen strokes at the beginning and end of strokes;
2. The background paper image, which contains dark colored and/or patterned background should be removed, and
3. Handwriting produced by a wide variety of pens such as a fountain pen, ballpoint pen, fiber-tip pen, and pencil is retained.

In Chapter 4, ***Recall*** and ***Precision*** are used to evaluate a new thresholding algorithm.

2.3 Review of Segmentation Methods

❖ **Definition 6: Segmentation** – In image analysis, **segmentation** is the partitioning of a digital image into multiple regions (sets of pixels), according to a given criterion. The goal of segmentation is typically to locate objects of interest and is sometimes considered a computer vision problem.

Grey-level thresholding is the simplest segmentation process, although numerous segmentation techniques have previously been proposed for document images. These segmentation techniques can be categorized into five classes:

1. Shape-based Techniques
2. Model-based Techniques
3. Region-based Techniques
4. Neural Network-based Techniques
5. Other Techniques

These techniques are reviewed below.

2.3.1 Shape-based Techniques

In **shape-based techniques**, the handwriting shape or texture information is analysed to determine the segmentation points. Handwriting segmentation using shape-based techniques normally results in an unnatural shape due to overlapping or touching stroke reconstruction. Shape-based segmentation techniques for handwritten images are summarised in Table 2.9.

Sabourin & Plamondon (1988) [73] presented a technique based on the extraction of textured regions characterized with local uniformity in the orientation of the gradient for handwritten signature images.

Fujisawa et al. (1992) [30] proposed a new method, which was based on the extraction of connected pattern components, and then spatial interrelations between components were measured to group them into meaningful character patterns, from images. Stroke shapes are analyzed in the case of touching characters. Machii et al. (1993) [56] presented a stroke information based method, which works well in online handwritten data segmentation.

In Strathy et al.'s technique (1993) [87], the contour chains were subdivided into four regions: valleys, mountains, holes, and open regions. Individual points of interest in the outer contour were then identified. The separating path was assumed to pass between some pair of these significant contour points. Naoi et al. (1994) [59] investigated a global shape interpolation based method, which can provide good results for handwritten characters overlapping a border. Congedo et al. (1995) [18] proposed a multiple segmentation algorithm based on contiguous row partition which works sequentially on the binary image until an acceptable segmentation is obtained for handwritten numeric strings. Amara et al. (1996) [2] presented a new method of recursive estimation of linear segments or circles parameters by a Kalman extended filtering method for handwritten drawing images.

Shi et al. (1997) [82] investigated a new system, composed of several document analysis modules: a pre-processing module, a segmentation module based on a thorough stroke analysis using contour representation of the strokes, and a recognition module. The new system has been proposed for unconstrained handwritten numeral strings.

Bishnu & Chaudhuri (1999) [5] presented a new method of recursive contour following in one of the zones across the height of the word to determine the extents within which the main portion of the character lies for handwritten *Bangla* (an national language of Bangladesh and the second most popular language in India) text. Kim et al. (2001) [45] proposed a gap-based method for segmenting of handwritten Korean text lines into separate words.

Mestetskii et al. (2002)[58] investigated a new technique based on approximating a binary raster image with a set of polygons and building a continuous skeleton of those polygons. These polygons and skeletons are then used to extract lines, remove spots and artifacts, extract words from lines and extract strokes from words. Feldbach & Tonnies (2003) [29] presented a new method, which combines semantic a-Priori-Knowledge with local shape features for historical documents (handwritten dates).

In Suwa & Naoi’s (2004) [85] technique, candidate segmentation paths are calculated using both graph theory techniques and heuristic rules. The boundaries of the digits are calculated to make the width of the touching strokes uniform.

Table 2.9 Shape - based Segmentation Techniques

Technique Author	Year	Main Features	Major Field
Sabourin & Plamondon	1988	Based on the extraction of textured regions characterized with local uniformity in the orientation of the gradient	Handwritten signature image
Fujisawa et al.	1992	Based on extraction of connected pattern components Stroke shapes are analyzed in the case of touching characters.	Form images
Machii et al.	1993	Strokes information based method	On-line handwritten data
Strathy et al.	1993	The contour chains are subdivided into four regions: valleys, mountains, holes, and open regions. Individual points of interest in the outer contour are then identified. The separating path is assumed to pass between some pair of these significant contour points (SCPs).	Touching unconstrained handwritten digits
Naoi et al.	1994	Global shape interpolation based method	Handwritten characters overlapping a border
Congedo et al.	1995	A multiple segmentation algorithms based on contiguous row partition work sequentially on the binary image until an acceptable segmentation is obtained	Handwritten numeric strings
Amara et al.	1996	A method of recursive estimation of linear segments or circles parameters by a Kalman extended filtering method	Handwritten Drawing
Shi et al.	1997	The system is composed of several document analysis modules: pre-processing module, segmentation module based on a thorough stroke analysis using contour representation of the strokes and a recognition module.	Unconstrained Handwritten Numeral Strings
Bishnu & Chaudhuri	1999	A method of recursive contour following in one of the zones across the height of the word to find out the extents within which the main portion of the character lies.	Handwritten (Bangla Language)
Kim et al.	2001	**A** gap-based method for the segmentation of handwritten Korean text lines into separate words	Handwritten Korean word
Mestetskii et al.	2002	Based on approximating a binary raster image with a set of polygons and building a continuous skeleton of those polygons. Polygons and skeletons are then used in extraction of lines, removing of spots and artifacts, extraction of words from lines and extraction of strokes from words.	Handwritten Documents
Feldbach & Tonnies	2003	Combining Semantic A-Priori-Knowledge with Local Shape Features	Handwritten Historical Documents from the 18th and 19th century
Suwa & Naoi	2004	Candidate segmentation paths are calculated using both graph theory techniques and heuristic rules.	Simply and multiply connected digits

Feldbach & Tonnies (2003) [29] presented a new method, which combines semantic apriori knowledge with local shape features for historical documents (handwritten dates).

In Suwa & Naoi's (2004) [85] technique, candidate segmentation paths are calculated using both graph theory techniques and heuristic rules. The boundaries of the digits are calculated to make the width of the touching strokes uniform.

The shape-based segmentation techniques are based on analysis of edges and texture etc information. This kind of technique works well on documents that have independent handwritten words placed all over the article but always fail on overlapping or touching words.

2.3.2 Model-based Techniques

Model-based techniques are, as the name suggests, based on creating models of segment handwriting. The model-based techniques normally focus on separating handwriting from other obvious interference marks cutting across the text. The similarity of these stroke characteristics with overlapping words on adjacent lines, this type of techniques is unlikely to provide good results.

A summary of published model-based techniques is given in Table 2.10.

Table 2.10 Model - based segmentation techniques

Technique Author	Year	Main Features	Major Field
Su et al.	1997	Based on multiple direction projection planes model	Printed documents that have interference marks and cutting across text
Cheung et al.	2002	Model-based, bi-directional matching Segmentation	*bb* and *bs* datasets in the CEDAR database with 633 handwritten city name images

Su et al. (1997) [86] proposed a new method based on a multiple direction projection plane model, which works well on interference marks and lines cutting across the text in printed documents.

Cheung et al. (2002) [14] presented a model-based, bi-directional matching segmentation method for segmenting handwritten city names written in the address on an envelope.

Model-based segmentation techniques depend on the characteristics of the images under investigation. These techniques work well for images that conform to the assumption of the model.

2.3.3 Region-based Techniques

Region-based techniques are based on analysis of the information in different classes of regions of the image.

A summary of region-based techniques is given in Table 2.11.

Chen & Wang’s method (2000) [13] used background and foreground analysis to segment handwritten connected numeral strings.

Breuel (2001) [7] proposed an algorithm, which evaluated a large set of curved cuts through the image of the input string using dynamic programming and selected a small “optimal” subset of cuts for segmentation.

Zhao et al. (2001) [111] proposed a two-stage approach for handwritten Chinese characters. A string was first coarsely segmented according to the background skeleton and vertical projection after image pre-processing. The segmentation paths are evaluated at the fine segmentation stage.

Table 2.11 Summary of Region-Based segmentation techniques

Technique Author	Year	Main Features	Major Field
Cheriet et al.	1992	Background region based method	Connected digits
Chen & Wang	2000	Use background and foreground analysis to segment handwritten connected numeral string	Connected numeral string; 150 images of handwritten for test
Breuel	2001	The CPSC algorithm evaluates large set of curved cuts through the image of the input string using dynamic programming and selects a small “optimal” subset of cuts for segmentation.	Letters in Roman alphabets,
Zhao et al.	2001	A two-stage approach: A string is first coarsely segmented according to the background skeleton and vertical projection after a proper image pre-processing. At the fine segmentation stage that follows, the segmentation paths are evaluated.	Handwritten Chinese Character
Sadri et al.	2004	Foreground and background features based method	Unconstrained Handwritten Numeral Strings

Sadri et al. (2004) [74] presented a foreground and background features based method for unconstrained handwritten numeral strings.

Region-based techniques focus on connected handwriting, but do not provide good results on overlapping strokes on adjacent lines.

2.3.4 Neural Network-based Techniques

Neural network-based methods are summarised in Table 2.12.

Yamamoto et al. (1993) [102] proposed a Hopfield neural network based technique, which can segment handwritten Japanese character strings. Japanese characters are difficult to segment into one meaningful string due to the irregular character size and disposition. This method is essentially a pre-processing method, based on expressing

Japanese characters as energy function in the Hopfield neural network so that the network can perform segmentation. This method was reported to achieve a correct segmentation rate of 82.8%.

Table 2.12 Neural Network - based Segmentation Techniques

Technique Author	Year	Main Features	Major Field
Yamamoto et al.	1993	Hopfield neural network based	Handwritten Japanese character string
Eastwood et al.	1997	Neural network based	Handwritten words
Hamid et al.	2001	Pre-segmentation points for connected blocks of characters. A neural network is subsequently used to verify the accuracy of these segmentation points.	Arabic handwritten text image
Chen & Zhen	2002	HMM based method	Handwritten Chinese character
Daekeun & Gyeonghwan	2003	A neural network based method	Handwritten numeral strings

Eastwood et al. (1997) [24] proposed neural network training method for handwritten word segmentation. The neural network was training using a large database of hand-segmented images. However, it cannot work well on overlapping or connected handwritten words.

In Hamid & Haraty's (2001) method [37], pre-segmentation points for connected blocks of characters were identified first. Then a neural network was subsequently used to verify the accuracy of the segmentation points.

Chen & Zhen (2002) [15] proposed an HMM based method for handwritten Chinese characters. The method first adopted the HMM method to produce the segmentation paths and applied two rules to reduce the redundant paths, then the left candidate paths dissected the text line in radicals or pseudo-radical components. In the second stage, the proposed method used three new criteria: aspect ratio, gap ratio and longer edge criteria to calculate the clustering cost matrix and used a dynamic programming technique to produce the optimal clustering scheme. This method proved very effective for offline handwritten Chinese text segmentation.

Daekeun & Gyeonghwan (2003) [19] proposed a neural network based method for various types of touching characters observed frequently in numeral strings. Potential segmentation points were located using the neural network by active interpretation of the features collected from the primitives. Also, the run-length coding was used to represent images. However, this method was not effective for overlapping words on adjacent lines.

The neural network-based techniques can work on different languages but most of them cannot work well for connected handwriting, and even less so for overlapping words.

2.3.5 Other Techniques

Other techniques are based on shape analysis, word/stroke models and region classification and have also been used to segment text.

Table 2.13 summarises some of the other segmentation techniques.

Plamondon & Privitera (1999) [69] proposed a segmentation method that partly mimics the cognitive-behavioral process used by humans to recover motor-temporal information from the image of a handwritten word.

Veloso et al. (2000) [95] presented a morphological based method for handwritten words, and Tripathy & Pal (2004) [93] proposed the water reservoir approach – the concept is applied to unconstrained Oriya handwritten text. (The alphabet of the modern

Oriya script consists of 11 vowels and 41 consonants.) The basic characters of Oriya script are shown in Figure 2.3.

Table 2.13 Other Segmentation Techniques

Technique Author	Year	Main Features	Major Field
Plamondon & Privitera	1999	A segmentation method that partly mimics the cognitive-behavioral process used by human subjects to recover motor-temporal information from the image of a handwritten word.	Data bases of handwritten word images representing names of international cities written by six different writers where each writer was asked to write six different city names
Veloso et al.	2000	A morphological based method	Handwritten words
Tripathy & Pal	2004	Water reservoir – concept based scheme	Unconstrained Oriya handwritten text

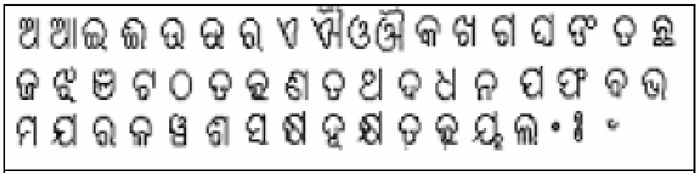


Figure 2.3 Basic Character of Oriya Script

2.4 Summary

The quality of the thresholding result when separating foreground from background is decisive for subsequent analysis of the document content. It requires retention of the full information content on a clear white background. In some applications, such as forensic document analysis, or scholastic analysis of the writing style, we are interested in the detailed greyscale or color variations of the pen strokes or printing. In others, such as studying the content of the documents, a binary image is sufficient.

In the review of the publications, no existing thresholding algorithm reported in the literature produces good performance for all types of document images, although this is an important requirement for a Document Management System.

In Chapters 3, 4 and 5, several knowledge-based techniques, which demonstrate superior performance for a wider range of document image types, are specified and evaluated.

Whilst the reviewed separation techniques have proven effective at segmenting words correctly if the handwritten text lines are not overlapping or touching, none has been shown able to produce consistently good results on the wide range of document images containing touching or overlapping handwritten strokes. The florid handwriting frequently encountered in historical documents exhibits extravagant loops in ascenders, descenders and upper case letters. These often result in touching or overlapping of words on adjacent lines.

In Chapter 6, a new segmentation algorithm, called the ICA (Independent Component Analysis) Segmentation Algorithm is proposed and investigated for this difficult task.

Chapter 3

Improvement of the QIR Algorithm

As observed in Chapter 2, global thresholding techniques provide fast and straightforward implementation in software or hardware for some forms of images.

The Quadratic Integral Ratio Technique (QIR) [83] (Solihin & Leedham, 1999) mentioned in Chapter 2, is an outstanding global two-stage thresholding approach based on histogram shape.

In order to achieve better and comprehensive research in document image segmentation, especially for degraded historical images, an improved QIR algorithm is examined in this chapter and a performance comparison of these two algorithms is described later in the experimental results.

3.1 Review of QIR Algorithm

Solihin & Leedham (1999) presented two global techniques: Native Integral Ratio (NIR) and Quadratic Integral Ratio (QIR). These two histogram shape based methods were developed to form a new class of global thresholding techniques: Integral Ratio. As described in Section 2.2.1.1, the Integral Ratio Class classifies the pixels of an image into 3 classes: foreground, background, and fuzzy class. In the first stage, parameters *A* and *C* (which separate foreground, fuzzy and background pixels respectively as illustrated in

Figure 3.1) are estimated using Integral Ratio Function; in the second stage, a final threshold value *T* was found in the range between *A* and *C*.

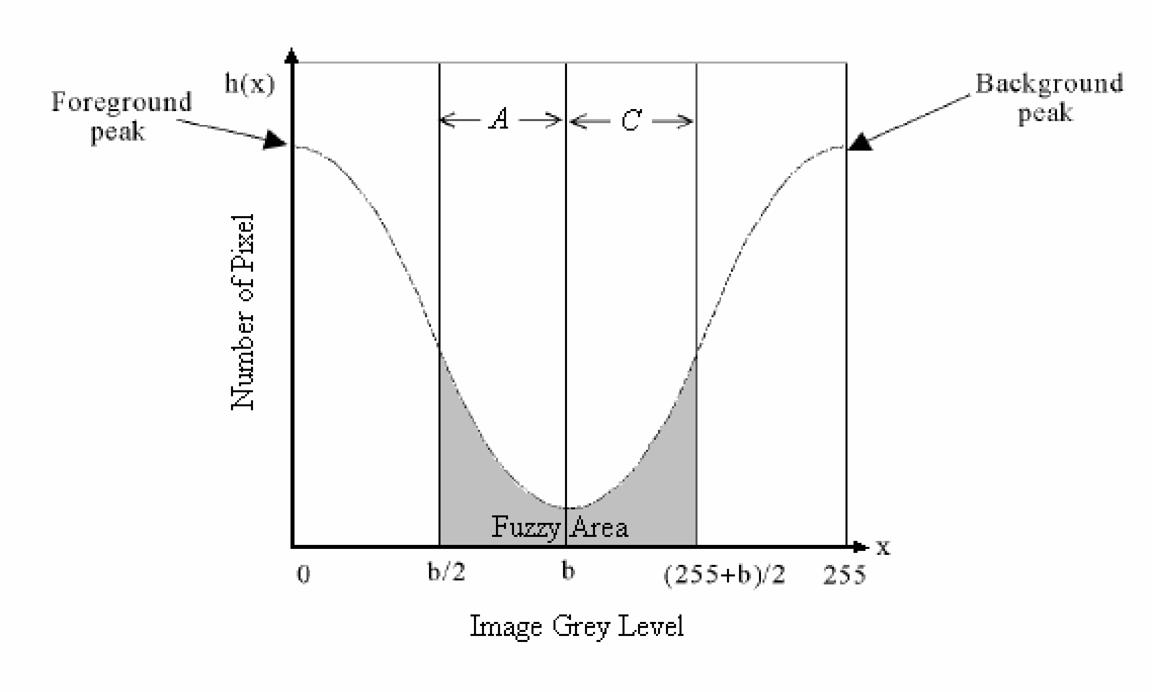
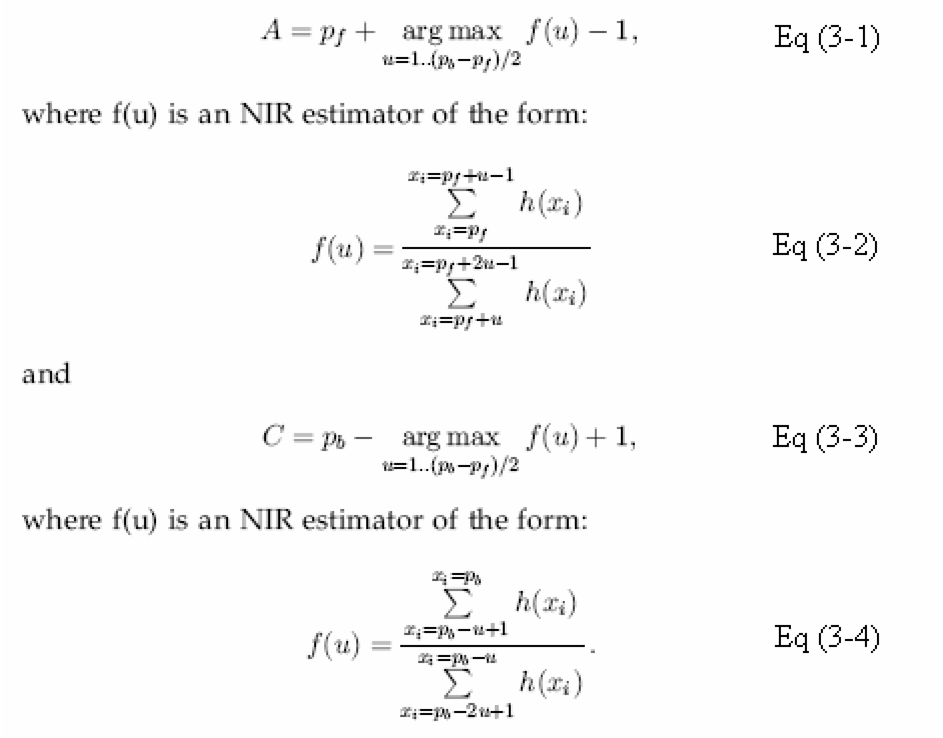


Figure 3.1 Fuzzy Area Determined by Parameter A and C

The NIR and QIR used different Integral Ratio functions to find parameters *A* and *C* in their first stage.

NIR works on real histogram to determine *A* and *C* based on Integral Ration functions as shown in Eq (3-1) and Eq (3-3). QIR deals with a quadratic approximation (as shown in Eq (3-5)) of the real intensity histogram instead of working on the real histogram as in NIR. As shown in Figure 3.2, three points are selected as P1 = (x_1,y_1), P2 = (x_2,y_2), and P3 = (x_3,y_3), and a quadratic curve can be generated by Eq (3-5) ~ Eq (3-9). The value of *U* can be calculated based on Eq (3-10) so that the final value *A* can be generated from Eq (3-11). Value *C* can be obtained in a similar procedure.



$$A = p_f + \underset{u=1..(p_b-p_f)/2}{\arg\max} f(u) - 1, \qquad \text{Eq (3-1)}$$

where f(u) is an NIR estimator of the form:

$$f(u) = \frac{\sum_{x_i=p_f}^{x_i=p_f+u-1} h(x_i)}{\sum_{x_i=p_f+u}^{x_i=p_f+2u-1} h(x_i)} \qquad \text{Eq (3-2)}$$

and

$$C = p_b - \underset{u=1..(p_b-p_f)/2}{\arg\max} f(u) + 1, \qquad \text{Eq (3-3)}$$

where f(u) is an NIR estimator of the form:

$$f(u) = \frac{\sum_{x_i=p_b-u+1}^{x_i=p_b} h(x_i)}{\sum_{x_i=p_b-2u+1}^{x_i=p_b-u} h(x_i)}. \qquad \text{Eq (3-4)}$$

where p_f is foreground peak value and p_b is background peak value as shown in Figure 3.1.

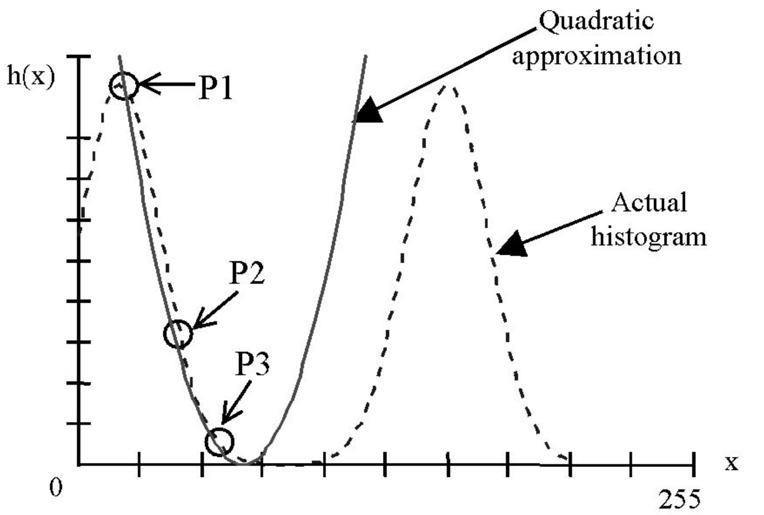


Figure 3.2 Quadratic Approximation of Foreground Curve in Histogram

$$h(x) = a(x-b)^2 + c \qquad \text{Eq (3-5)}$$

$$a' = \frac{\frac{y_3-y_1}{x_3-x_1} - \frac{y_2-y_1}{x_2-x_1}}{x_3-x_2} \qquad \text{Eq (3-6)}$$

$$b' = \frac{y_2-y_1}{x_2-x_1} - a'(x_2+x_1) \qquad \text{Eq (3-7)}$$

$$c' = y_1 - bx_1 - ax_1^2 \qquad \text{Eq (3-8)}$$

$$a = a' \quad b = -\frac{b'}{2a'} \quad c = \frac{4a' - b'^2}{4a'} \qquad \text{Eq (3-9)}$$

$$U = \frac{3(ab^2+c) - \sqrt{3}\sqrt{a^2b^4 + 4ab^2c + 3c^2}}{2ab}. \qquad \text{Eq (3-10)}$$

$$A = U + x_1 \qquad \text{Eq (3-11)}$$

where (x_1,y_1) (x_2,y_2) and (x_3,y_3) are P1, P2 and P3's coordinates respectively.

In the second stage, NIR and QIR used the quality of pen inking to find the final threshold value.

$$T = T_2 = \begin{cases} C - 1/2(C-A), & \text{if the writing implement is a felt tipped pen} \\ C - 1/10(C-A), & \text{if the writing implement is a ballpoint pen} \\ C, & \text{if the writing implement is a pencil} \\ C - 1/10(C-A), & \text{if the writing implement is not specified.} \end{cases}$$

In Solihin & Leedham's experiments, QIR performed better than NIR but both of them use a hasty method to determine the final threshold value. Because of the limitation of histogram-based global thresholding, NIR and QIR do not work well in some images that do not have the obvious bimodal peaks in the histogram. (A bimodal histogram has two peaks and the two peaks correspond to the object pixels and the background pixels respectively.) An improved QIR method is proposed for non-obvious bimodal peaks observed in the histograms of historical handwritten document images.

3.2 Proposed Improved QIR

Many images do not show obvious bimodal peaks in their histogram. For this kind of image, QIR cannot clearly locate the boundary between the foreground and the fuzzy area. An incorrectly chosen fuzzy area will affect the definition of the final threshold value. The original QIR algorithm fails in many cases due to degraded historical document images rarely contain obvious bimodal peaks in their histogram.

The outline of the improved QIR thresholding algorithm is:

1. Fuzzy area classification: Integral Ratio (IR) classification (first stage of original QIR technique);
2. Global cut point determination: Second Integral Ratio (IR) classification (instead of finding the final threshold value depending on the quality of pen inking).

3.2.1 Pre-processing

Although the histogram of document images can describe global information about an image, the unstable characteristic of document images will result in an unruly histogram. This problem means histogram-based thresholding methods cannot always produce good results, for example it may produce over-thresholded results. QIR also suffers from this problem as it seeks to find two peaks and model them. To improve the QIR method for thresholding document images, a new pre-processing method is proposed in this section: Window-based Enhancement Method (for enhancing contrast of the document to obtain a sharper histogram).

Enhancing the image before applying the histogram-based global thresholding algorithm will help to produce a more obvious bimodal histogram. The objective of this window-based enhancement is to enhance the contrast of words to the background of the document image so that the histogram becomes a bimodal without changing the major contents of the image. It is based on the average stroke width of words in the input document image.

The stroke width based window-based enhancement method can be divided into two steps.

(1) Enhancing the original image with a window function.

Using a $(w_1 \times W) \times (w_2 \times W)$ window to enhance the image by finding the maximum and minimum grey value in the window, where W is the average stroke width of the document words, w_1 and w_2 are weight values for adjusting the size of region to be enhanced, and default setting $w_1 = w_2 = 4$ according to the experience.

(2) Comparing the pixels' values minus the minimum grey value and the maximum grey value minus the pixels' values. If the former is larger, the *Pixel* is closer to the highest grey value than the lowest value in this window; hence we need to set the threshold value of the *Pixel* to the highest grey value. If the former is smaller, then set the *Pixel* to the lowest grey value.

There are numerous enhancement methods in image processing area. This method is chosen here is because it is fast and effective.

The enhancement of the original document image can provide a more obvious two-peak histogram, which can help to find the better cutting point for fuzzy area classification in the first stage of the QIR algorithm. However, the enhancement enhances both the useful and noise information in the original image. IR (Integral Ratio) class [83] is used to remove the unwanted noise after enhancement which is based on fuzzy area classification.

IR class is used to find the pixels which are classified in background class. If the pixel's grey-level is in the background class, it will be set to background value (for example, B=230 for 256 grey-level historical image).

3.2.2 Fuzzy Area Classification and Global Threshold Determination

In binarization of a historical handwritten document image, the main goal is to retain the useful information (the handwriting) and discard the unwanted information (background and noise). In order to achieve this goal, a good global cut point must be determined in the fuzzy area.

The second stage is the main improvement in the Improved QIR compared to the original QIR. The global cut point determination of the Improved QIR does not depend on the quality of pen inking, but classifies the fuzzy area by Integral Ratio (IR) classification.

In the first stage of the Improved QIR, the histogram is classified as three classes: foreground, background and fuzzy classes by using IR class. The boundary of the fuzzy area is determined by parameters *A* and *C*.

In the second stage, IR class is used again to determine the best global threshold in the fuzzy class. Equations Eq (3-12) & Eq (3-13) based on the IR (Integral Ratio) class are used to estimate the best threshold of the histogram in the fuzzy area:

$$f(t) = \frac{\int_{A}^{A+t} h(x)dx}{\int_{A+t}^{C} h(x)dx} \qquad \text{Eq (3-12)}$$

$$T = \operatorname{argmax} f(t) \qquad \text{Eq (3-13)}$$

where *A* and *C* have already been calculated in the first step.

Eq (3-13) is solved to get the final threshold value *T'*= *A* + *T*.

Some experimental results of the Improved QIR compared to Original QIR are shown in the next section.

3.3 Experimental Results

The experiment results of Original QIR and Improved QIR are shown in this section. The first part compares the binary results of original QIR algorithm which was applied on bimodal and single peak histogram images. The experiment results of the pre-processing and Improved QIR are shown in the second part.

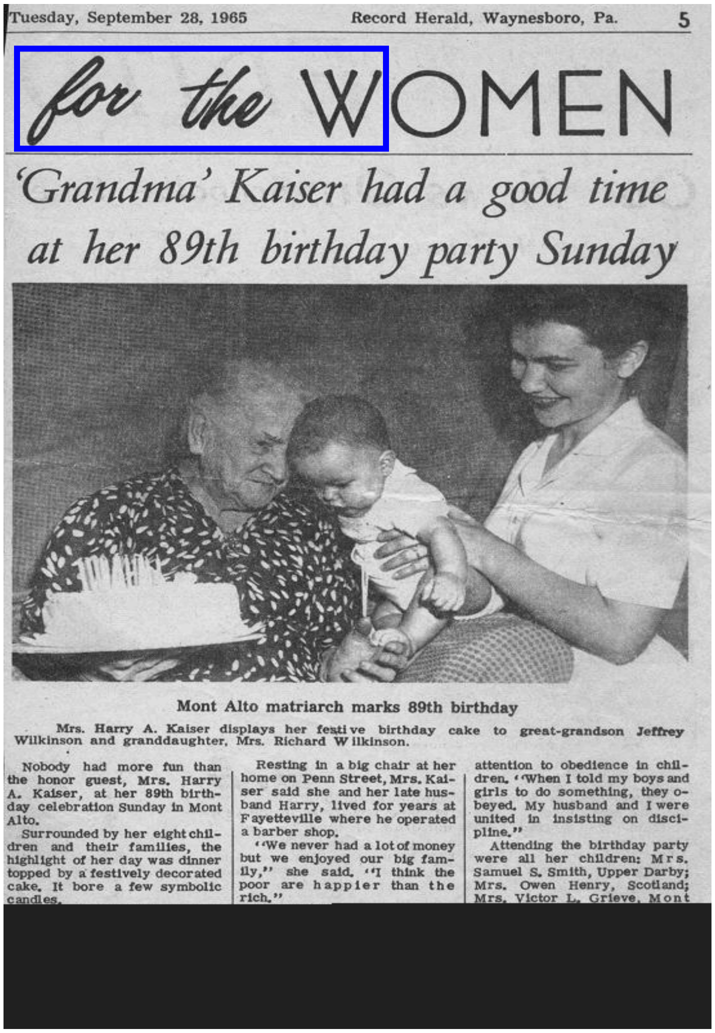
3.3.1 Experimental Results using the Original QIR Algorithm

QIR algorithm is a histogram-based thresholding algorithm. It works well on bimodal histogram image, but cannot provide good result on those images which include single peak in histogram.

3.3.1.1 Application on a Bimodal Histogram Image

The performance of the QIR algorithm is highly dependent on whether the images have obvious bimodal histograms. Figure 3.3 shows an example of a general newspaper image, which has obvious bimodal peaks in the histogram as illustrated in Figure 3.4.

There are two obvious peaks in the histogram (Figure 3.4) at grey levels of approximately 90 and 215. In order to see the detailed information of the words, the small black box region in Figure 3.3 is zoomed in as shown in Figure 3.5.



Tuesday, September 28, 1965 Record Herald, Waynesboro, Pa. 5

for the WOMEN

'Grandma' Kaiser had a good time at her 89th birthday party Sunday

Mont Alto matriarch marks 89th birthday

Mrs. Harry A. Kaiser displays her festi ve birthday cake to great-grandson Jeffrey Wilkinson and granddaughter, Mrs. Richard W ilkinson.

Nobody had more fun than the honor guest, Mrs. Harry A. Kaiser, at her 89th birthday celebration Sunday in Mont Alto.

Surrounded by her eight children and their families, the highlight of her day was dinner topped by a festively decorated cake. It bore a few symbolic candles.

Resting in a big chair at her home on Penn Street, Mrs. Kaiser said she and her late husband Harry, lived for years at F ayetteville where he operated a barber shop.

''We never had a lot of money but we enjoyed our big family,'' she said. ''I think the poor are happier than the rich.''

attention to obedience in children. ''When I told my boys and girls to do something, they obeyed. My husband and I were united in insisting on discipline.''

Attending the birthday party were all her children: Mrs. Samuel S. Smith, Upper Darby; Mrs. Owen Henry, Scotland; Mrs. Victor L. Grieve, Mont

Figure 3.3 Original grey-scale image

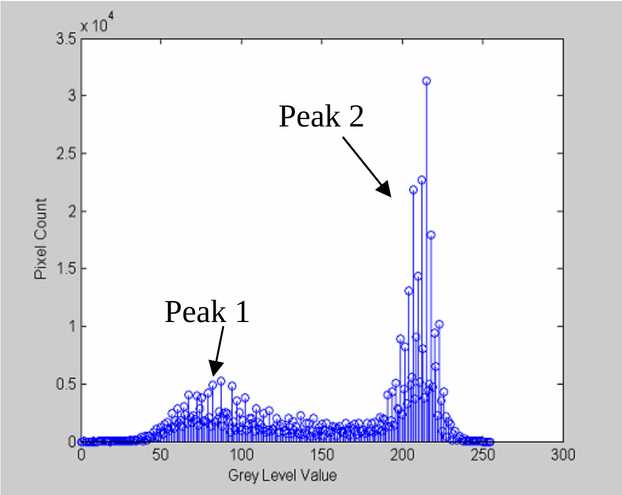


Figure 3.4 Histogram of Figure 3.3

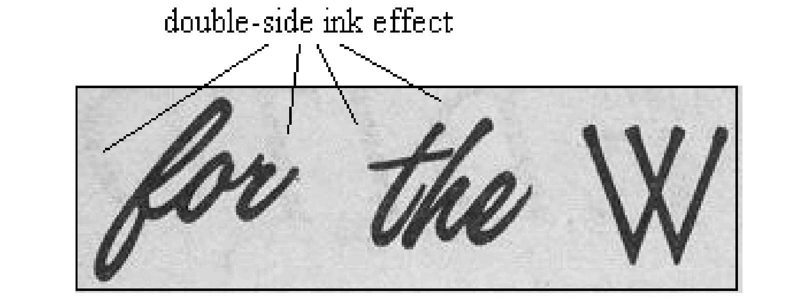


Figure 3.5 Enlarged Region of Figure 3.3

In Figure 3.5, there is a small amount of double-sided noise that affects the quality of the image. (Note: this becomes less pronounced when printed using a laser printer so is difficult to reproduce here.) The QIR thresholding algorithm's binary result of Figure 3.5 is shown in Figure 3.6.

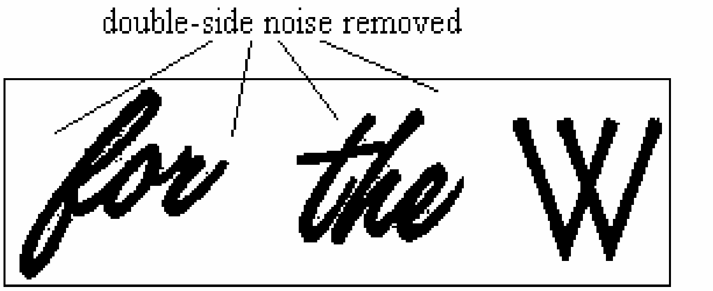


Figure 3.6 Binary Result_1 of Figure 3.5 Using Original QIR Algorithm

However, for images, which have single or less obvious peaks in the histogram, the QIR algorithm is not able to locate the best cutting point within the fuzzy area. The final binary result is not ideal. An example for a single-peak histogram is illustrated in the next section.

3.3.1.2 Application on a Single Peak Histogram Image

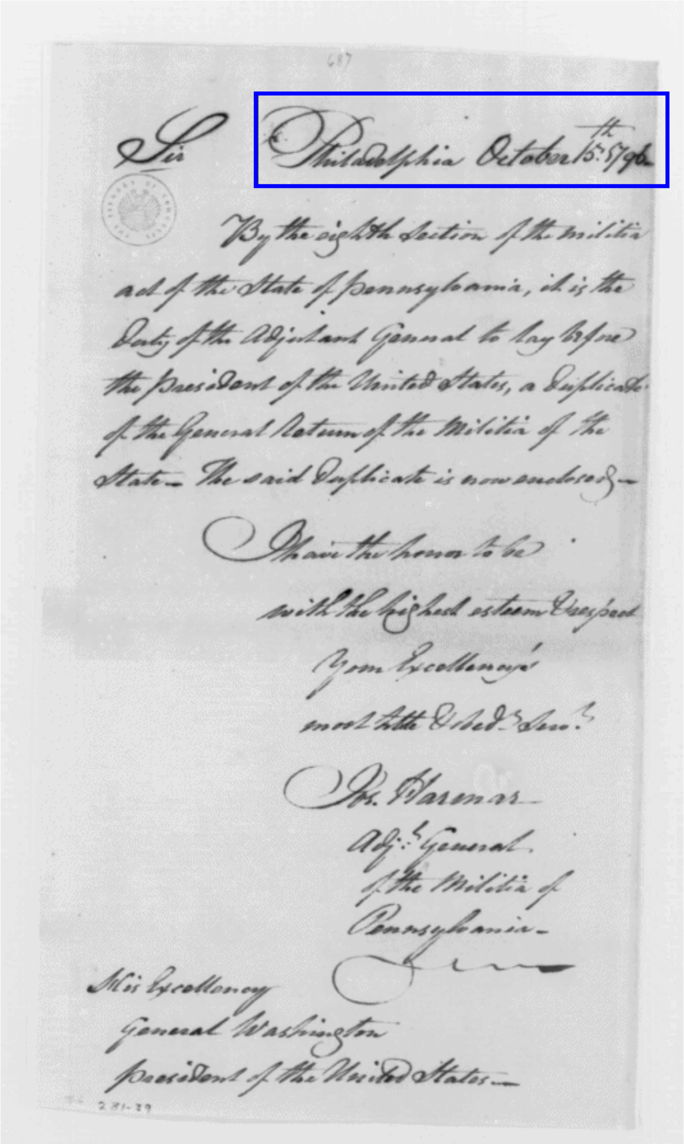
As mentioned in Chapter 1, there are several difficulties in thresholding: faint stroke, small dot, faint hole in words and digits. If an image does not have bimodal histogram peaks, even though the background of the image is very clear, the thresholding results of histogram-based global techniques will not be acceptable. Such as in the example is shown in Figures 3.7 ~ 3.10, which is an image with clear background, but its histogram is a single peak as in Figure 3.9. To observe the thresholding result in detail, the squared region in Figure 3.7 is shown in detail in Figure 3.8.

Figure 3.8 and it’s binary result (Figure 3.10) shows that Original QIR thresholding is less effective on some types of stokes: faint stroke, small dot, and faint digit hole.

Figure 3.9 has one obvious peak at the background (at greyscale 215) and one non-obvious peak at the foreground of the histogram (somewhere between greyscale 100 and 150). The binary result of Figure 3.8 after using the QIR thresholding algorithm is shown in Figure 3.10.

The binary result in Figure 3.10 shows that some faint strokes are lost; some smaller spaces inside words or digits are also not retained, resulting in the blurring of the words or the digits.

The original QIR algorithm produces good performance for images where there are obvious bimodal peaks in histogram. However, it cannot find idea final global threshold value in the second stage for those images which have unimodal peak histogram as shown in Figures 3.7 ~ 3.10 .



687

Sir Philadelphia October 15th 1796

By the eighth Section of the Militia act of the State of Pennsylvania, it is the Duty of the Adjutant General to lay before the President of the United States, a Duplicate of the General Return of the Militia of the State — The said Duplicate is now enclosed —

I have the honor to be
with the highest esteem & respect
Your Excellency's
most obt. & obed. Servt.
Jos. Harmar
Adjt. General
of the Militia of
Pennsylvania —

His Excellency
General Washington
President of the United States —

281-39

Figure 3.7 Second Grey-scale Image

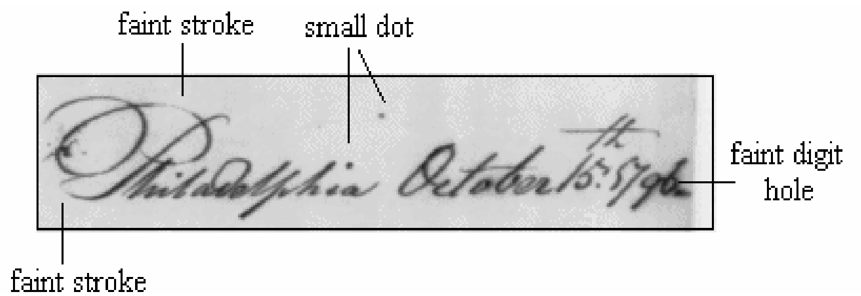


Figure 3.8 Enlarged Region of Figure 3.7

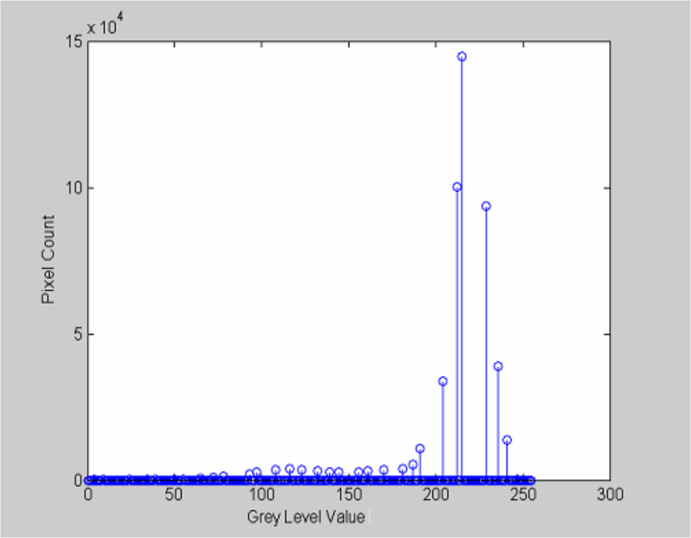


Figure 3.9 Histogram of Figure 3.7

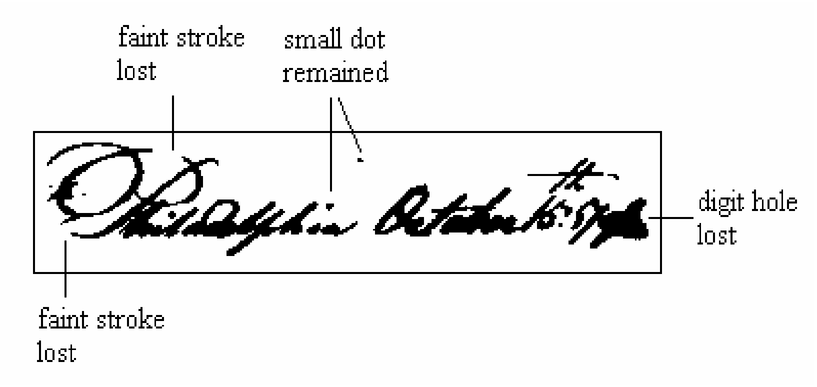


Figure 3.10 Binary Result of Figure 3.8 using the original QIR Algorithm

3.3.2 Experimental Results using the Improved QIR Algorithm

Most cases of degraded historical handwritten document images do not have an obviously bimodal histogram. For example, the single histogram peak of 'Image_A' (Figure 3.11(a)) which contains noise and handwritten characters is shown in Figure 3.12.

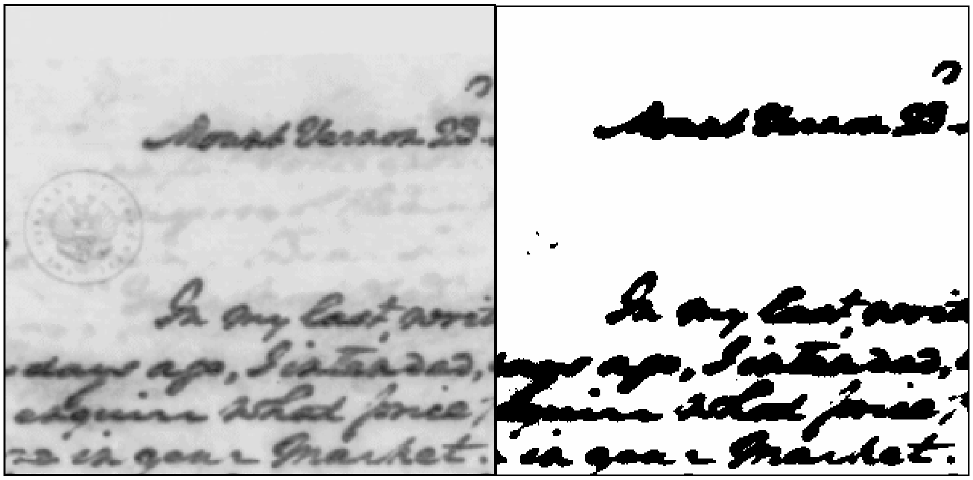


Figure 3.11 (a) Original Historical Image 'Image_A' (b) QIR Binarization Result

The original handwritten image on Figure 3.11(a) has many obvious occurrences of double-sided inking. Because of the long history of this particular document, ink has penetrated from the other side of the paper. This is a quite common noise effect in historical handwritten documents. The histogram of the original image is illustrated in Figure 3.12.

The histogram of 'Image_A' contains no obvious bimodal peak. This means that the object pixels are easily confused with the noise pixels. In this kind of image, the original QIR cannot clearly locate the boundary between the foreground and the fuzzy area. The binarization result is shown in Figure 3.11(b). The incorrectly chosen fuzzy area will affect the definition of the final threshold value. QIR fails in single-modal histogram images in many cases. The improved QIR thresholding algorithm can solve these problems.

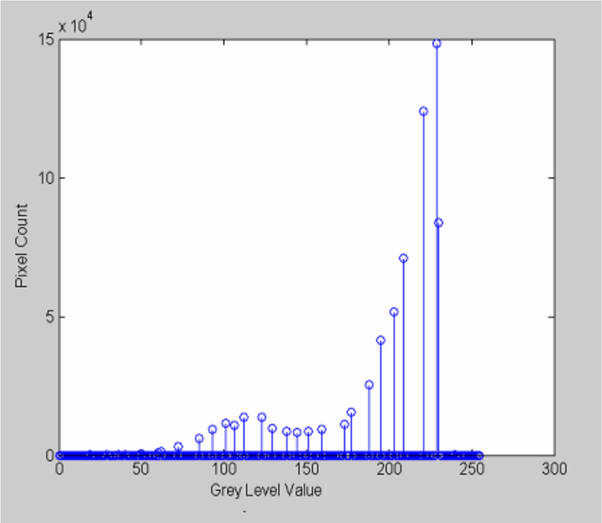


Figure 3.12 Histogram of the Original image in Figure 3.11(a) ‘Image_A’

3.3.2.1 Enhancing the Original Image

In order to produce obvious bimodal histogram, an enhancement pre-processing is applied to original image. The histogram of the enhanced original image is shown in Figure 3.13. Compared to Figure 3.12, Figure 3.13 has a more obvious bi-model histogram (peaks at greyscale values of approximately 90 and 230). This helps QIR to find a better cutting point in the first stage. The enhanced ‘Image_A’ is showed in Figure 3.14. It shows that both the useful and noise information in the original image are enhanced.

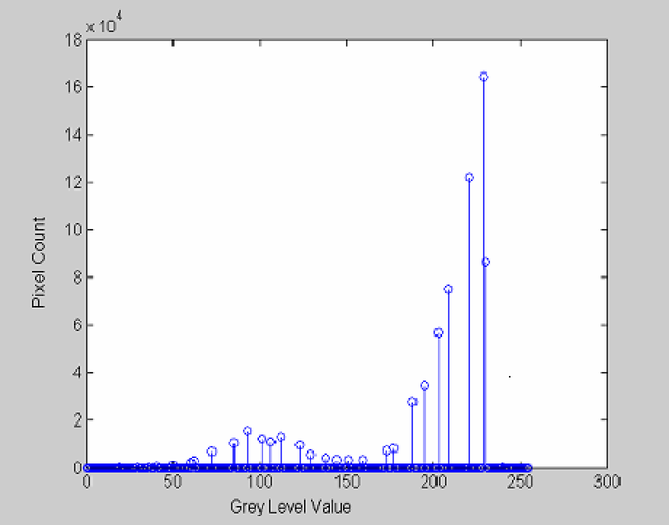


Figure 3.13 Histogram of The Enhanced Image

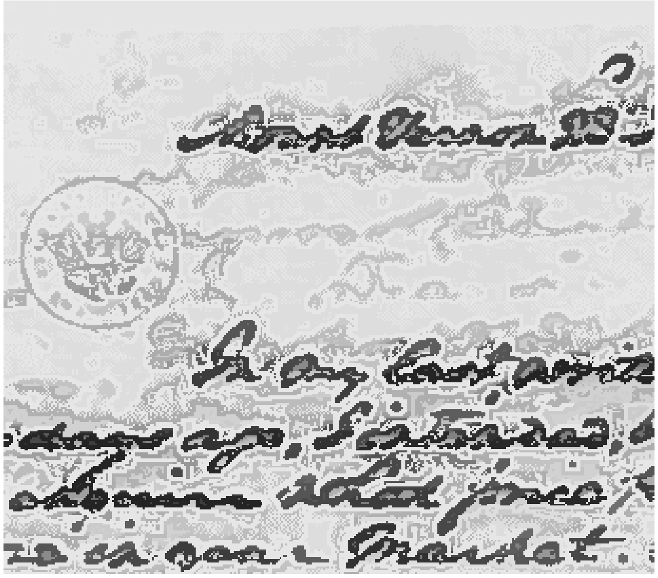
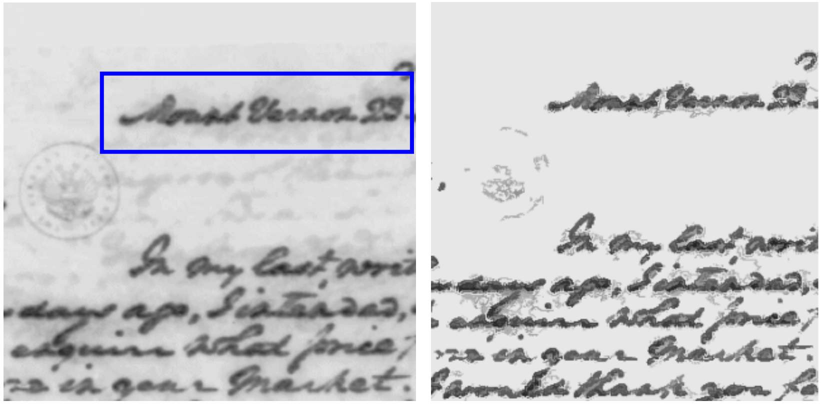


Figure 3.14 Enhanced 'Image_A'

3.3.2.2 Noise Removal using Integral Ratio (IR) Class

The Integral Ratio (IR) Class can be used to find background area from image histogram so that the pixels in this area can be set to grey level 220. This removes the noise from the background. The original and processed 'Image_A' are illustrated in

Figure 3.15. From the processed image in Figure 3.15, it can be observed that most of the unwanted double-side noise was removed.



(a) (b)

Figure 3.15 (a) Original 'Image_A'; (b) Enhanced Image (Right)

3.3.2.3 Using Integral Ratio (IR) Class in the second step of the QIR Algorithm

The comparison of results from the original QIR and the Improved QIR (with and without pre-processing) are shown in Figure 3.16. Improved QIR (Figure 3.16(c)) compared to Original QIR (Figure 3.16(a)) can keep more descenders' strokes and loops between strokes; however over-thresholding happens in some faint strokes location (Figure 3.16(c)). In order to further demonstrate the Improved QIR Algorithm compared with the original QIR Algorithm, the pre-processing method is used for both original QIR and Improved QIR as the first step.

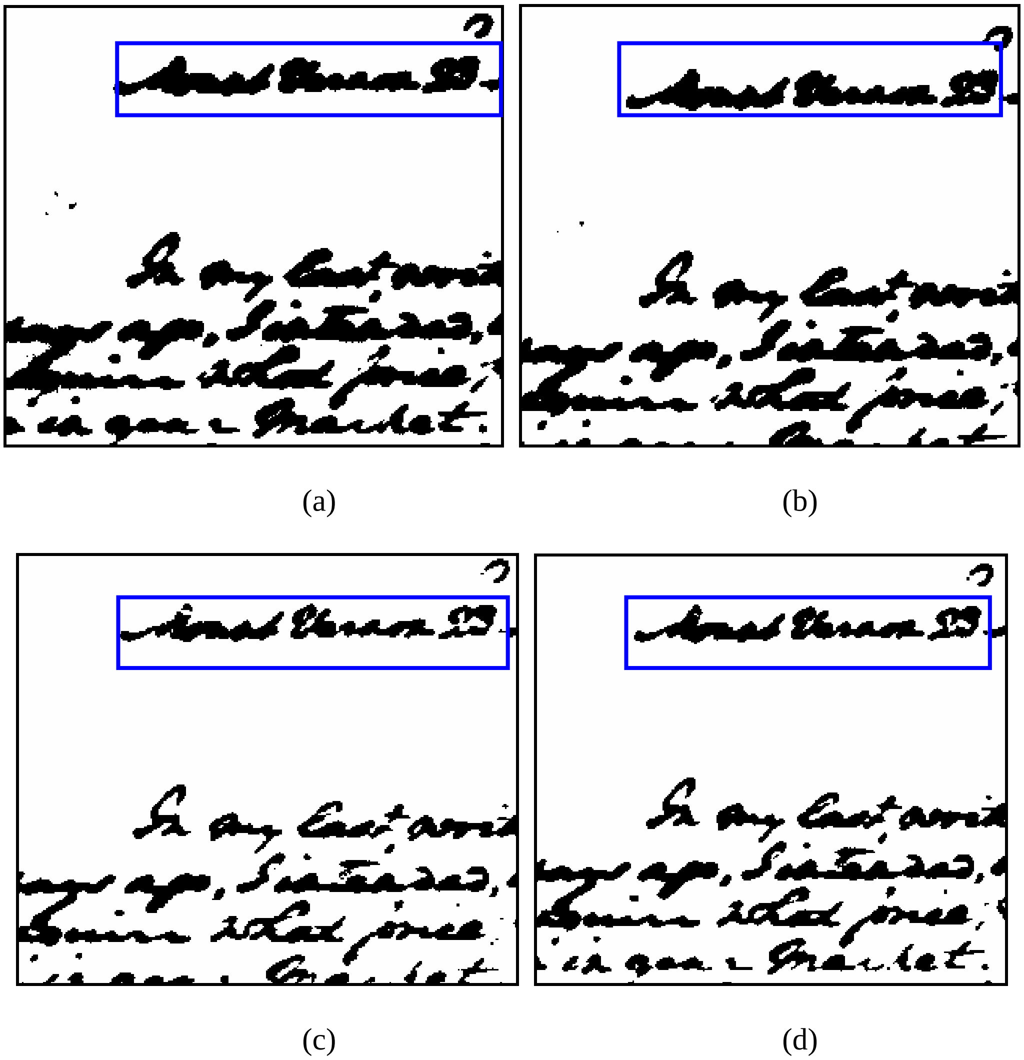
It is apparent that the Improved QIR with pre-processing (Figure 3.16 (d)) can retain more stroke information and reduce noise around handwritten characters. This will improve later processing of the handwriting.

Figure 3.17(a) shows the enlarged region of Figure 3.15(a). There is some double-side noise, which is hard to remove; faint strokes and faint word holes are also very difficult to retain.

Let’s compare how the original QIR and Improved QIR work on this section of the image. The original QIR retains most of the useful information as shown in Figure 3.16(a) but some detailed information such as the holes in loops has been lost and many strokes are connected. These make the binary result in Figure 3.17(a) quite hard to recognize. Figure 3.18(b) shows the binary result of original QIR with the proposed pre-processing method. The result shows that there are still some strokes connected together, which are not easy to recognize. The binary result of the Improved QIR is shown in Figure 3.17(c). The Improved QIR retains weak stroke and holes in loops that make the result much more clearly because of the fuzzy area classification. Figure 3.17(d) shows the binary result of Improved QIR with pre-processing method. As shown, the pre-processing method can avoid over-thresholding on faint strokes.

Figure 3.18 shows another comparison of the original QIR and the improved QIR on historical images. In this example, the original QIR provides a too high final threshold so that many faint strokes are broken. For the improved QIR algorithm, it keeps more useful information such as faint strokes and holes between connected strokes.

The Original QIR and Improved QIR were applied to 15 historical handwritten images. Two of them are shown in the Appendix 1 and others are shown in an attached CD. The images are aged and poor quality grey-scale image, which are degraded making the task difficult. The experimental results show that the Improved QIR has better performance than the Original QIR in correctly keeping thin strokes.



(a) (b)

(c) (d)

Figure 3.16 (a) Original QIR Binarization (b) Qriginal QIR with Pre-processing (c) Improved QIR (d) Improved QIR with Pre-processing

3.4 Summary

In summary, for improving the binarization result, three steps were added in the Improved QIR global thresholding algorithm:

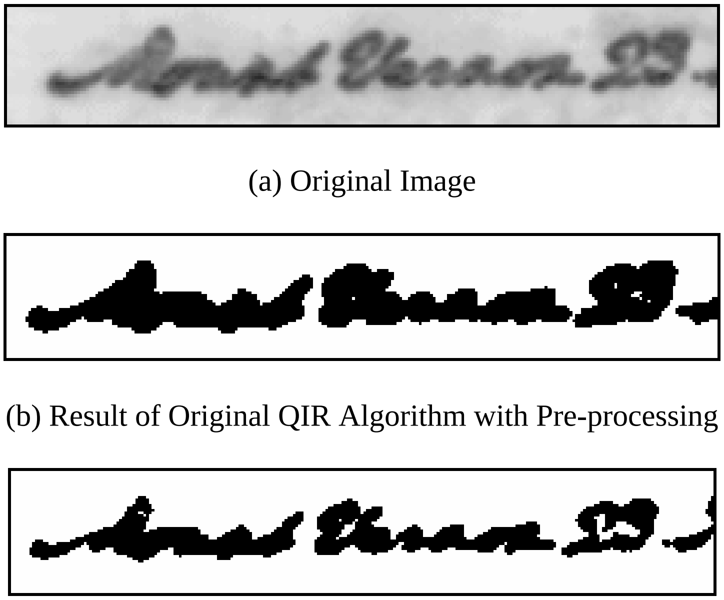
1. Enhance the image before binarization;
2. Double-side noise removal by using the IR Class;

3. Use the IR Class in the second step of the QIR algorithm.

The Original QIR and Improved QIR were applied to 15 historical handwritten images. Two of them are shown in the Appendix 1 and others are shown in an attached CD. The images are aged and poor quality grey-scale image, which are degraded making the task difficult. The experimental results show that the Improved QIR has better performance than the Original QIR in correctly keeping thin strokes.

Although QIR and the Improved QIR work better than other global thresholding algorithms [83], they cannot work well in some very hard low-contrast and noisy handwritten images, and global thresholding algorithm cannot correctly extract most of local information.

In order to extract the most useful information in a hard degraded historical image, a local based mean-gradient thresholding algorithm is considered in the next chapter.

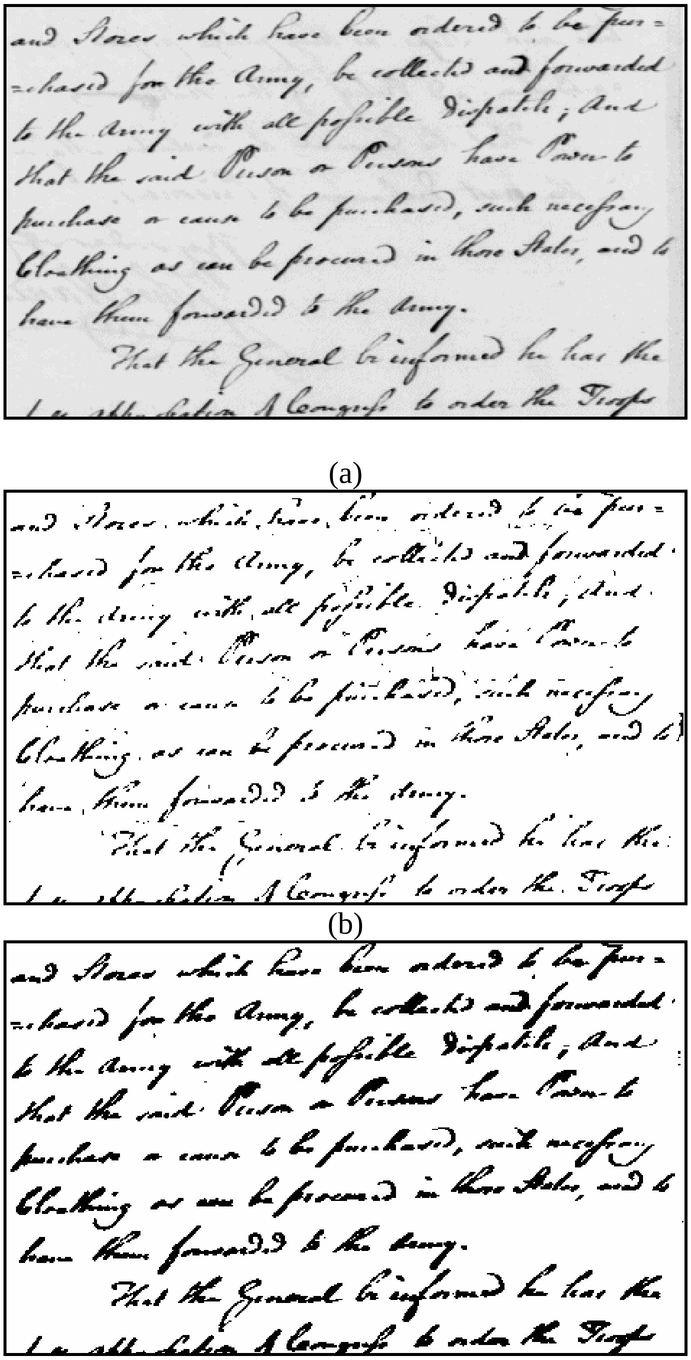


(a) Original Image

(b) Result of Original QIR Algorithm with Pre-processing

(c) Result of Improved QIR Algorithm with Pre-processing

Figure 3.17 Enlarged Region of Figure 3.16



and Stores which have been ordered to be pur=
=chased for the Army, be collected and forwarded
to the Army with all poſsible diſpatch; and
that the said Person or Persons have Power to
purchase or cause to be purchased, such neceſsary
Cloathing as can be procured in those States, and to
have them forwarded to the Army.
That the General be informed he has the
of Congreſs to order the Troops

(a)

(b)

(c)

Figure 3.18 (a) Original Historical Image_B; (b) Effect of Original QIR; (c) Effect of Improved QIR

Chapter 4

Mean-Gradient Thresholding Technique

Converting a scanned grey scale image into a binary image while retaining the foreground (or regions of interest) and removing the background is an important step in many image analysis systems, including Document Management Systems.

As discussed in Chapters 1 and 2, original documents are often dirty due to smearing and smudging of text and aging, and frequently, after scanning, exhibit a unimodal greyscale histogram meaning that global thresholding techniques will generally fail to produce satisfactory results.

In this chapter, a local mean-gradient global technique is proposed for four popular but different and difficult document image types: historical images, form images, cheque images and newspaper images to produce effective thresholding.

This chapter describes initial pre-processing, the proposed local adaptive Mean Gradient Technique, followed by experimental results and conclusion.

The mean-gradient thresholding methods described in this chapter was published in *Proc. 7th Int. Conf. on Document Analysis and Recognition*, Edinburgh, Scotland, Vol. 2, pp 859 -865, 2003.

4.1 Local Characteristic Analysis

A number of global and local thresholding techniques have been previously proposed as described in Chapter 2. All the reported thresholding methods have been proven effective in constrained pre-processing environments with predictable images. However, none of them has demonstrated to be effective in all cases of general document image processing.

In order to propose an effectively local adaptive threshold method, the analysis of local characteristics of grey-scale document image is necessary.

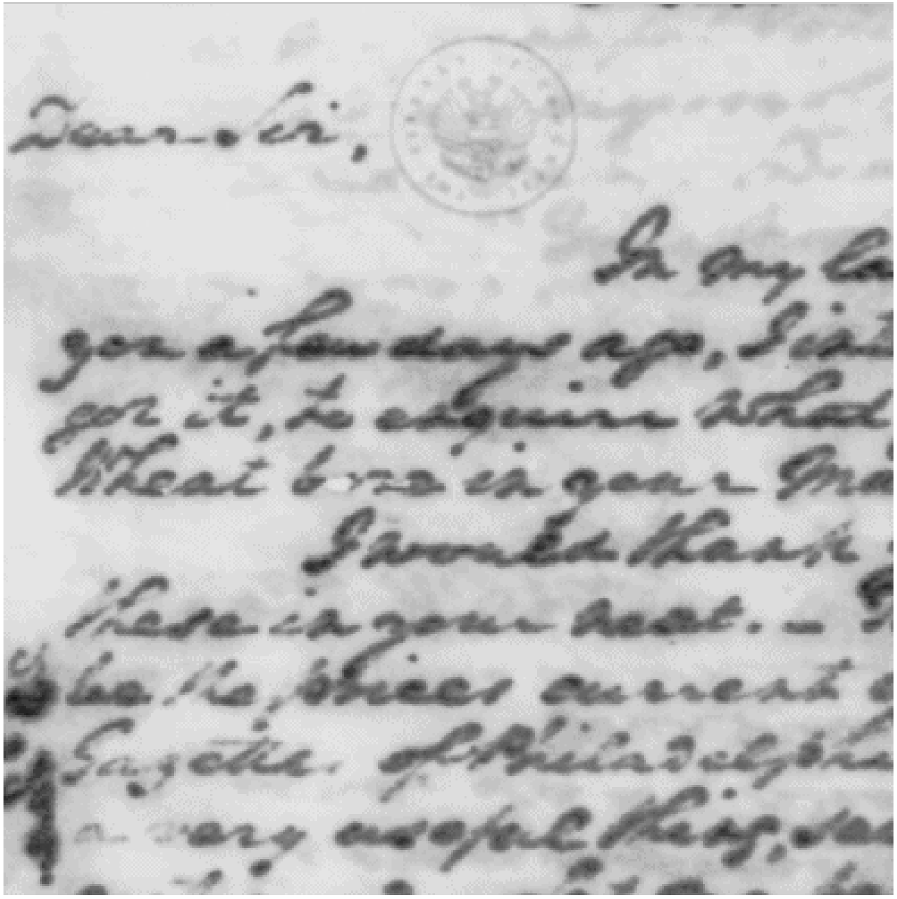
4.1.1 Grey-Scale Image Analysis

It was shown in Chapters 2 and 3 that the use of histogram-based global thresholding is limited in practice, because most documents have content producing non-obvious peaks in the histogram. Because of this, the local information is a very important feature to threshold a grey-scale image. This section focuses on the characteristics of a grey-scale image.

Figure 4.1 is a grey-scale historical handwritten image for analyzing the properties of a grey-scale scanned image. On Figure 4.1, it can be seen that the more information in a local area, the more variation in intensities. The variances can be used to represent the local information. The image shown in Figure 4.1 is divided into 16 blocks in Figure 4.2 in order to show the variation in each local area.

Variance is the main characteristic of a grey-scale image. The definitions of variance and standard deviation are described in the next section.

In order to analyze the local information of a grey-scale image, we firstly divide the image into some smaller blocks, and observe the variation of grey level for each block. These variations can be treated as one form of local information.



Dear Sir,
In my la
you a few days ago, I
got it, to enquire what
Wheat bore in your Ma
I would thank
there is your seat. —
the Three current
Gazette of Philadelphia
a very useful thing,

Figure 4.1 Grey-scale Scanned Image: ‘Image_B’

4.1.2 Definition of Variance and Standard Deviation

The variance is a measurement of how spread-out a distribution is. The spread of a variable is the degree by which scores on the variable differ from each other. It is computed as the average squared deviation of each number from its mean. The formula for the variance in a population is $\sigma^2 = \frac{\sum (X - \mu)^2}{N - 1}$ where μ is the mean and N is the number of scores.

- ❖ **Definition 7 –** Variance: If $\mu = E(X)$ is the expected value (mean) of the random variable *X*, then the variance is $\operatorname{var}(X) = E\left((X-\mu)^2\right)$.

It is the expected value of the square of the deviation of *X* from its own mean. In other words, it can be expressed as "The average of the square of the distance of each data point from the mean". The variance of random variable *X* is typically designated as $\operatorname{var}(X), \sigma_X^2$, or simply σ^2.

When the variance is computed in a sample, the statistic $S^2 = \frac{\sum (X-M)^2}{N-1}$ (where *M* is the mean of the sample) can be used. *S* is a biased estimate of σ ; however, $S^2 = \frac{\sum (X-M)^2}{N-1}$ is an unbiased estimate of σ^2.

Since samples are usually used to estimate parameters, S^2 is the most commonly used measurement of variance. The standard deviation is the square root of the variance. It is the most commonly used measurement of spread.

An important attribute of the standard deviation as a measure of spread is that if the mean and standard deviation of a normal distribution are known, it is possible to compute the percentile rank associated with any given score. Standard deviation can be used to describe the variation of a local area.

Figure 4.2 show that standard deviation describes the variation in each of the 16 image blocks shown in Figure 4.1. Analyzing the standard deviation value in each block, we can observe that there is only a little amplitude change of the standard deviation values between blocks. This shows that the standard deviation cannot obviously describe the difference between each block in detail, especially for those blocks which include heavy strokes. It's difficult to separate the blocks which include only heavy strokes from the blocks include both faint and heavy strokes. In this case, another parameter is needed to describe the strokes' characteristics.

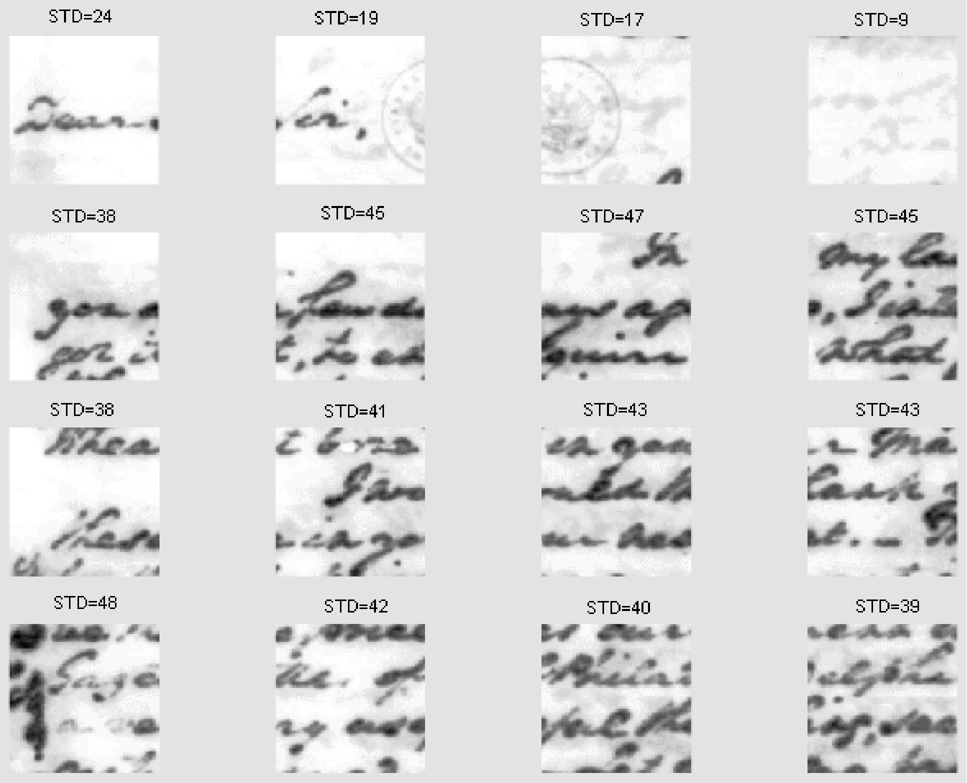


Figure 4.2 The Standard Deviation (STD) in Each Block

4.1.3 Definition of Mean-Gradient Value

- ❖ **Definition 8 - Texture Gradient** is the change of image intensity (measured or perceived) along some direction in the image, often corresponding to either a change in distance or surface orientation in the 3D world containing the objects creating the texture - Shapiro & Stockman (2001) [81].

From the definition above, the gradient of the intensity image *I(x,y)* is:

$$\nabla I(x,y) = \left[{\partial I(x,y)}\big/{\partial x}, {\partial I(x,y)}\big/{\partial y} \right] \qquad \text{Eq (4-1)}$$

The mean-gradient of the intensity image *I(x,y)* in directions *x* and *y* is:

$$G = \sum_{x=0}^{i-1} \sum_{y=0}^{j-1} \frac{\left[{\partial I(x,y)}\big/{\partial x}, {\partial I(x,y)}\big/{\partial y} \right]}{x \times y} \qquad \text{Eq (4-2)}$$

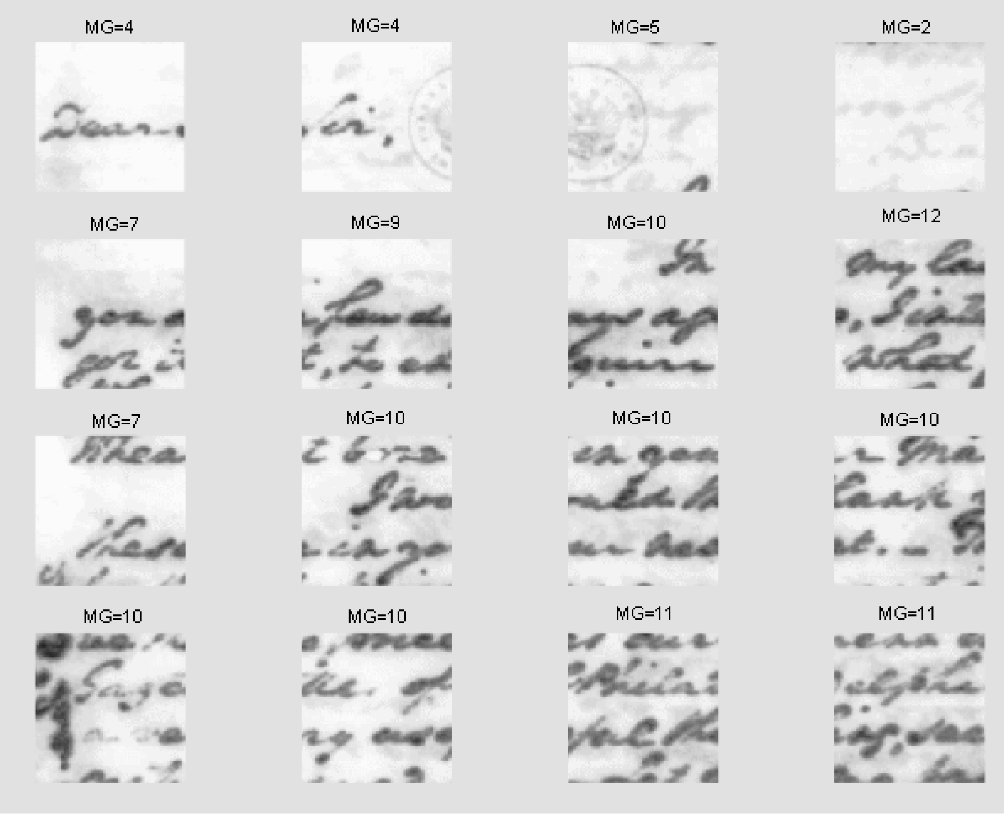


Figure 4.3 Mean-Gradient (MG) of Each Block

The mean-gradient value for each block of Figure 4.2 is produced using Eq (4-2) as shown in Figure 4.3. From Figure 4.3, it can be observed that the more characters (with heavy strokes) there are in a block, the higher the gradient value. Compared to standard deviation, the mean gradient is more sensitive to the variation between blocks, and one important property is that mean-gradient is sensitive to edges, so that it can correctly show the valuation of the strokes of characters. The more heavy stokes in the block, the higher gradient value. Comparing Figure 4.2 and Figure 4.3, it can be seen that the standard deviation is less sensitive to strokes.

For assessment the Mean-Gradient method was applied to 40 document images: 10 historical document images, 10 form document images, 10 cheque document images and 10 newspaper document images. The results of the analysis show that the property of the mean-gradient is sensitivity to edges and noise (All these images are shown in Appendix 2 and the attached CD). In order to remove the unwanted noise and patterns, a new background subtraction method for removing noise can be used.

4.2 Pre-processing Method

Scanned images may contain many unknown artifacts and unwanted patterns as shown in Figure 4.5. In order to remove the unwanted artifacts and patterns as much as possible, a background subtraction technique can be used during pre-processing.

This method consists of two steps. Firstly, the background of an image is modeled by removing the handwriting from the original image using a closing algorithm, as illustrated in Gonzalez & Woods (1993) [33], with a small disk as a structuring element. The closing algorithm when applied to a greyscale image where the characters of interest are darker than the background tends to remove these darker areas. Figure 4.4 shows an original grey-scale cheque ‘Image_C’. The background of Figure 4.4 is illustrated in Figure 4.5.

Secondly, the background is subtracted from the original document image leaving only the handwriting of interest. Figure 4.6 is the final result after subtracting Figure 4.5

from Figure 4.4. Figure 4.6 shows that most of the unwanted pattern and noise are removed. However, this method sometimes causes the removal of some wanted details, especially weak strokes in the document images.

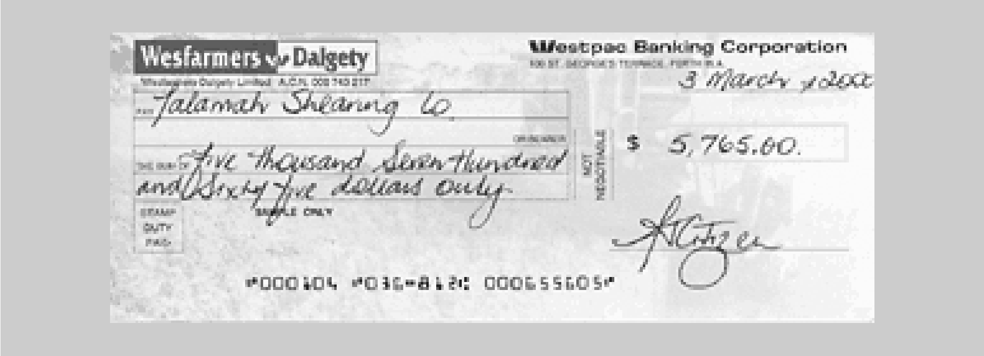


Figure 4.4 Original Grey-scale Cheque ‘Image_C’

Figure 4.5 Unwanted Artifacts and Patterns of ‘Image_C’

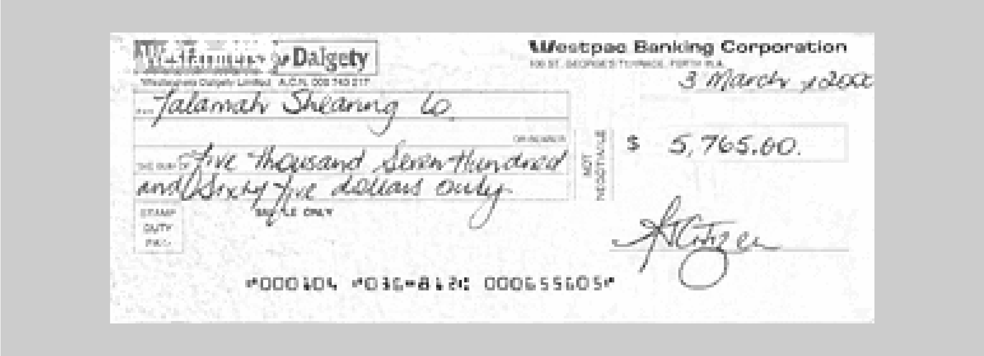

Dalgety
Westpac Banking Corporation
3 March 2000
Talamah Shearing Co.
$ 5,765.60.
Five thousand Seven Hundred
and Sixty five dollars only
NOT NEGOTIABLE
STAMP DUTY PAID

Figure 4.6 Noise and Pattern Background Removed ‘Image_C’

4.3 Proposed Local Adaptive Mean-Gradient Technique

After examining the local information of a grey-scale image and the noise removal methods, a new local adaptive thresholding technique is proposed.

The outline of the mean-gradient thresholding algorithm is:

1. Pre-processing: Window-based enhancement and Background Subtraction
2. Mean value and Mean-Gradient value calculation of each local area
3. Thresholding

4.3.1 Pre-processing

(1) Image Enhancing. In Chapter 3, a simple enhancement method was described. This method can also be used in our new mean-gradient algorithm to obtain a clear enhanced image.

(2) Background Subtraction. This method is effective for noise removal when the intensity on an image is not constant.

These two pre-processing steps can produce a clear, higher contrast grey-scale image for subsequent processing.

4.3.2 Local Adaptive Mean-Gradient Thresholding Technique

This Mean-Gradient thresholding technique can be regarded as a new variant of Niblack’s local thresholding method, Niblack (1986).

The mean-Gradient thresholding technique is based on the mean greyscale value and mean-gradient value. The threshold value for a block (local area) is calculated as:

$$T(x,y) = M(x,y) + k \times G(x,y) \qquad \text{Eq (4-3)}$$

where the parameters *M(x,y)* and *G(x,y)* are the local mean and local mean-gradient respectively calculated in a window centred at *(x,y)*, and *k* is a user defined negative weighted value. The formula for mean-gradient can be found as Eq (4-2) in Section 4.1.3.

Mean-Gradient is sensitive to noise, and as such, the technique can be improved by adding a pre-condition when selecting a threshold level: a constant *C*, which is the local contrast value equal to the maximum grey value minus the minimum grey value in the window. The weighted mean-gradient value can be combined with local mean value to detect stroke details from document images when the local contrast is very low ($C \le R$).

In summary:

If $C \le R$,

then $T(x,y) = M(x,y) + k \times G(x,y)$;

Else $T(x,y) = 0.5M(x,y)$, where k = -1.5, R = 40.

Parameters R = 40 and k = -1.5 are the values which can produce best thresholding result during the experimental testing of the proposed technique applied to four different types of images.

More experimental results are described in the following section.

4.4 Experimental Results

4.4.1 Qualitative Comparison of thresholding Methods

The mean-gradient thresholding method was applied to four types of images – historical document, newspapers, forms, and cheques. These were used to test the performance of the method because the images have varying resolutions, sizes, as well as contrast to ensure realistic comparison of performance of the method with other techniques.

The threshold results are separately presented in four groups, each containing ten images:

1. Historical handwritten Document Image
2. Form Document Image
3. Newspaper Document Image
4. Cheque Document Image

There are four published techniques that will be compared between these four groups of images:

1. Improved Niblack's Algorithm, Zhang & Tan (2001);
2. Original QIR's Algorithm, Solihin & Leedham (1999);
3. Yanowitz and Bruckstein's Algorithm, Yanowitz & Bruckstein (1989);
4. Mean-Gradient Algorithm, Leedham et al. (2003).

Techniques 1 to 3 are published global or local techniques which are outstanding techniques in thresholding research area. Figures 4.8 to 4.11 show the comparison of these four thresholding algorithms. The detected foreground area is restored as the greyscale value from the original images. The area determined as background is set to

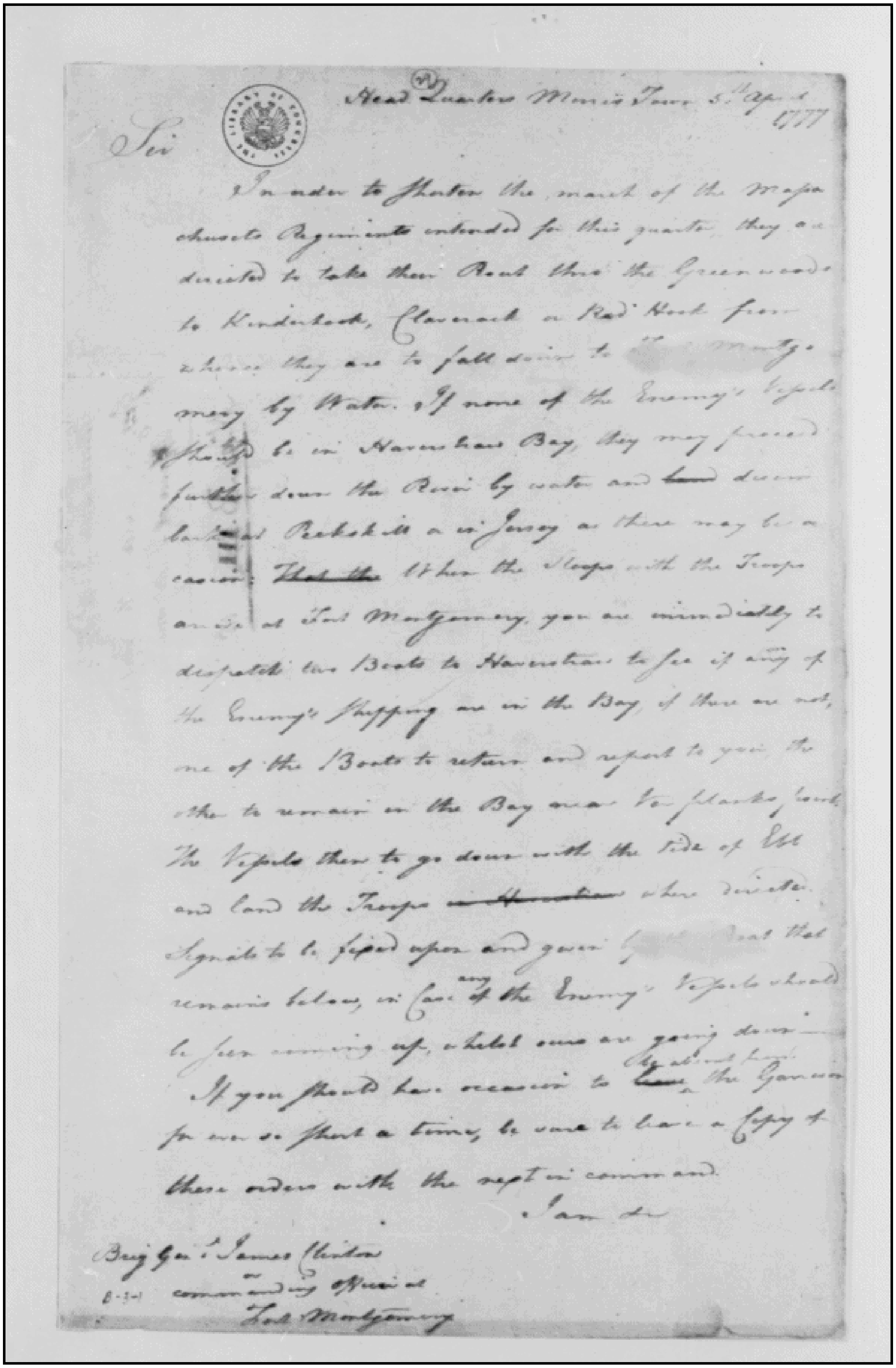
230 in all cases for better visual comparison. Each group of figures consists of: one original grey-scale image from that group and four greyscale restored images obtained using the above four algorithms.

4.4.1.1 Group 1: Historical Handwritten Document Image

As discussed in an earlier chapter, historical documents will frequently contain a significant amount of noise in the background due to the long storage time and handling. Nevertheless, some strokes were originally connected and some become connected due to spreading of the ink to form overlapped strokes. This makes the original handwritten word very difficult to recognize. Weak strokes and faint words will always affect the reading quality. To provide better visual comparison of how well the thresholding methods separate the useful foreground from the noisy background, all detected foregrounds are restored to the greyscale value of the original input images in the same position; the pixels classified as background are all set to 230 for good visual comparison. Figures 4.7(b) to (e) are the greyscale restored images to show how well the four techniques work on the historical handwritten image shown in Figure 4.7(a).

Figure 4.7(b) shows that Improved Niblack’s Algorithm cannot retain all weak strokes, and it filled some holes in words. The connected strokes make the reading quality worse.

In Figure 4.7(c), the QIR Algorithm did not successfully retain weak words, and filled the holes in words. It performed well for faint words. As with the Improved Niblack’s Algorithm, it was not able to correctly segment the overlapped words.



Head Quarters Morris Town 5th April
1777

Sir

In order to shorten the march of the Massa-
chusetts Regiments intended for this quarter, they are
directed to take their Rout thro' the Green woods
to Kinderhook, Claverack or Red Hook from
whence they are to fall down to [illegible] Montgo-
mery by Water. If none of the Enemy's Vessels
should be in Haverstraw Bay, they may proceed
further down the River by water and ~~land~~ down
landed at Peekskill or in Jersey as there may be oc-
casion. ~~That the~~ When the Sloops with the Troops
arrive at Fort Montgomery, you are immediately to
dispatch two Boats to Haverstraw to see if any of
the Enemy's Shipping are in the Bay, if there are not,
one of the Boats to return and report to you, the
other to remain in the Bay near Verplanks point.
The Vessels then to go down with the tide of Ebb
and land the Troops ~~at Haverstraw~~ where directed.
Signals to be fixed upon and given [illegible] boat that
remains below, in case any of the Enemy's Vessels should
be seen coming up, whilst ours are going down —

If you should have occasion to ~~leave~~ be absent from the Garrison
for ever so short a time, be sure to leave a Copy of
these orders with the next in command.

I am &c

Brig: Genl. James Clinton
b-3-1 commanding officer at
Fort Montgomery

Figure 4.7(a) Historical Image

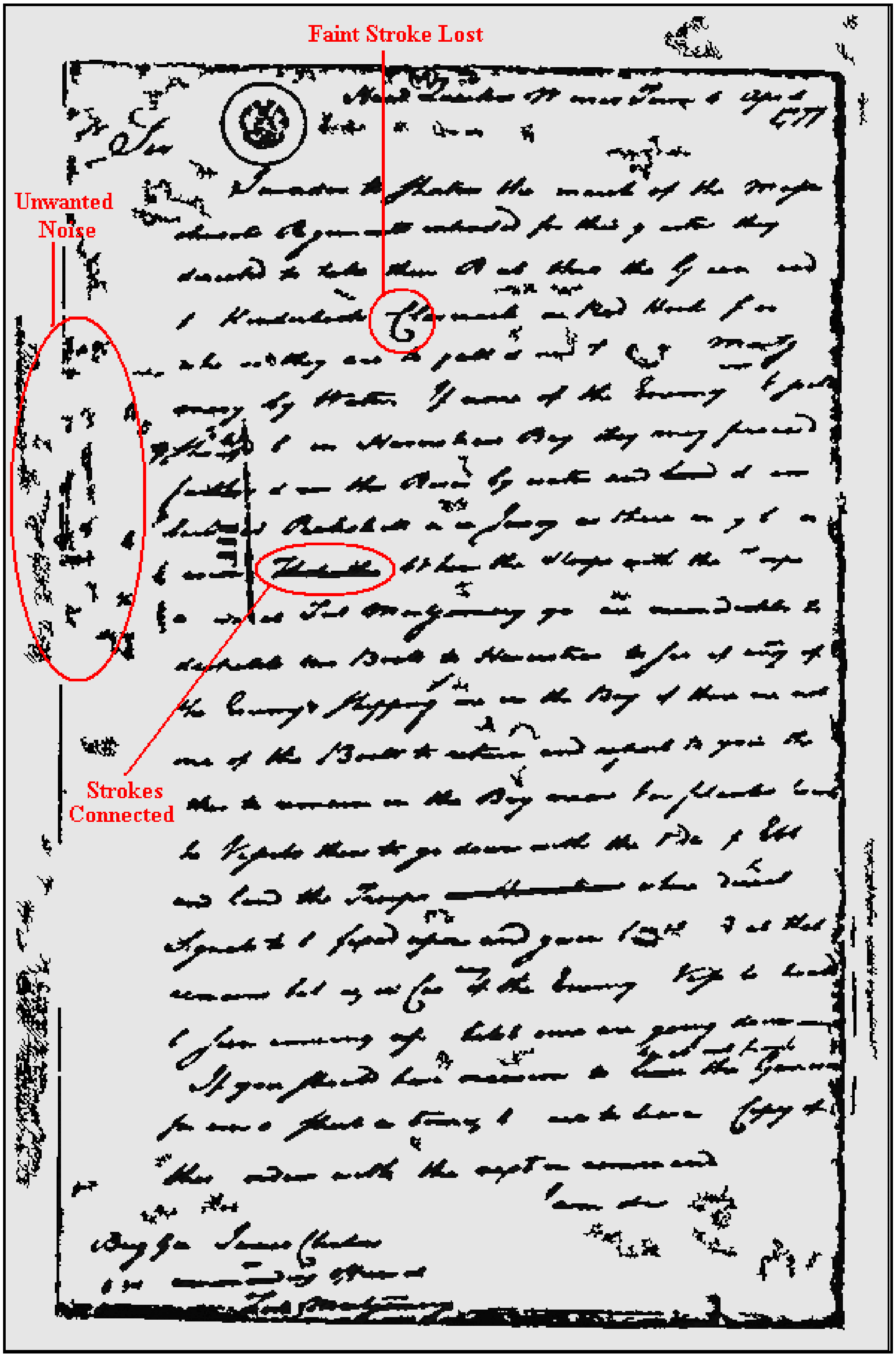

Faint Stroke Lost
Unwanted Noise
Strokes Connected

Figure 4.7 (b) Effect of the Improved Niblack Algorithm

on the image shown in Figure 4.7(a)

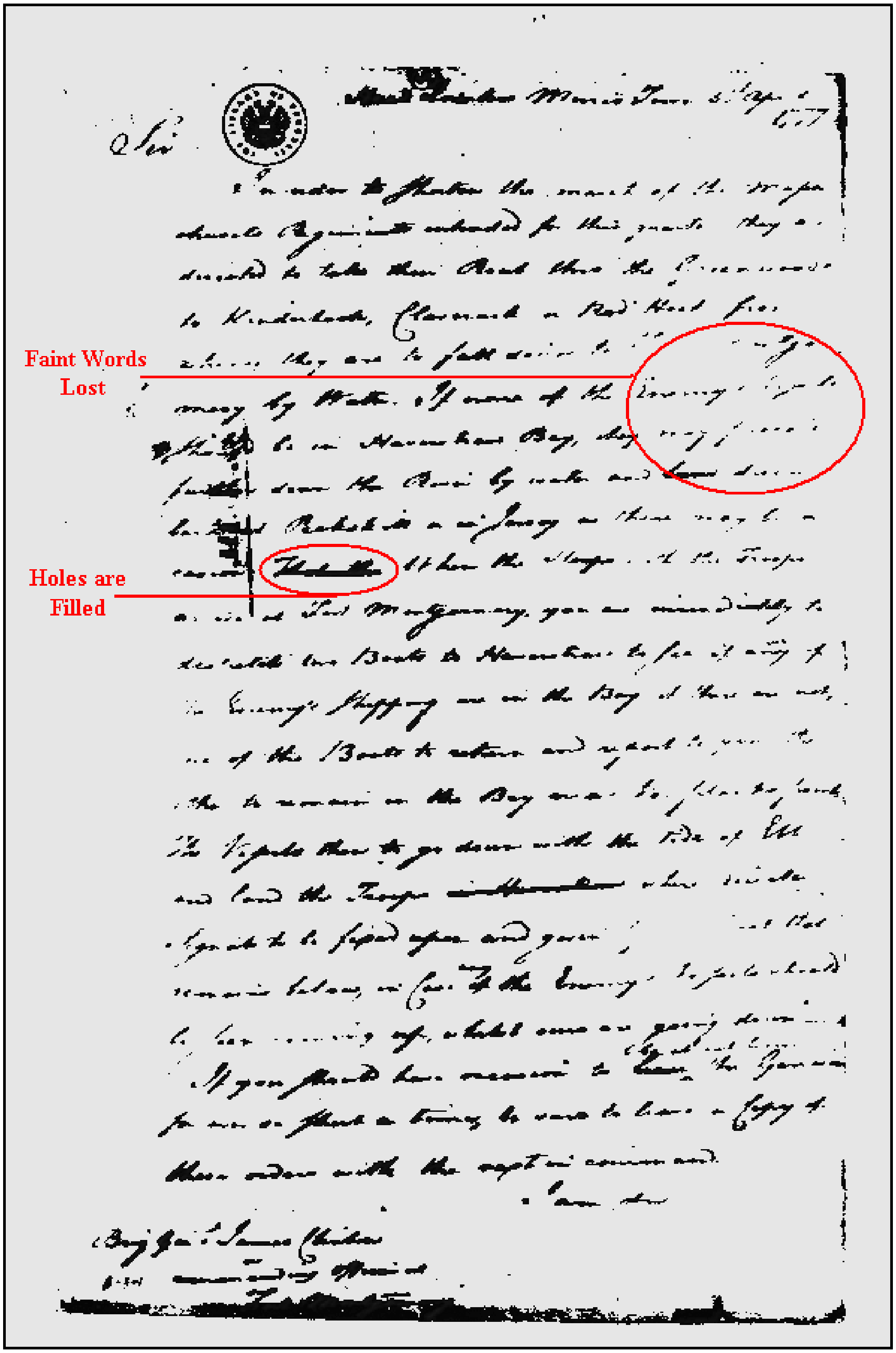


Figure 4.7 (c) Effect of the Original QIR Algorithm

on the image shown in Figure 4.7(a)

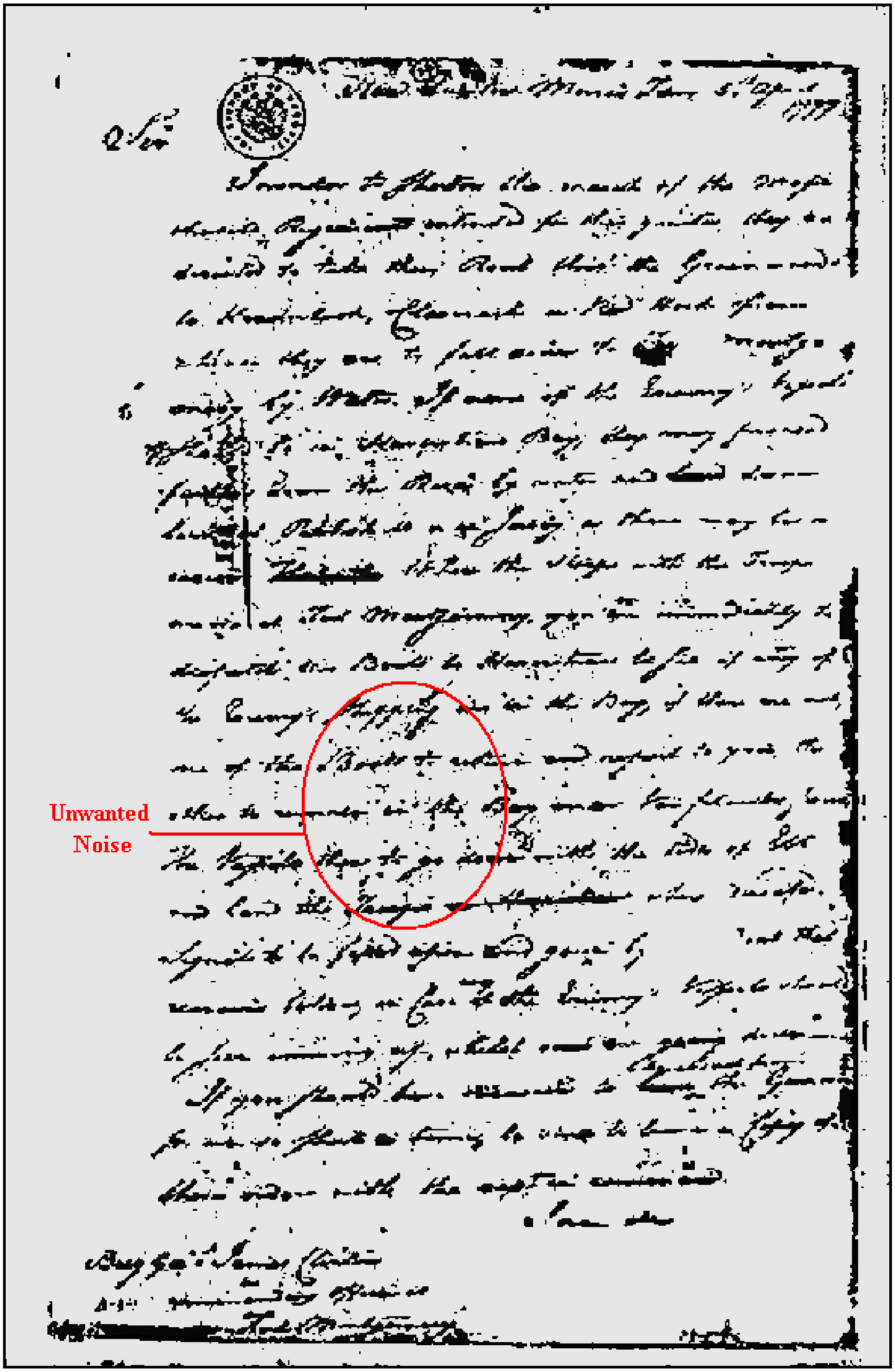


Figure 4.7 (d) Effect of the Yanowitz and Bruckstein's Algorithm on the image shown in Figure 4.7(a)

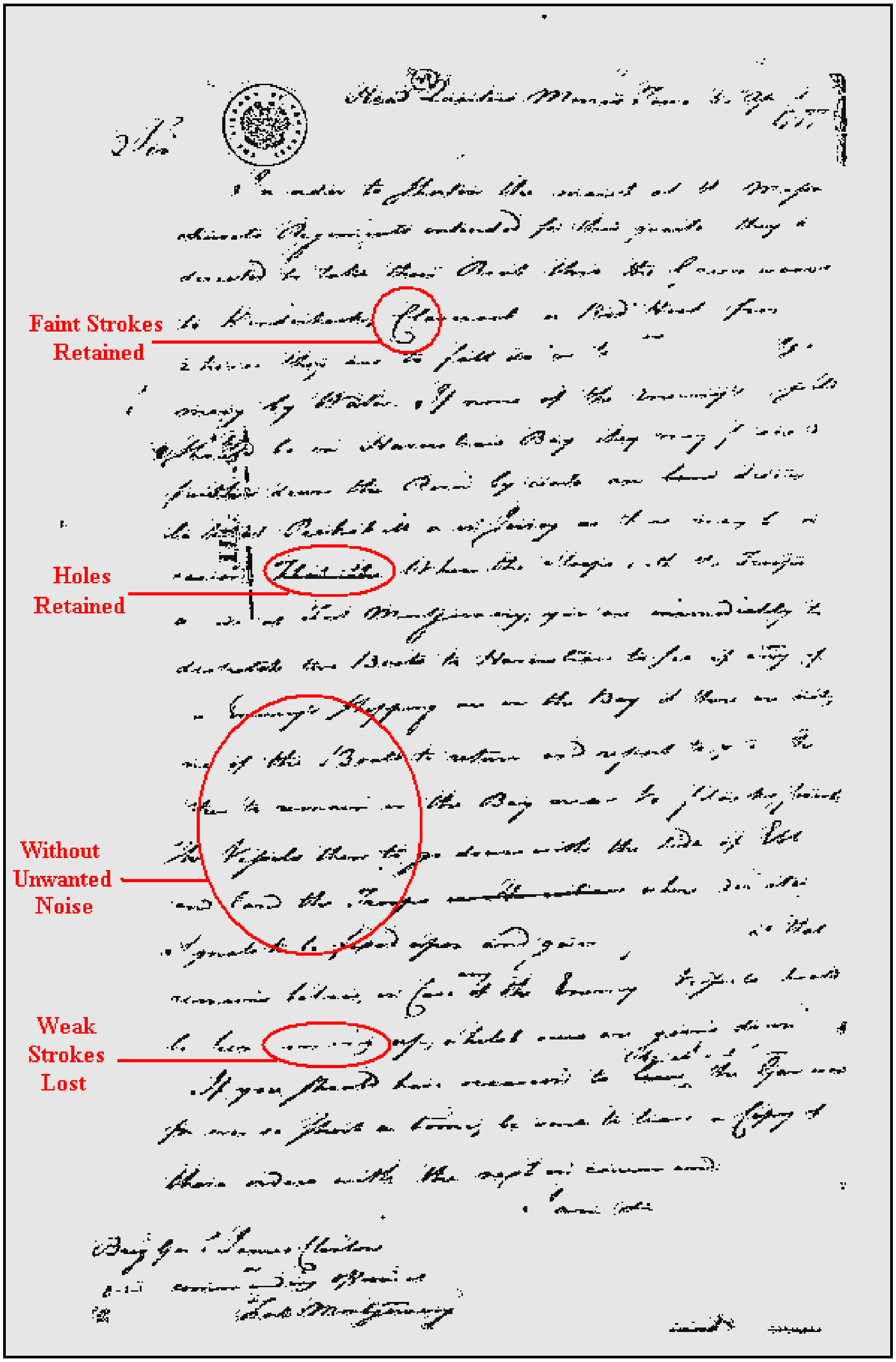


Figure 4.7 (e) Effect of the Mean-Gradient Algorithm

on the image shown in Figure 4.7(a)

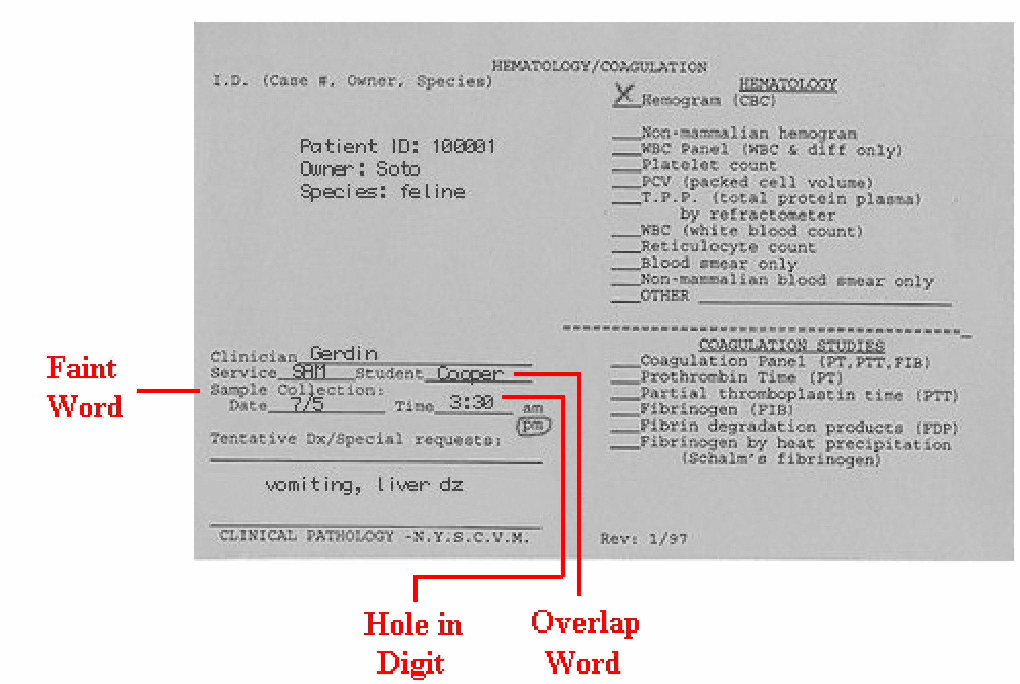
Yanowitz and Bruckstein's Algorithm removed some text along with the unwanted noise (See Figure 4.7 (d)). This affected the reading quality. Yanowitz and Bruckstein's Algorithm retained holes in words, and faint words. However, it is not possible to detect the overlapped strokes without noise. The Mean-Gradient technique (Figure 4.7 (c)) retained the holes in words and faint words, and correctly detected the overlapped strokes with less noise. However, it did not retain the weak connected strokes in some words.

4.4.1.2 Group 2: Form Document Image

The challenge of separating foreground from the form images is exacerbated by lots of short lines underneath and overlapped with the printed or handwritten words. This makes words difficult to clearly detect.

Figure 4.8(a) shows a printed form document image with typed words. The printed text appears faint, and the typewritten words overlap with lines underneath them. Figures 4.9(b) to (e) show the restored foreground greyscale images using the four different thresholding methods.

After applying the improved Niblack's algorithm (See Figure 4.8(b)), the faint strokes were connected together so that is difficult to read the words. The Improved Niblack's Algorithm failed to clearly detect the words with overlapping lines, and faint strokes are broken in some words.


Faint
Word
Hole in
Digit
Overlap
Word

HEMATOLOGY/COAGULATION

I.D. (Case #, Owner, Species)

Patient ID: 100001
Owner: Soto
Species: feline

HEMATOLOGY

☒ Hemogram (CBC)

___Non-mammalian hemogram
___WBC Panel (WBC & diff only)
___Platelet count
___PCV (packed cell volume)
___T.P.P. (total protein plasma) by refractometer
___WBC (white blood count)
___Reticulocyte count
___Blood smear only
___Non-mammalian blood smear only
___OTHER ___

Clinician Gerdin
Service SAM Student Cooper
Sample Collection:
Date 7/5 Time 3:30 am [✓ pm]

Tentative Dx/Special requests:

vomiting, liver dz

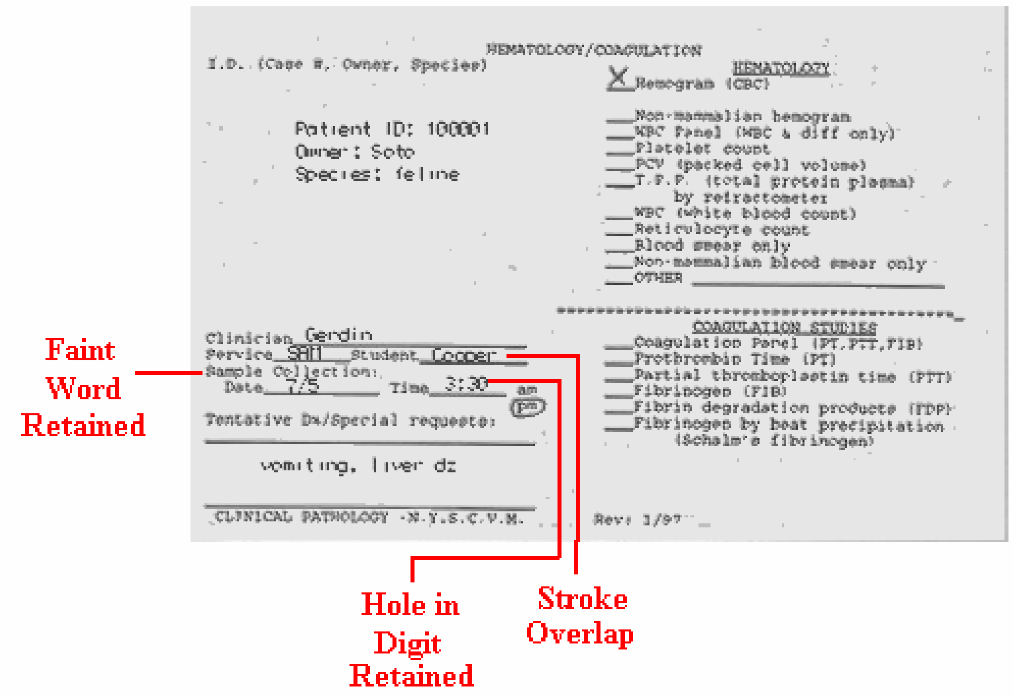
COAGULATION STUDIES

___Coagulation Panel (PT,PTT,FIB)
___Prothrombin Time (PT)
___Partial thromboplastin time (PTT)
___Fibrinogen (FIB)
___Fibrin degradation products (FDP)
___Fibrinogen by heat precipitation (Schalm's fibrinogen)

CLINICAL PATHOLOGY -N.Y.S.C.V.M.

Rev: 1/97

Figure 4.8 (a) Original Form Image


Faint
Word
Retained
Hole in
Digit
Retained
Stroke
Overlap

HEMATOLOGY/COAGULATION

I.D. (Case #, Owner, Species)

Patient ID: 100001
Owner: Soto
Species: feline

HEMATOLOGY

☒ Hemogram (CBC)

___Non-mammalian hemogram
___WBC Panel (WBC & diff only)
___Platelet count
___PCV (packed cell volume)
___T.P.P. (total protein plasma) by refractometer
___WBC (white blood count)
___Reticulocyte count
___Blood smear only
___Non-mammalian blood smear only
___OTHER ___

Clinician Gerdin
Service SAM Student Cooper
Sample Collection:
Date 7/5 Time 3:30 am [✓ pm]

Tentative Dx/Special requests:

vomiting, liver dz

COAGULATION STUDIES

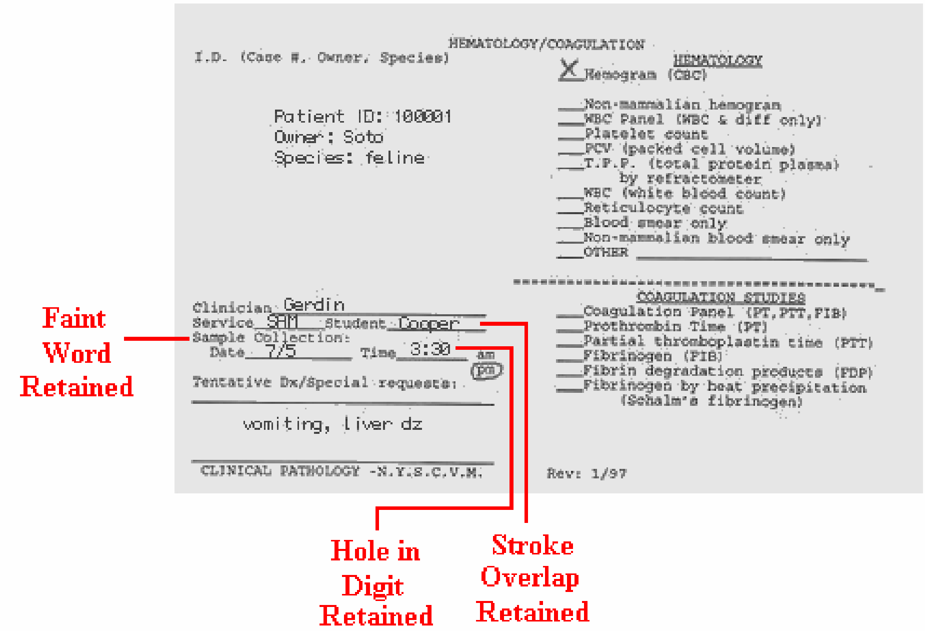
___Coagulation Panel (PT,PTT,FIB)
___Prothrombin Time (PT)
___Partial thromboplastin time (PTT)
___Fibrinogen (FIB)
___Fibrin degradation products (FDP)
___Fibrinogen by heat precipitation (Schalm's fibrinogen)

CLINICAL PATHOLOGY -N.Y.S.C.V.M.

Rev: 1/97

Figure 4.8 (b) Effect of the Improved Niblack's Algorithm

on the image shown in Figure 4.8(a)



HEMATOLOGY/COAGULATION

I.D. (Case #, Owner, Species)

Patient ID: 100001
Owner: Soto
Species: feline

HEMATOLOGY

X Hemogram (CBC)

- __Non-mammalian hemogram
- __WBC Panel (WBC & diff only)
- __Platelet count
- __PCV (packed cell volume)
- __T.P.P. (total protein plasma) by refractometer
- __WBC (white blood count)
- __Reticulocyte count
- __Blood smear only
- __Non-mammalian blood smear only
- __OTHER ______

COAGULATION STUDIES

- __Coagulation Panel (PT,PTT,FIB)
- __Prothrombin Time (PT)
- __Partial thromboplastin time (PTT)
- __Fibrinogen (FIB)
- __Fibrin degradation products (FDP)
- __Fibrinogen by heat precipitation (Schalm's fibrinogen)

Clinician Gerdin
Service SAM Student Cooper
Sample Collection:
Date 7/5 Time 3:30 am [✓ pm]
Tentative Dx/Special requests:

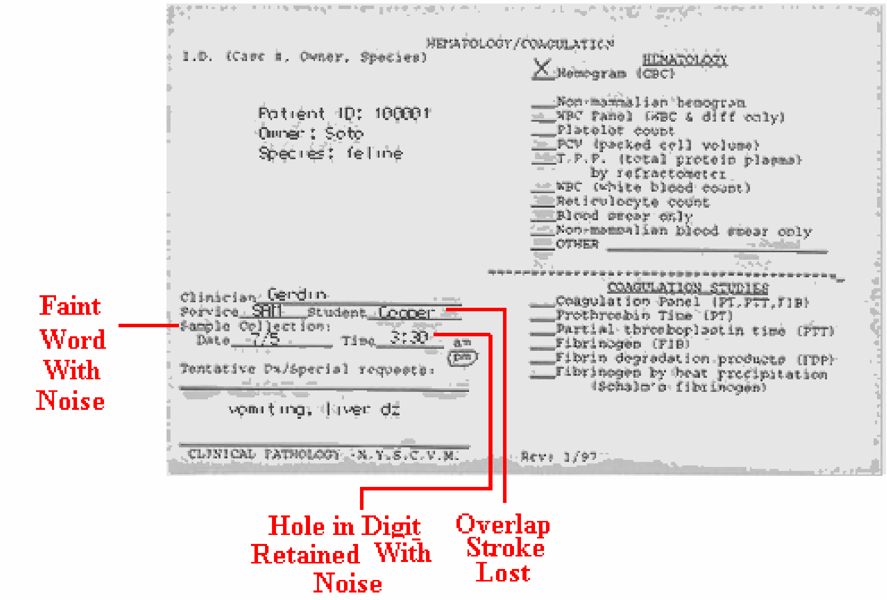
vomiting, liver dz

CLINICAL PATHOLOGY -N.Y.S.C.V.M.

Rev: 1/97

Figure 4.8 (c) Effect of the Original QIR Algorithm

on the image shown in Figure 4.8(a)



HEMATOLOGY/COAGULATION

I.D. (Case #, Owner, Species)

Patient ID: 100001
Owner: Soto
Species: feline

HEMATOLOGY

X Hemogram (CBC)

- __Non-mammalian hemogram
- __WBC Panel (WBC & diff only)
- __Platelet count
- __PCV (packed cell volume)
- __T.P.P. (total protein plasma) by refractometer
- __WBC (white blood count)
- __Reticulocyte count
- __Blood smear only
- __Non-mammalian blood smear only
- __OTHER ______

COAGULATION STUDIES

- __Coagulation Panel (PT,PTT,FIB)
- __Prothrombin Time (PT)
- __Partial thromboplastin time (PTT)
- __Fibrinogen (FIB)
- __Fibrin degradation products (FDP)
- __Fibrinogen by heat precipitation (Schalm's fibrinogen)

Clinician Gerdin
Service SAM Student Cooper
Sample Collection:
Date 7/5 Time 3:30 am [✓ pm]
Tentative Dx/Special requests:

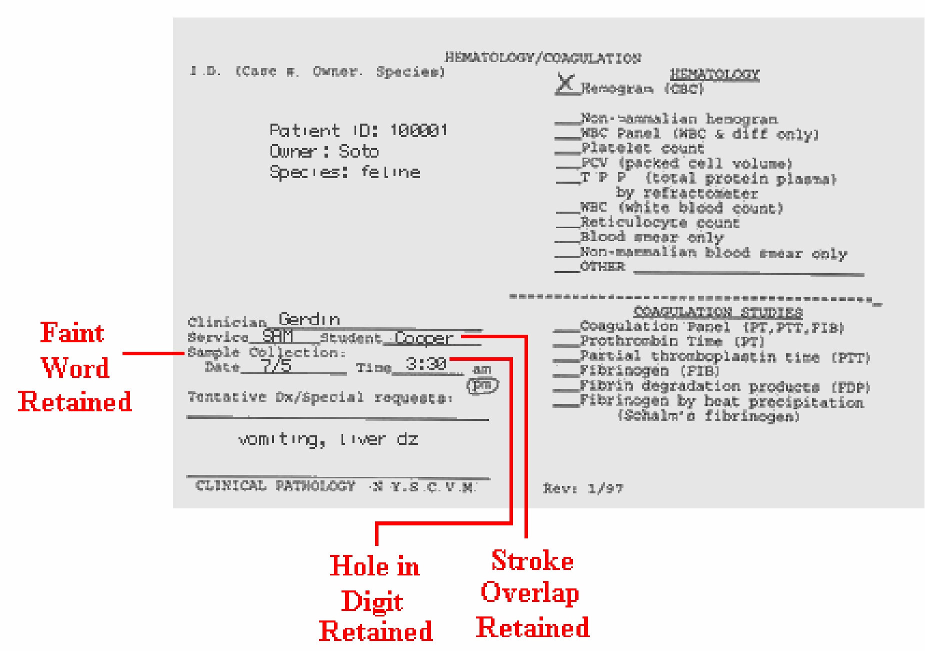
vomiting, liver dz

CLINICAL PATHOLOGY -N.Y.S.C.V.M.

Rev: 1/97

Figure 4.8 (d) Effect of the Yanowitz's Algorithm

on the image shown in Figure 4.8(a)



HEMATOLOGY/COAGULATION

I.D. (Case #. Owner. Species)

Patient ID: 100001
Owner: Soto
Species: feline

HEMATOLOGY

X Hemogram (CBC)

___Non-mammalian hemogram
___WBC Panel (WBC & diff only)
___Platelet count
___PCV (packed cell volume)
___T P P (total protein plasma) by refractometer
___WBC (white blood count)
___Reticulocyte count
___Blood smear only
___Non-mammalian blood smear only
___OTHER ____________

Clinician Gerdin
Service SAM Student Cooper
Sample Collection:
Date 7/5 Time 3:30 am [✓ pm]

Tentative Dx/Special requests:

vomiting, liver dz

CLINICAL PATHOLOGY N.Y.S.C.V.M.

COAGULATION STUDIES

___Coagulation Panel (PT,PTT,FIB)
___Prothrombin Time (PT)
___Partial thromboplastin time (PTT)
___Fibrinogen (FIB)
___Fibrin degradation products (FDP)
___Fibrinogen by heat precipitation (Schalm's fibrinogen)

Rev: 1/97

Figure 4.8 (e) Effect of the Mean-Gradient Algorithm

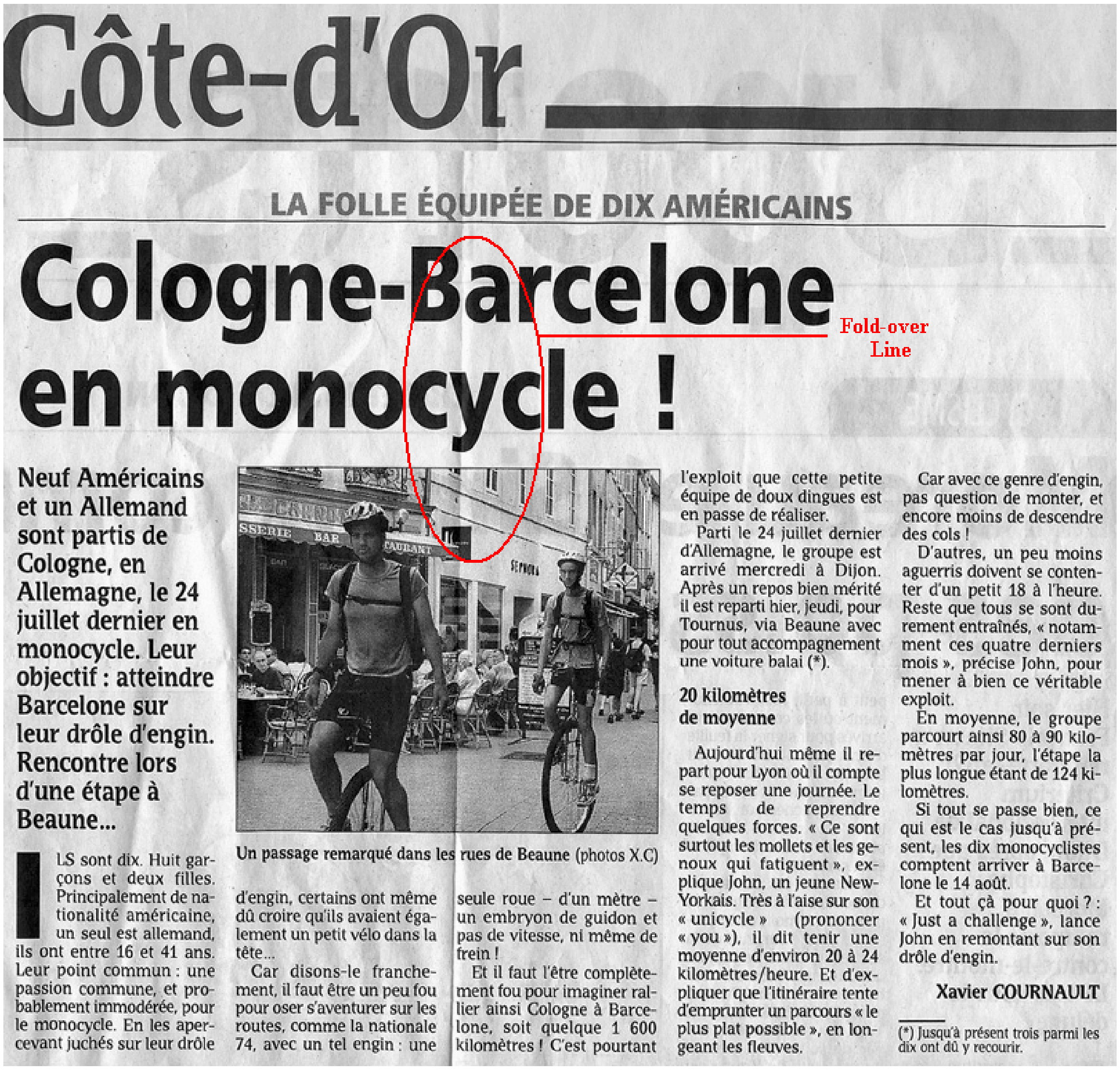
on the image shown in Figure 4.8(a)

Figure 4.8(c) shows the result using the QIR algorithm. Some strokes of the faint words overlapped. For the overlapping words with lines, the QIR Algorithm correctly detected parts of strokes. Compared with Figure 4.8(b), Figure 4.8(c) contains more useful information and removed unwanted impulse noise. In Figure 4.8(d), it can be seen that Yanowitz's Algorithm also retained faint words but did not clearly detect the words, which were overlapped by lines.

The Mean-Gradient technique, Figure 4.8(e) clearly retained the faint words' strokes, and correctly detected the strokes, which connected with form lines. Although Figure 4.8(c) still shows some broken strokes so that some words cannot be recognized, the qualitative impression is that the Mean-Gradient technique works better than the other three techniques.

4.4.1.3 Group 3: Newspaper Document Image

A newspaper is usually folded, so that fold-over lines are frequently observed in the images of scanned newspaper. Let us examine the results that the techniques produced for this case as illustrated in Figure 4.9.


Fold-over
Line

Côte-d'Or

LA FOLLE ÉQUIPÉE DE DIX AMÉRICAINS

Cologne-Barcelone en monocycle !

Neuf Américains et un Allemand sont partis de Cologne, en Allemagne, le 24 juillet dernier en monocycle. Leur objectif : atteindre Barcelone sur leur drôle d'engin. Rencontre lors d'une étape à Beaune...

Un passage remarqué dans les rues de Beaune (photos X.C)

ILS sont dix. Huit garçons et deux filles. Principalement de nationalité américaine, un seul est allemand, ils ont entre 16 et 41 ans. Leur point commun : une passion commune, et probablement immodérée, pour le monocycle. En les apercevant juchés sur leur drôle d'engin, certains ont même dû croire qu'ils avaient également un petit vélo dans la tête...

Car disons-le franchement, il faut être un peu fou pour oser s'aventurer sur les routes, comme la nationale 74, avec un tel engin : une seule roue – d'un mètre – un embryon de guidon et pas de vitesse, ni même de frein !

Et il faut l'être complètement fou pour imaginer rallier ainsi Cologne à Barcelone, soit quelque 1 600 kilomètres ! C'est pourtant l'exploit que cette petite équipe de doux dingues est en passe de réaliser.

Parti le 24 juillet dernier d'Allemagne, le groupe est arrivé mercredi à Dijon. Après un repos bien mérité il est reparti hier, jeudi, pour Tournus, via Beaune avec pour tout accompagnement une voiture balai (*).

20 kilomètres de moyenne

Aujourd'hui même il repart pour Lyon où il compte se reposer une journée. Le temps de reprendre quelques forces. « Ce sont surtout les mollets et les genoux qui fatiguent », explique John, un jeune New-Yorkais. Très à l'aise sur son « unicycle » (prononcer « you »), il dit tenir une moyenne d'environ 20 à 24 kilomètres/heure. Et d'expliquer que l'itinéraire tente d'emprunter un parcours « le plus plat possible », en longeant les fleuves.

Car avec ce genre d'engin, pas question de monter, et encore moins de descendre des cols !

D'autres, un peu moins aguerris doivent se contenter d'un petit 18 à l'heure. Reste que tous se sont durement entraînés, « notamment ces quatre derniers mois », précise John, pour mener à bien ce véritable exploit.

En moyenne, le groupe parcourt ainsi 80 à 90 kilomètres par jour, l'étape la plus longue étant de 124 kilomètres.

Si tout se passe bien, ce qui est le cas jusqu'à présent, les dix monocyclistes comptent arriver à Barcelone le 14 août.

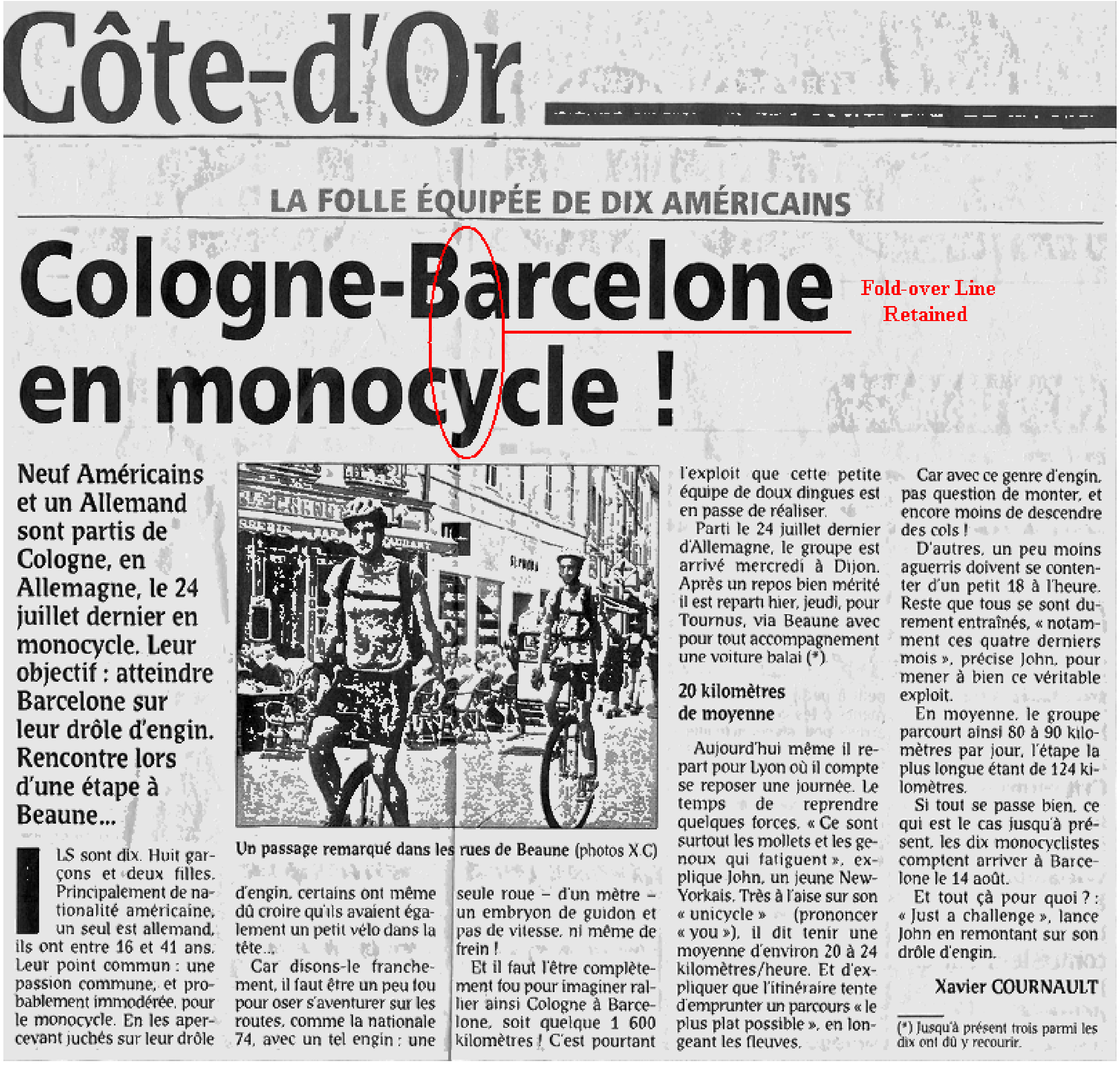
Et tout çà pour quoi ? : « Just a challenge », lance John en remontant sur son drôle d'engin.

Xavier COURNAULT

(*) Jusqu'à présent trois parmi les dix ont dû y recourir.

Figure 4.9(a) Section of an Original Newspaper Image

The use of the Improved Niblack's Algorithm in Figure 4.9(b) retained the dark and weak fold-over lines. It is sensitive to noise so that some impulse noise remains here and there in the restored greyscale image. This leads to poor reading quality.


Fold-over Line
Retained

Côte-d'Or

LA FOLLE ÉQUIPÉE DE DIX AMÉRICAINS

Cologne-Barcelone en monocycle !

Neuf Américains et un Allemand sont partis de Cologne, en Allemagne, le 24 juillet dernier en monocycle. Leur objectif : atteindre Barcelone sur leur drôle d'engin. Rencontre lors d'une étape à Beaune...

Un passage remarqué dans les rues de Beaune (photos X.C)

ILS sont dix. Huit garçons et deux filles. Principalement de nationalité américaine, un seul est allemand, ils ont entre 16 et 41 ans. Leur point commun : une passion commune, et probablement immodérée, pour le monocycle. En les apercevant juchés sur leur drôle d'engin, certains ont même dû croire qu'ils avaient également un petit vélo dans la tête...

Car disons-le franchement, il faut être un peu fou pour oser s'aventurer sur les routes, comme la nationale 74, avec un tel engin : une seule roue – d'un mètre – un embryon de guidon et pas de vitesse, ni même de frein !

Et il faut l'être complètement fou pour imaginer rallier ainsi Cologne à Barcelone, soit quelque 1 600 kilomètres ! C'est pourtant l'exploit que cette petite équipe de doux dingues est en passe de réaliser.

Parti le 24 juillet dernier d'Allemagne, le groupe est arrivé mercredi à Dijon. Après un repos bien mérité il est reparti hier, jeudi, pour Tournus, via Beaune avec pour tout accompagnement une voiture balai (*).

20 kilomètres de moyenne

Aujourd'hui même il repart pour Lyon où il compte se reposer une journée. Le temps de reprendre quelques forces. « Ce sont surtout les mollets et les genoux qui fatiguent », explique John, un jeune New-Yorkais. Très à l'aise sur son « unicycle » (prononcer « you »), il dit tenir une moyenne d'environ 20 à 24 kilomètres/heure. Et d'expliquer que l'itinéraire tente d'emprunter un parcours « le plus plat possible », en longeant les fleuves.

Car avec ce genre d'engin, pas question de monter, et encore moins de descendre des cols !

D'autres, un peu moins aguerris doivent se contenter d'un petit 18 à l'heure. Reste que tous se sont durement entraînés, « notamment ces quatre derniers mois », précise John, pour mener à bien ce véritable exploit.

En moyenne, le groupe parcourt ainsi 80 à 90 kilomètres par jour, l'étape la plus longue étant de 124 kilomètres.

Si tout se passe bien, ce qui est le cas jusqu'à présent, les dix monocyclistes comptent arriver à Barcelone le 14 août.

Et tout çà pour quoi ? : « Just a challenge », lance John en remontant sur son drôle d'engin.

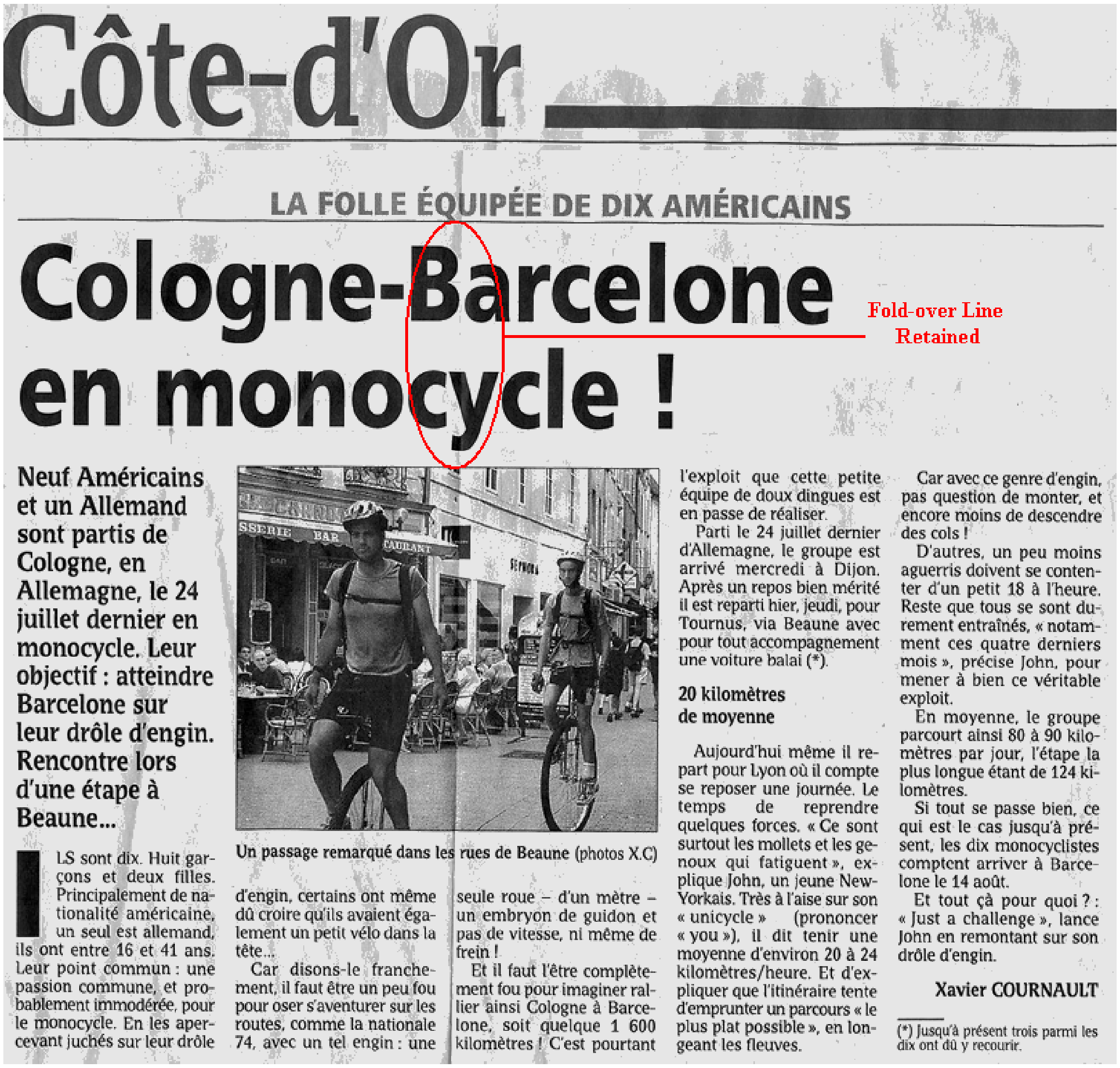
Xavier COURNAULT

(*) Jusqu'à présent trois parmi les dix ont dû y recourir.

Figure 4.9 (b) Effect of the Improved Niblack's Algorithm

on the image shown in Figure 4.9(a)

In Figure 4.9(c), the QIR algorithm also did not remove unwanted fold-over lines or detect the words which overlap with the fold-over line.



Fold-over Line Retained

Côte-d'Or

LA FOLLE ÉQUIPÉE DE DIX AMÉRICAINS

Cologne-Barcelone en monocycle !

Neuf Américains et un Allemand sont partis de Cologne, en Allemagne, le 24 juillet dernier en monocycle. Leur objectif : atteindre Barcelone sur leur drôle d'engin. Rencontre lors d'une étape à Beaune...

ILS sont dix. Huit garçons et deux filles. Principalement de nationalité américaine, un seul est allemand, ils ont entre 16 et 41 ans. Leur point commun : une passion commune, et probablement immodérée, pour le monocycle. En les apercevant juchés sur leur drôle d'engin, certains ont même dû croire qu'ils avaient également un petit vélo dans la tête...

Car disons-le franchement, il faut être un peu fou pour oser s'aventurer sur les routes, comme la nationale 74, avec un tel engin : une seule roue – d'un mètre – un embryon de guidon et pas de vitesse, ni même de frein !

Et il faut l'être complètement fou pour imaginer rallier ainsi Cologne à Barcelone, soit quelque 1 600 kilomètres ! C'est pourtant l'exploit que cette petite équipe de doux dingues est en passe de réaliser.

Parti le 24 juillet dernier d'Allemagne, le groupe est arrivé mercredi à Dijon. Après un repos bien mérité il est reparti hier, jeudi, pour Tournus, via Beaune avec pour tout accompagnement une voiture balai (*).

20 kilomètres de moyenne

Aujourd'hui même il repart pour Lyon où il compte se reposer une journée. Le temps de reprendre quelques forces. « Ce sont surtout les mollets et les genoux qui fatiguent », explique John, un jeune New-Yorkais. Très à l'aise sur son « unicycle » (prononcer « you »), il dit tenir une moyenne d'environ 20 à 24 kilomètres/heure. Et d'expliquer que l'itinéraire tente d'emprunter un parcours « le plus plat possible », en longeant les fleuves.

Car avec ce genre d'engin, pas question de monter, et encore moins de descendre des cols !

D'autres, un peu moins aguerris doivent se contenter d'un petit 18 à l'heure. Reste que tous se sont durement entraînés, « notamment ces quatre derniers mois », précise John, pour mener à bien ce véritable exploit.

En moyenne, le groupe parcourt ainsi 80 à 90 kilomètres par jour, l'étape la plus longue étant de 124 kilomètres.

Si tout se passe bien, ce qui est le cas jusqu'à présent, les dix monocyclistes comptent arriver à Barcelone le 14 août.

Et tout çà pour quoi ? : « Just a challenge », lance John en remontant sur son drôle d'engin.

Xavier COURNAULT

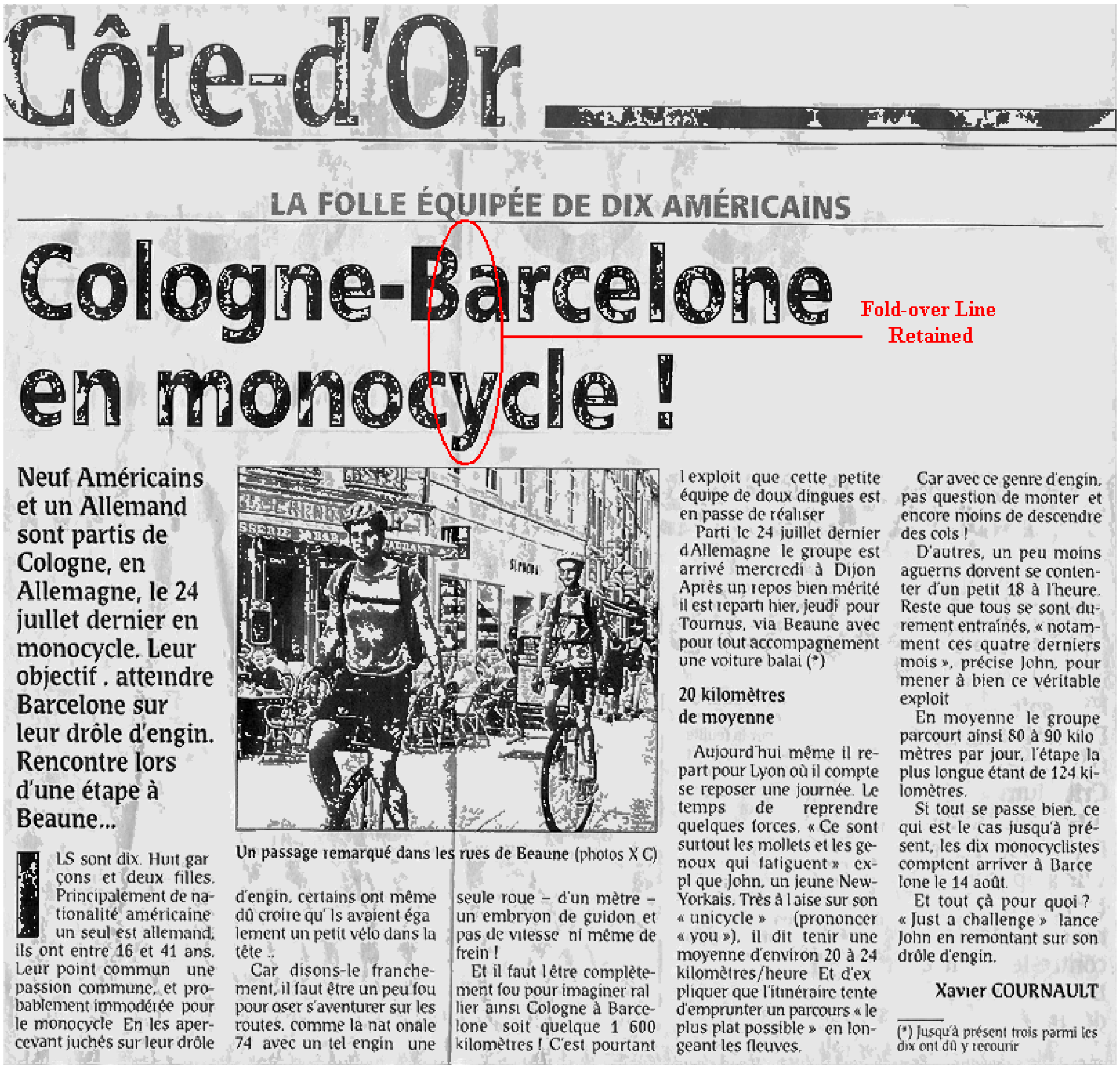
Un passage remarqué dans les rues de Beaune (photos X.C)

(*) Jusqu'à présent trois parmi les dix ont dû y recourir.

Figure 4.9 (c) Effect of the Original QIR Algorithm

on the image shown in Figure 4.9 (a)

Yanowitz and Bruckstein's Algorithm also failed to remove the fold over lines. Figure 4.9(d) shows that the algorithm can detect the text that overlapped with the foldover lines, but it cannot remove the foldover lines, and it also cannot retain all the details of some words.


Fold-over Line
Retained

Côte-d'Or

LA FOLLE ÉQUIPÉE DE DIX AMÉRICAINS

Cologne-Barcelone en monocycle !

Neuf Américains et un Allemand sont partis de Cologne, en Allemagne, le 24 juillet dernier en monocycle. Leur objectif . atteindre Barcelone sur leur drôle d'engin. Rencontre lors d'une étape à Beaune...

Un passage remarqué dans les rues de Beaune (photos X C)

ILS sont dix. Huit garçons et deux filles. Principalement de nationalité américaine un seul est allemand, ils ont entre 16 et 41 ans. Leur point commun une passion commune, et probablement immodérée pour le monocycle En les apercevant juchés sur leur drôle d'engin, certains ont même dû croire qu' ls avaient éga lement un petit vélo dans la tête ..

Car disons-le franchement, il faut être un peu fou pour oser s'aventurer sur les routes, comme la nat onale 74 avec un tel engin une seule roue – d'un mètre – un embryon de guidon et pas de vitesse ni même de frein !

Et il faut l être complètement fou pour imaginer ral lier ainsi Cologne à Barcelone soit quelque 1 600 kilomètres ! C'est pourtant l exploit que cette petite équipe de doux dingues est en passe de réaliser

Parti le 24 juillet dernier dAllemagne le groupe est arrivé mercredi à Dijon Après un repos bien mérité il est reparti hier, jeudi pour Tournus, via Beaune avec pour tout accompagnement une voiture balai (*)

20 kilomètres de moyenne

Aujourd'hui même il repart pour Lyon où il compte se reposer une journée. Le temps de reprendre quelques forces. « Ce sont surtout les mollets et les genoux qui fatiguent » explique John, un jeune New-Yorkais. Très à l aise sur son « unicycle » (prononcer « you »), il dit tenir une moyenne d'environ 20 à 24 kilomètres/heure Et d'ex pliquer que l'itinéraire tente d'emprunter un parcours « le plus plat possible » en longeant les fleuves.

Car avec ce genre d'engin, pas question de monter et encore moins de descendre des cols !

D'autres, un peu moins aguerris doivent se contenter d'un petit 18 à l'heure. Reste que tous se sont durement entraînés, « notamment ces quatre derniers mois », précise John, pour mener à bien ce véritable exploit

En moyenne le groupe parcourt ainsi 80 à 90 kilo mètres par jour, l'étape la plus longue étant de 124 kilomètres.

Si tout se passe bien, ce qui est le cas jusqu'à présent, les dix monocyclistes comptent arriver à Barce lone le 14 août.

Et tout çà pour quoi ? « Just a challenge » lance John en remontant sur son drôle d'engin.

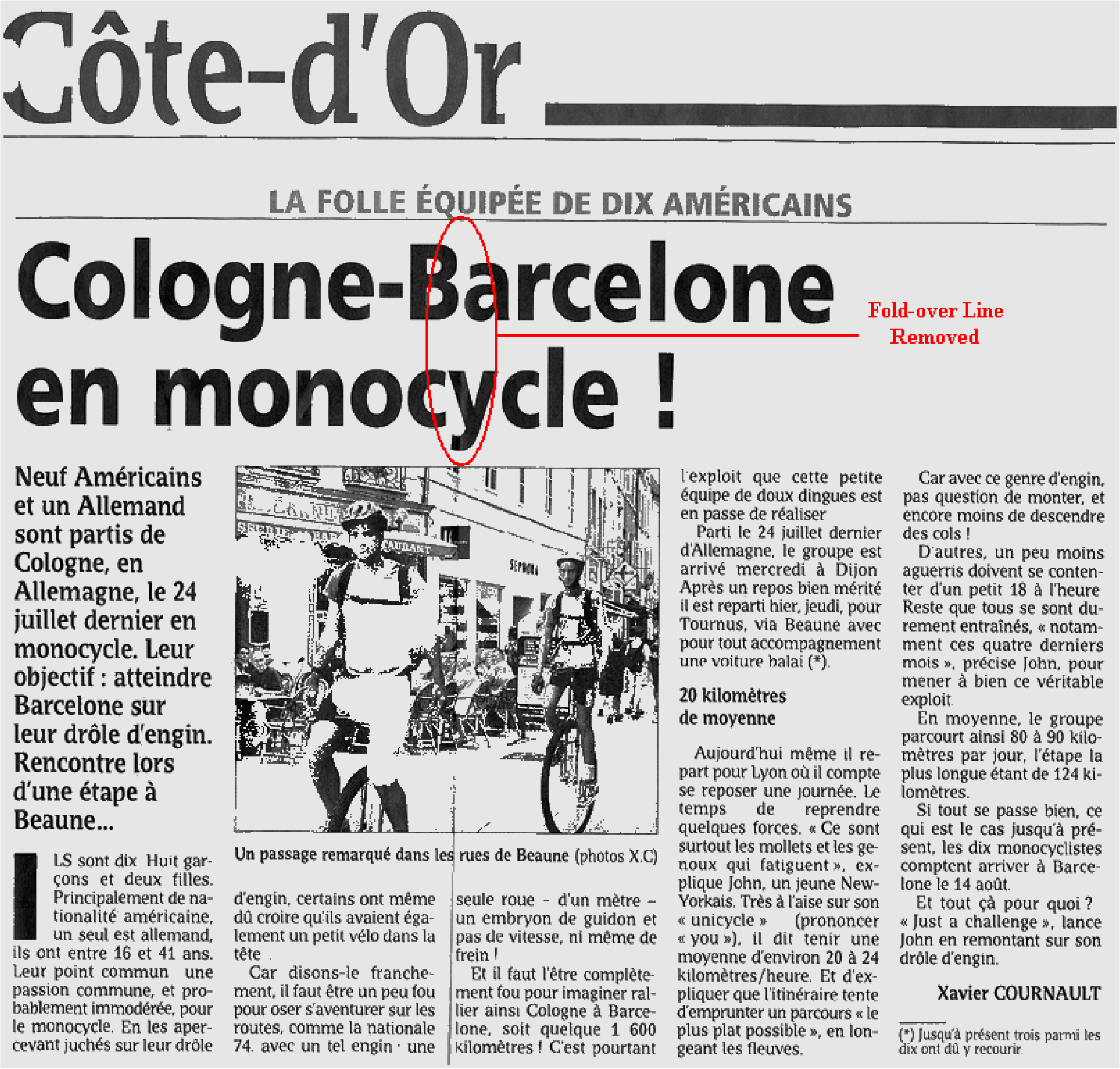
Xavier COURNAULT

(*) Jusqu'à présent trois parmi les dix ont dû y recourir

Figure 4.9 (d) Effect of the Yanowitz and Bruckstein's Algorithm

on the image shown in Figure 4.9(a)

The result of the mean-gradient method in Figure 4.9(e) shows a clear binary image without fold-over lines. The Mean-Gradient technique removed unwanted noise of the fold-over lines and retains most of the word details. Compared to the other three algorithms, Mean-Gradient works well for removing fold-over lines and correctly retains words in newspaper images.



Côte-d'Or

LA FOLLE ÉQUIPÉE DE DIX AMÉRICAINS

Cologne-Barcelone en monocycle !

Neuf Américains et un Allemand sont partis de Cologne, en Allemagne, le 24 juillet dernier en monocycle. Leur objectif : atteindre Barcelone sur leur drôle d'engin. Rencontre lors d'une étape à Beaune...

Un passage remarqué dans les rues de Beaune (photos X.C)

ILS sont dix Huit garçons et deux filles. Principalement de nationalité américaine, un seul est allemand, ils ont entre 16 et 41 ans. Leur point commun une passion commune, et probablement immodérée, pour le monocycle. En les apercevant juchés sur leur drôle d'engin, certains ont même dû croire qu'ils avaient également un petit vélo dans la tête .

Car disons-le franchement, il faut être un peu fou pour oser s'aventurer sur les routes, comme la nationale 74. avec un tel engin · une seule roue - d'un mètre - un embryon de guidon et pas de vitesse, ni même de frein !

Et il faut l'être complètement fou pour imaginer rallier ainsı Cologne à Barcelone, soit quelque 1 600 kilomètres ! C'est pourtant l'exploit que cette petite équipe de doux dingues est en passe de réaliser

Parti le 24 juillet dernier d'Allemagne, le groupe est arrivé mercredi à Dijon Après un repos bien mérité il est reparti hier, jeudi, pour Tournus, via Beaune avec pour tout accompagnement une voiture balai (*).

20 kilomètres de moyenne

Aujourd'hui même il repart pour Lyon où il compte se reposer une journée. Le temps de reprendre quelques forces. « Ce sont surtout les mollets et les genoux qui fatiguent », explique John, un jeune New-Yorkais. Très à l'aise sur son « unicycle » (prononcer « you »), il dit tenir une moyenne d'environ 20 à 24 kilomètres/heure. Et d'expliquer que l'itinéraire tente d'emprunter un parcours « le plus plat possible », en longeant les fleuves.

Car avec ce genre d'engin, pas question de monter, et encore moins de descendre des cols !

D'autres, un peu moins aguerris doivent se contenter d'un petit 18 à l'heure Reste que tous se sont durement entraînés, « notamment ces quatre derniers mois », précise John, pour mener à bien ce véritable exploit.

En moyenne, le groupe parcourt ainsi 80 à 90 kilomètres par jour, l'étape la plus longue étant de 124 kilomètres.

Si tout se passe bien, ce qui est le cas jusqu'à présent, les dix monocyclistes comptent arriver à Barcelone le 14 août.

Et tout çà pour quoi ? « Just a challenge », lance John en remontant sur son drôle d'engin.

Xavier COURNAULT

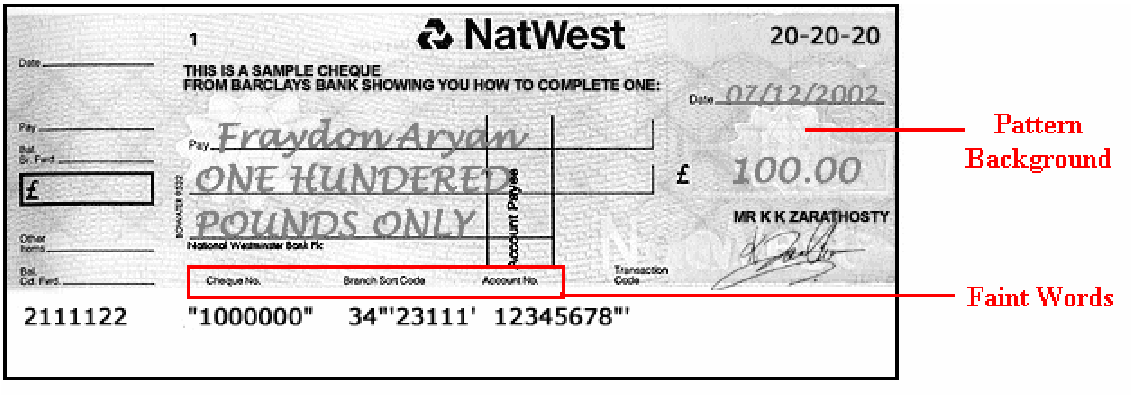
(*) Jusqu'à présent trois parmi les dix ont dû y recourir.

Figure 4.9 (e) Effect of the Mean-Gradient Algorithm

on the image shown in Figure 4.9(a)

4.4.1.4 Group 4: Cheque Document Image

Figure 4.10(a) shows a scanned image of a cheque. Frequently, a cheque can include all kinds of patterned background, which are designed by the banks. One of the important requirements in binarization of cheque images is to detect and extract the text from the patterned backgrounds. The four techniques are compared below using the image in Figure 4.10(a) of a patterned background scanned cheque.



Date
Pay
Bal. Br. Fwd
£
Other Items
Bal. Cd. Fwd.

1 NatWest 20-20-20
THIS IS A SAMPLE CHEQUE
FROM BARCLAYS BANK SHOWING YOU HOW TO COMPLETE ONE:
Date 07/12/2002
Pay Fraydon Aryan
ONE HUNDERED
POUNDS ONLY
Account Payee
£ 100.00
MR K K ZARATHOSTY
National Westminster Bank Plc
Cheque No. Branch Sort Code Account No. Transaction Code
2111122 "1000000" 34"'23111' 12345678"'

Figure 4.10(a) Original Cheque Image

The Improved Niblack's Algorithm detected the text that overlapped with the pattern background as shown in Figure 4.10(b). It lost details of some useful words and still contains a considerable amount of unwanted pattern and noise.

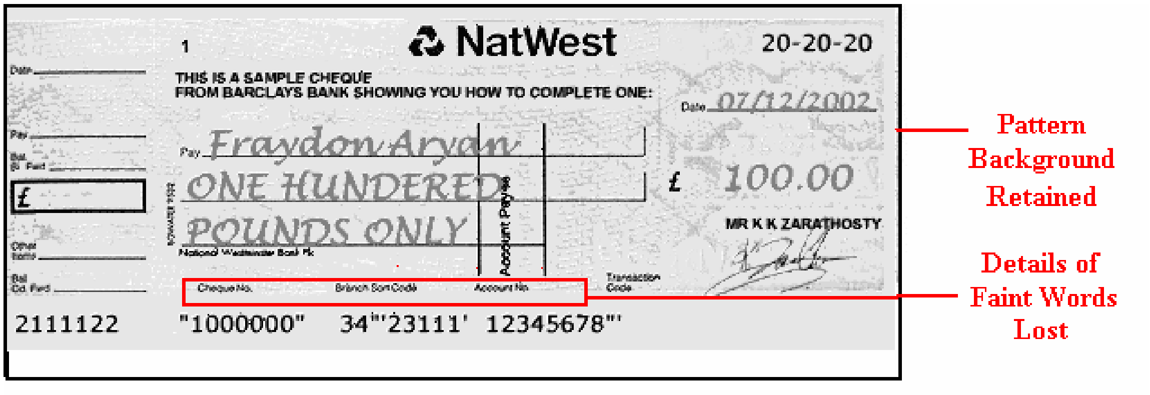


Figure 4.10 (b) Effect of the Improved Niblack Algorithm

on the image shown in Figure 4.10(a)

Figure 4.10(c) shows that the QIR method also cannot remove the patterned background, while the detected text overlapped with the patterns, which leads to the result that some words will be confused in the future word segmentation or recognition.

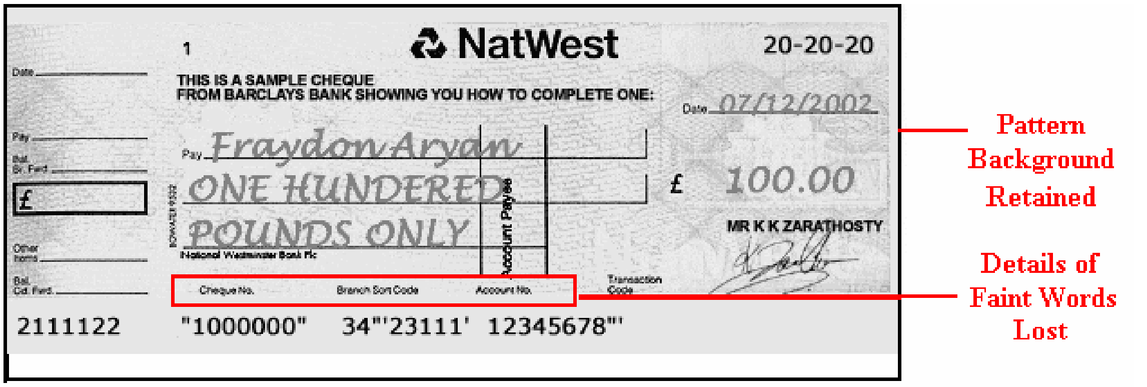


Figure 4.10 (c) Effect of the Original QIR Algorithm

on the image shown in Figure 4.10(a)

Yanowiz's Algorithm failed to remove patterns from the image, and its ability to be edge sensitive makes it confuse the boundaries of useful words with useless pattern, so that many detected words have sharp edges. Yanowiz's Algorithm is not suitable for processing this kind of cheque image.

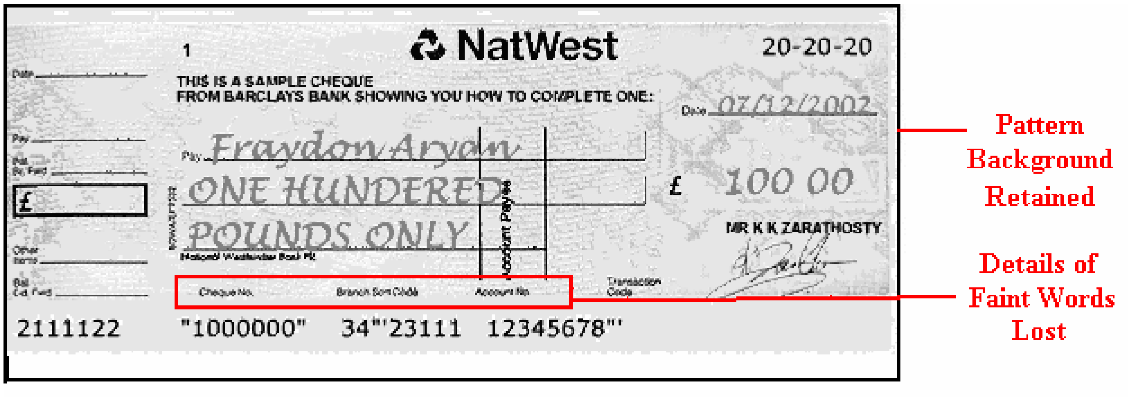


Figure 4.10 (d) Effect of the Yanowiz's Algorithm

on the image shown in Figure 4.10(a)

In Figure 4.10(e), the unwanted patterned background is almost removed by the Mean-Gradient Method, and the faint strokes were retained well for soft reading.

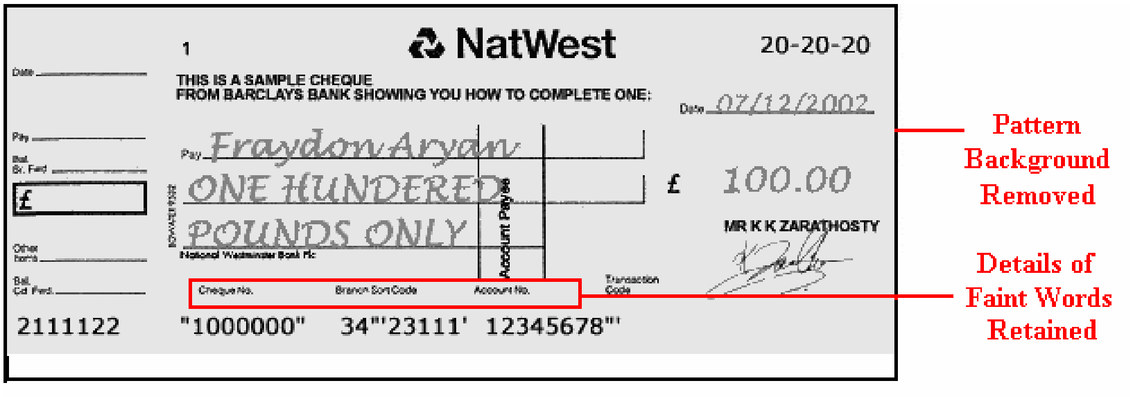


Figure 4.10 (e) Effect of the Mean-Gradient Algorithm

on the image shown in Figure 4.10(a)

Compared to the other three methods, the Mean-Gradient Method retains more detail of the handwriting while removing the background in different types of document image. For the further following steps of segmentation and recognition, what we really are interested in is the retained greyscale pixels and the Mean-Gradient Method can provide the most useful foreground information, but the least useful background.

4.4.2 Quantitative Comparison of Thresholding Methods

A comparison of the restored greyscale image results for qualitative comparison was detailed in the previous section. In this section, two performance metrics are used to quantify the quality of the four thresholding methods on the four different types of document images.

The well known Information Retrieval standard measures, precision and recall (**Definitions 4 and 5 in Section 2.2.4**), were used to compare the performance of the proposed method with the other three thresholding algorithms.

Precision and Recall are defined as follows:

Precision = (Correctly Detected Words / Total Detected Words)×100%,

Recall = (Correctly Detected Words / Total Actual Words)×100%

The Precision and Recall calculations in this thesis are defined as word – based revaluation methods, which can be used to compare the performance of different thresholding methods for degraded binary handwriting document images. According to Solihin & Leedham's [83] suggestions, there are 3 main aspects to describe the quality of detected binary words. 1. Whether all details of handwriting are retained or not, eg: faint skate-on and skate-off pen strokes at the beginning and the end of strokes; 2. is the patterned/noisy background removed? 3. are all handwriting retained? Each word in the original greyscale image is compared using human vision with the corresponding word(s) in the binary image(s) using the conditions mentioned above.

The Precision value describes how many correct words remain in the binary image, and the Recall value describes how many correct words are detected out from original image. The Precision and Recall values can be used to quantity, the effectiveness of document thresholding algorithm. Before the evaluation calculation, the criterion for correct detection is an important step, Solihin & Leedham (1999).

The results of Precision and Recall values for evaluating the Mean-Gradient Technique with other three techniques are showed in Tables 4.1 to 4.4.

Table 4.1 Precisions and Recall Evaluations for Historical Document Images

Image No Algorithm	1 (215 Words)		2 (120 Words)		3 (80 Words)		4 (230 Words)		5 (443 Words)		6 (183 Words)	
	P (%)	R (%)	P (%)	R (%)	P (%)	R (%)	P (%)	R (%)	P (%)	R (%)	P (%)	R (%)
Mean-Gradient	96	92	73	69	96	96	96	92	73	69	96	96
QIR	78	75	64	53	72	72	78	75	64	53	72	72
Yanowiz's	51	51	68	68	98	98	51	51	68	68	98	98
Improved Niblack	37	37	40	40	76	76	37	37	40	40	76	76

Image No Algorithm	7 (121 Words)		8 (85 Words)		9 (91 Words)		10 (103 Words)		Average	
	P (%)	R (%)	P (%)	R (%)	P (%)	R (%)	P (%)	R (%)	P (%)	R (%)
Mean-Gradient	88.	87	77	77	98	98	100.00	100.00	**89**	**87**
QIR	84	82	80	80	98	98	100.00	100.00	81	79
Yanowiz's	93	92	85	85	100	100	96	96	84	84
Improved Niblack	91	91	79	79	73	73	88	88	74	74

P.S: P = Precision; R = Recall.

In Table 4.1, the ranked order of average Precision and Recall values from high to low are: Mean-Gradient Algorithm, Yanowitz's Algorithm, Original QIR Algorithm and

Improved Niblack's Algorithm. Compared to the other three algorithms, the Mean-Gradient Algorithm can detect more words in the original historical images.

In Table 4.2, the order of average Precision and Recall values for the Form Document Images from high to low are: Mean-Gradient Algorithm, Original QIR Algorithm, Improved Niblack's Algorithm and Yanowitz's Algorithm. Compared to the other three algorithms, the Mean-Gradient Algorithm produces cleaner and less noisy result in the original form images.

Table 4.2 Precisions and Recall Evaluations For Form Document Images

Image No **Algorithm**	**1** (101 Words)		**2** (199 Words)		**3** (68 Words)		**4** (101 Words)		**5** (49 Words)		**6** (160 Words)	
	P *(%)*	**R** *(%)*	**P** *(%)*	**R** *(%)*	**P** *(%)*	**R** *(%)*	**P** *(%)*	**R** *(%)*	**P** *(%)*	**R** *(%)*	**P** *(%)*	**R** *(%)*
Mean-Gradient	93	93	54	52	70	67	85	84	100	100	96	96
QIR	94	94	95	95	73	73	75	75	91	91	95	94
Yanowiz's	88	88	73	73	14	14	4	4	0	0	25	25
Improved Niblack	63	63	80	80	75	75	12	12	94	94	80	64

Image No **Algorithm**	**7** (418 Words)		**8** (342 Words)		**9** (45 Words)		**10** (164 Words)		**Average**	
	P *(%)*	**R** *(%)*	**P** *(%)*	**R** *(%)*	**P** *(%)*	**R** *(%)*	**P** *(%)*	**R** *(%)*	**P** *(%)*	**R** *(%)*
Mean-Gradient	99	99	94	94	100	100	88	86	**88**	**87**
QIR	16	4	22	7	100	100	98	98	76	73
Yanowiz's	17	17	21	21	100	100	2	2	34	34
Improved Niblack	27	27	30	30	100	100	93	93	65	64

P.S: P = Precision; R = Recall.

In Table 4.3, the order of average Precision values for Cheque Document Image from high to low are: Mean-Gradient Algorithm, Original QIR Algorithm, Improved

Niblack's Algorithm and Yanowitz's Algorithm. Recall values for Cheque Document Image from high to low are: Mean-Gradient Algorithm, Improved Niblack's Algorithm and Yanowitz's Algorithm and Original QIR Algorithm. Compared to the other three algorithms, the Mean-Gradient Algorithm can detect more words in the original cheque images from complex pattern background.

Table 4.3 Precisions and Recall Evaluation for Cheque Document Images

Image No Algorithm	1 (21 Words)		2 (51 Words)		3 (32 Words)		4 (31 Words)		5 (24 Words)		6 (35 Words)	
	P **(%)**	**R** **(%)**	**P** ***(%)***	**R** ***(%)***	**P** ***(%)***	**R** ***(%)***	**P** ***(%)***	**R** ***(%)***	**P** ***(%)***	**R** ***(%)***	**P** ***(%)***	**R** ***(%)***
Mean-Gradient	95	95	79	76	93	93	83	83	100	100	83	74
QIR	95	95	47	21	90	90	42	22.58	37	37	6.6	2.8
Yanowiz's	52	52	0	0	14	14	59	59.38	79	79	70.	70
Improved Niblack	56	56	55	22	62	62	62	62.50	58	58	22	22

Image No Algorithm	7 (46 Words)		8 (33 Words)		9 (13 Words)		10 (72 Words)		**Average**	
	P **(%)**	**R** **(%)**	**P** ***(%)***	**R** ***(%)***	**P** ***(%)***	**R** ***(%)***	**P** ***(%)***	**R** ***(%)***	**P** ***(%)***	**R** ***(%)***
Mean-Gradient	97	97	96	96	92	92	92	80	91	89
QIR	79	41	100	51	11	72	96	88	60	45.
Yanowiz's	100	100	100	100	0	0	53	53	52	52
Improved Niblack	93	93	96	96	40	40	11	11	55	52

P.S: P = Precision; R = Recall.

In Table 4.4, the order of average Precision and Recall values for Newspaper Document Image from high to low is: Mean-Gradient Algorithm, Yanowitz's Algorithm, Improved Niblack's Algorithm and Original QIR Algorithm. Compared to the other three algorithms, the Mean-Gradient Algorithm can detect more words in the original newspaper images, which include noisy background.

Table 4.4 Precisions and Recall Evaluation for Newspaper Document Images

Image No / Algorithm	1 (629 Words)		2 (502 Words)		3 (312 Words)		4 (344 Words)		5 (531 Words)		6 (431 Words)		7 (380 Words)		8 (487 Words)		9 (643 Words)		10 (589 Words)		Average	
	P (%)	R (%)	P (%)	R (%)	P (%)	R (%)	P (%)	R (%)	P (%)	R (%)	P (%)	R (%)	P (%)	R (%)	P (%)	R (%)	P (%)	R (%)	P (%)	R (%)	P (%)	R (%)
Mean-Gradient	95	95	99	99	3	2	95	95	99	99	3	2	100	100	100	100	75	100	100	100	**77**	**79**
QIR	43	42	99	99	1	1	43	42	99	99	2	1	0	0	96	96	90	0	100	96	57.	48
Yanowiz's	91	89	94	94	7	7	91	89	94	94	8	7	99	99	98	98	75	99	99	98	75	77
Improved Niblack	90	90	83	83	3	3	90	90	83	83	3	3	41	41	97	97	60	41	41	97	59	63

P.S: P = Precision; R = Recall.

From the above, it can be concluded that the Mean-Gradient Algorithm produces good performance on several different image types. The other three thresholding techniques evaluated techniques only perform well on selected image types.

4.5 Summary

Yanowitz's technique uses a median filter in its pre-processing stage to eliminate the noise. The effect of this filter, however, reduces the handwritten contrast and fills holes in both handwriting and printed characters producing thickened characters. The resulting binary images therefore have worse quality since these characters are not distinguishable, especially if the original image has poor resolution as observed in the examined form and cheque images.

The Improved Niblack's method is simple, but it is sensitive to the constant values in the equation. It is difficult to find weighted values that produce good results for different images. The computational complexity of the Improved Niblack's method is: $O(N^2)$ for an $N \times N$ image.

Performance evaluation of several binarization methods has shown that the Yanowitz and Bruckstein's method was one of the best binarization methods. The computational complexity of the Yanowitz and Bruckstein's method is: $O(N^2)$ for an $N \times N$ image.

QIR works quite well on images that have two distinct peaks in their histograms, meaning high constant homogeneous images. Due to the fact that the technique depends on the bimodal histogram, it is not suitable for some images, such as those that have double-side effect as well as those that contain noisy backgrounds. The computational complexity of the QIR method is: $O(N)$ for an $N \times N$ image.

The Mean-Gradient technique works well to retain the high contrast strokes, and also it is generally successful in retaining the small holes in characters. Nevertheless it performs well in removing patterned backgrounds. The Mean-gradient values can be regarded as a variation of an image and describes the ripple gradient of an image. The computational complexity of the Mean-Gradient technique is: $O(N^2)$ for an $N \times N$ image.

Figure 4.8(e) shows the Mean-Gradient method is particularly effective at removing blotches and smudges in images while still maintaining the handwriting details. This is because large objects are considered as part of the background during the process, and therefore will be subtracted from the image. For some images with very light strokes, the mean-gradient technique misses some of them, resulting in broken handwriting.

The Mean-Gradient technique generally achieves better precision and recall than the other methods for all the categories of images. It subtracts out the ink seeping and double-sided effect in historical documents to produce clean binary images. The technique also performs well for high/low resolution as well as large/small printed characters present in newspapers, forms and cheques. However, with the usage of the background subtraction method for removing noise, it tends to over-threshold some of the weak handwriting, resulting in broken handwriting.

From the comparison of the four techniques, it is concluded that a simple thresholding technique, which works perfectly for all kinds of images, does not exist. However, in a document management system, there are many different document images, which need to be processed. This makes it a requirement of a thresholding technique to perform well on many image types. Most of exiting thresholding techniques apply same process on the whole image. They cannot apply the best suitable process for different image sub-region with different conditions. A feature vectors analysis based multi-stage local adaptive technique (Decompose Thresholding Approach) is proposed as a solution to this problem.

Chapter 5

Decompose Thresholding Approach

A number of techniques have previously been proposed for thresholding document images. The global and local adaptive techniques have met with varying degrees of success. While the existing methods apply the same process to the entire image even if the image contains different characteristics in different area, none of them is able to produce consistently good results on a wide range of degraded historical documents. A review of previously reported thresholding techniques was given in Chapter 2.

In this chapter a new thresholding structure called the **decompose-threshold approach** is proposed and compared against some existing global and local thresholding algorithms. The chapter first describes the decompose algorithm and its improved version followed by experimental analysis. A set of six thresholding algorithms, which have shown a superior level of performance on 'difficult' images, were chosen for incorporation in this work for performance comparison purposes.

The decompose thresholding method described in this chapter was published in *Proc. 17th Int. Conf. on Pattern Recognition*, Cambridge, United Kingdom, Vol.1, pp 445-448, 2004; *Proc. 9th Int. Workshop on Frontiers in Handwriting Recognition*, Tokyo, Japan, pp 239-244, 2004; and is accepted for publication in the *IEE Proceedings on Vision, Image and Signal Processing*.

5.1 The Proposed Decompose Algorithm

A local adaptive analysis method, called the Decompose Algorithm, uses local feature vectors to find the best approach to thresholding a local area. Appropriate algorithm(s) are selected or combined automatically for specific types of document image under investigation. The original image is recursively broken down into sub-regions using quad-tree decomposition until an appropriate thresholding method can be applied to each of the sub-region.

The proposed decompose algorithm is shown in Figure 5.1. The initial step of the algorithm tests whether the image has a bimodal histogram. If it does, then histogram-based global thresholding methods, which have been proven to have outstanding results for bimodal histogram images can be applied. Global methods are particularly effective at saving processing time (complexity is equal to O(n)). The decompose algorithm focuses on 8-bit grey-scale images of historical handwritten documents, which generally do not have a bimodal histogram. The decompose algorithm is a local adaptive analysis method, which uses local feature vectors to find the best approach for thresholding a local area. Appropriate weighted values are selected automatically for the specific types of document image regions under investigation. Due to the characteristics of pen strokes and spots on the paper, there are three main classes of sub-region that can be defined: **Faint Strokes Class**, **Heavy Strokes Class** and **Background Class**.

The outline of the decompose algorithm is:

1. Bimodal Testing
2. Decompose image into four equal size local regions
3. Extract feature vectors from each local region

4. Perform local region classification
5. Repeat steps 2, 3 and 4 until all regions are classified
6. Appropriate threshold methods are applied to each region
7. Smooth the edges of each region

If the image does not have a bimodal histogram, it is recursively decomposed using quad-tree decomposition into smaller regions until an appropriate local threshold method can be applied to each different class region (faint strokes class, heavy strokes class and background class). This continues until the whole image has been decomposed and an appropriate threshold technique assigned to each region. The new algorithm analyzes the feature information of the local regions of different sizes, and determines and applies different threshold methods to obtain the best result. There are three outstanding threshold methods, which are Yanowitz and Bruckstein's, Bersen's and improved Niblack's methods, are chosen. Section 5.1.5 describes more explanation on it.

The complexity of this algorithm can be sufficiently described by the function $g(n) = (1/d)n^2$ where n is the number of pixels in the input grey-scale historical image, d is the minimum size of sub region). Formally, this algorithm is "of order $O((1/5)n^2)$". This function expresses the worst-case scenario when the minimum size of sub region is set as 5 from particular results.

$g(n) = T_{BinodalTest} + T_{Decompose} + \max(T_{regionX})$ (where $_X \leq n^2/25$; $T_{Decompose} = T_{BinodalTest} = 1$). For each classified sub region, feature vectors extraction and local thresholding cost $N^2 \times 5$, where N^2 is the size of the sub region. The biggest sub region $((n/5) \times (n/5))$ cost the longest time for computing: $((n/5) \times (n/5)) \times 5 = (1/5)n^2$.

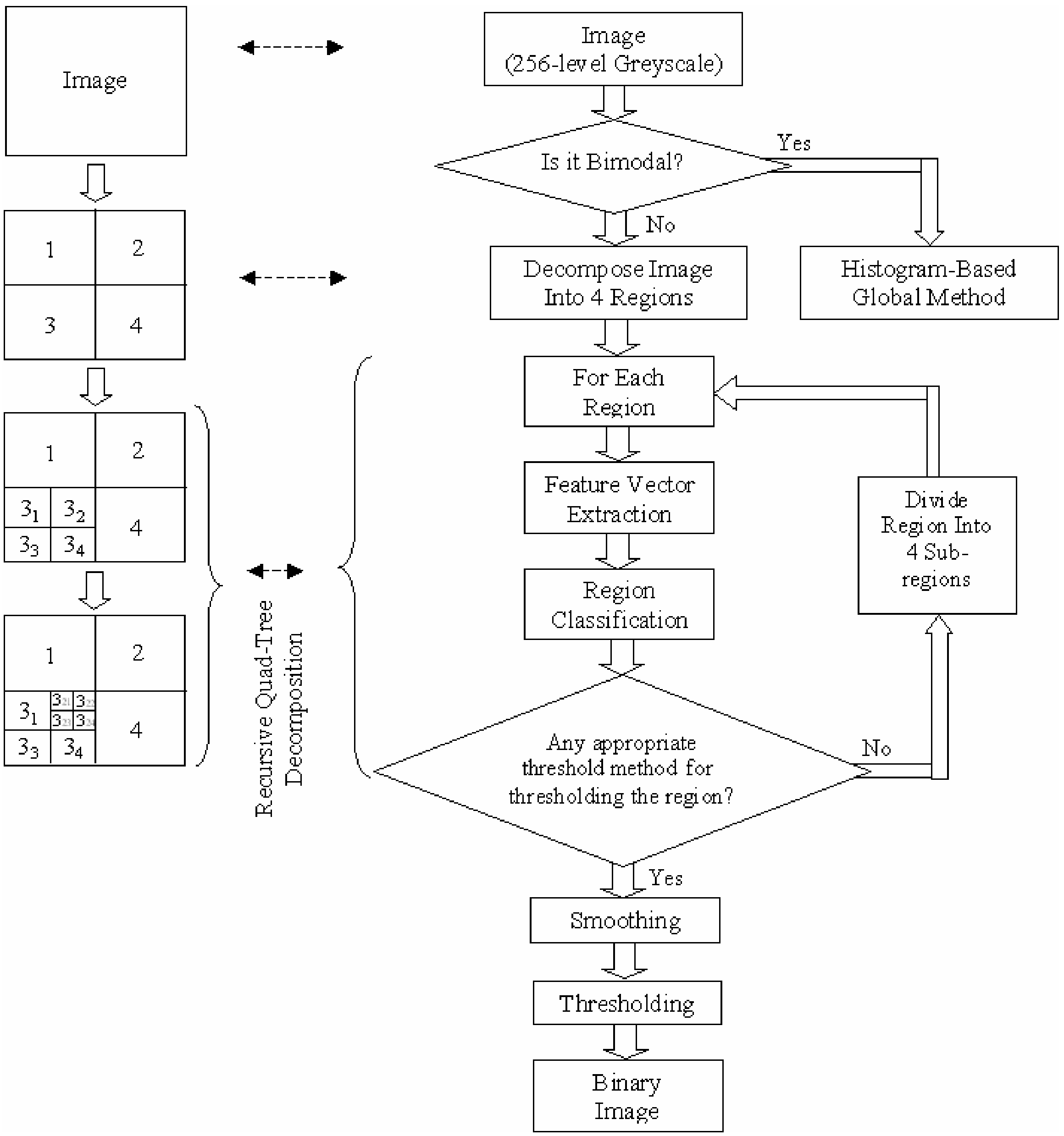


Image
1
2
3
4
1
2
3_1
3_2
3_3
3_4
4
1
2
3_1
3_{31}
3_{32}
3_{33}
3_{34}
3_3
3_4
4
Recursive Quad-Tree Decomposition
Image (256-level Greyscale)
Is it Bimodal?
Yes
No
Decompose Image Into 4 Regions
Histogram-Based Global Method
For Each Region
Feature Vector Extraction
Region Classification
Divide Region Into 4 Sub-regions
Any appropriate threshold method for thresholding the region?
No
Yes
Smoothing
Thresholding
Binary Image

Figure 5.1 Flow Chart of the Decompose Thresholding Algorithm

5.1.1 Image Bimodality Testing

A test for bimodality is performed on the histogram to determine whether a global thresholding method can be applied effectively. For an image, that has an obvious bimodal histogram, for example dark black ink on clean white paper, an existing histogram-based global thresholding method can usually be applied to the whole image. Figure 5.2 shows two examples of historical images illustrating the difference between 'unimodal' and 'bimodal' histograms. Figure 5.2(a) is the histogram of Figure 5.2(b), which is referred to as a unimodal historical document image. Figure 5.2(c) is the histogram of Figure 5.2(d), which is referred to as a bimodal historical document image. The unimodal historical image has various characteristics in different areas, which are difficult to binarize accurately using a global threshold value. The bimodal historical image, on the other hand, contains balances characteristics all over the whole image and is easier to threshold using an appropriate global threshold.

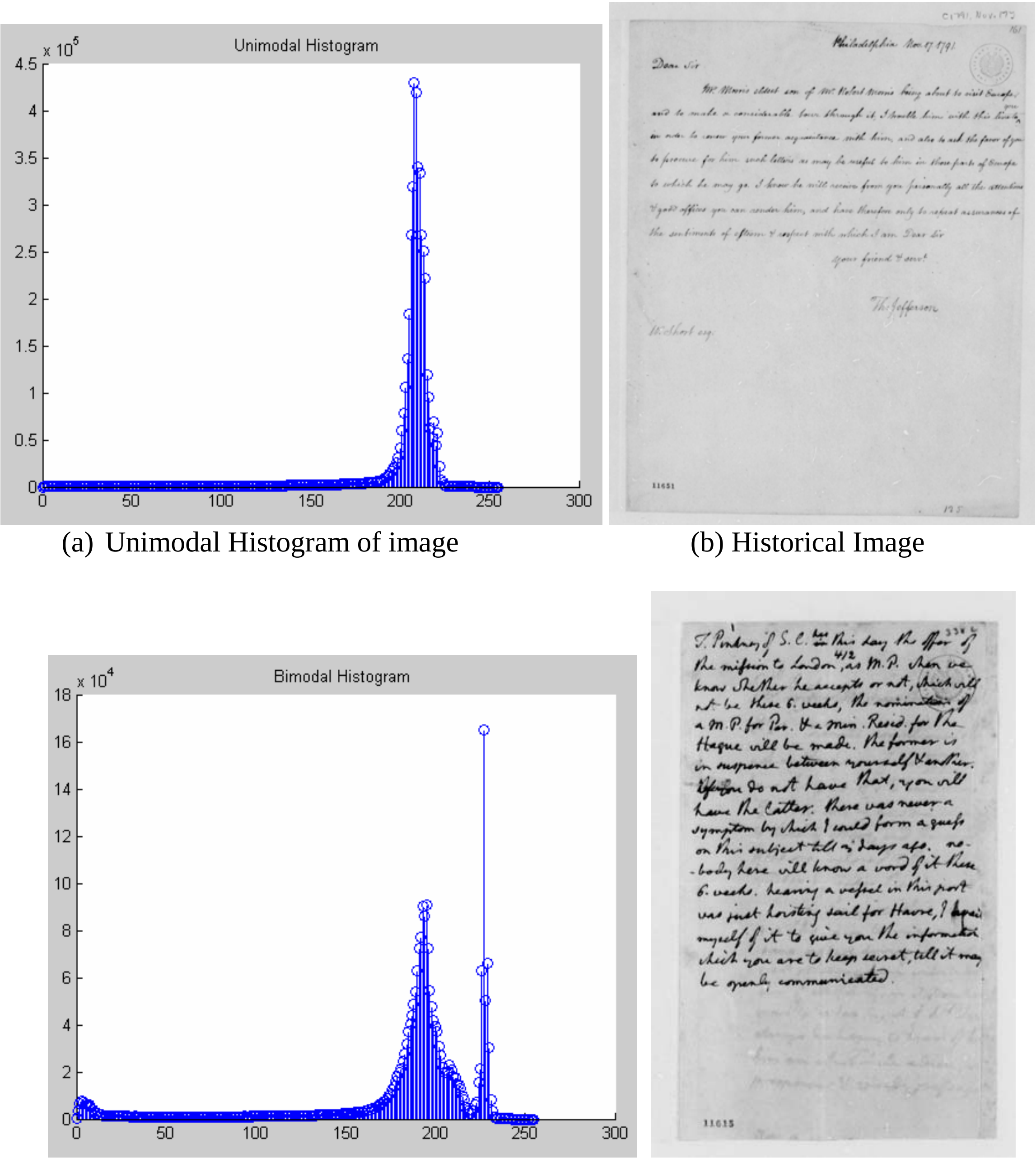
There are two branches following the bimodal testing. One is to threshold the input image using the original QIR [83] or other global method. The other branch is to apply the decompose approach and analyze local feature vectors to find the best approach for thresholding each sub-divided local area.

A simple histogram classification method is used. If the histogram of the image contains two obvious peaks, and the valley between the two peaks is located, then the histogram is bimodal; otherwise the histogram is unimodal (or multi-modal).

The bimodal testing is carried out in three steps:

1. Histogram smoothing
2. Peak number calculation
3. Testing whether there is a clear valley between the two peaks

If the peak number is equal to two and there is the valley between the two peaks, the image has a bimodal histogram.



C1791, Nov. 17

161

Philadelphia Nov. 17. 1791.

Dear Sir

Mr. Morris eldest son of Mr. Robert Morris being about to visit Europe, and to make a considerable tour through it, I trouble you with this line in order to renew your former acquaintance with him, and also to ask the favor of you to procure for him such letters as may be useful to him in those parts of Europe to which he may go. I know he will receive from you personally all the attentions & good offices you can render him, and have therefore only to repeat assurances of the sentiments of esteem & respect with which I am Dear Sir

your friend & servt

Th: Jefferson

W. Short esq.

11651

T. Pinckney of S.C. has this day the offer of the mission to London, 412 as M.P. when we know whether he accepts or not, which will not be these 6. weeks, the nomination of a M.P. for Par. & a Min. Resid. for the Hague will be made. the former is in suspence between yourself & another. if you do not have that, you will have the latter. there was never a symptom by which I could form a guess on this subject till 3 days ago. nobody here will know a word of it these 6. weeks. hearing a vessel in this port was just hoisting sail for Havre, I avail myself of it to give you the information which you are to keep secret, till it may be openly communicated.

11635

(a) Unimodal Histogram of image (b) Historical Image

(c)Bimodal Histogram of image (d) (d) Historical Image

Figure 5.2 Examples of Unimodal and Bimodal Histograms

5.1.2 Image Decomposition

Contrast variation is a major problem in historical document images, and causes many of the previously reported thresholding methods to fail. Contrast variation is therefore a characteristic that can be used to determine whether a thresholding algorithm can be applied or whether the image needs to be further decomposed.

The input image is first decomposed into four equal size sub-images if $C \geq T$ (where C is the contrast of the decomposed input image, T is currently empirically set at a greyscale value of 180 based on experimental results of test images.), and then the sub-image is further decomposed. It can be achieved in five steps:

1. Set the input image as a sub-image;
2. Calculate the mean value of the 24-neighbours (5×5 window) of the pixels that lie at coordinates $(2 \times M + 1, 2 \times N + 1)$ of the sub-image. Where

$$\begin{cases} M = 1, 2, \ldots, \frac{1}{2}(row\, number\, of\;\, sub\, image) \\ N = 1, 2, \ldots, \frac{1}{2}(column\, number\, of\;\, sub\, image) \end{cases}$$

3. Find the maximum mean value *Maximean* and minimum mean value *Minimean* of the sub-image;
4. Contrast = *Maximean – Minimean*;
5. If $C \geq T$, the sub-image is decomposed into four smaller equal-sized sub-images;
6. Repeat steps 2 to 4 until $C < T$, or when the sub-image reaches 64×64. A sub-image, which is smaller than 64×64, has insufficient information for further feature extraction.

5.1.3 Feature Extraction

Many feature vectors have been used in document image binarization. Most of them were applied to printed documents with clean (white) background but do not work well for degraded images. Three feature vectors are proposed in this paper, which focus on handwritten document images with messy background and faded writing. These feature vectors are extracted to measure useful information from the decomposed sub-images.

By observing handwritten document images, it can be seen that edge and variance measures change with stroke or word direction based on the characteristics of the handwriting. These two feature vectors are presented as **WDES (**Word Direction Based Edge Strength) and **WDVAR (**Word Direction Based Variance) are based on Word Direction based Grey level Co-occurrence Matrix (WDGLCM). Since degraded historical image strokes are frequently affected by noise, the **MG** (Mean-Gradient) feature vector is good at describing the effect of stroke noise. These three important features are considered in region classification.

5.1.3.1 Word Direction based Grey level Co-occurrence Matrix (WDGLCM) Feature Vectors

A Grey Level Co-occurrence Matrix (GLCM) contains information about the positions of pixels having similar grey level values. A co-occurrence matrix is a two-dimensional array, **G**, in which both the rows and the columns represent a set of possible image values.

A GLCM **Gd** is defined by first specifying a displacement vector **d**=(dx,dy) and counting all pairs of pixels separated by |**d**| having grey levels i and j.

The GLCM is defined by:

$$\mathbf{G_d[i,j] = n_{ij}} \qquad \text{Eq (5-1)}$$

where $\mathbf{n_{ij}}$ is the number of occurrences of the pixel values (i,j) lying at distance **d** in the image.

The co-occurrence matrix $\mathbf{P_d}$ has dimension $n \times n$, where n is the number of grey levels in the image.

Lam (1996) [49] described the grey-level gradient co-occurrence matrix (GLGCM). In GLGCM, $\mathbf{n_{ij}}$ was the number of occurrences of the pixel values (i,j) lying at distance **d** in the four directions: 0^o, 45^o, 90^o and 135^o.

For the three fragments of handwritten document images shown in Figure 5.3, it can be seen that there are three main directions in the slant of the handwritten words: left top to right bottom, top to bottom, and right top to left bottom. In counter-clockwise direction, they are at 45 (or 225) degrees, 0 (or 180) degrees, and 135 (or 315) degrees.

The three sub-images in Figure 5.3 are 64×64 pixel local areas. They are in 256-grey level TIFF image format. The comments in Figure 5.3, 'Direction = 1' means the direction of word stroke is G1 (refer to Figure 5.4), which is equal to 45 degrees; 'Direction = 4' means the direction of word stroke is G4, which is equal to 180 degrees; 'Direction = 3' means the direction of word stroke is G3, which is equal to 135 degrees.

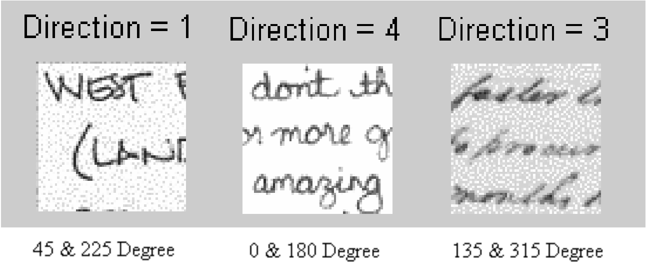


Figure 5.3 Three Main Word Directions

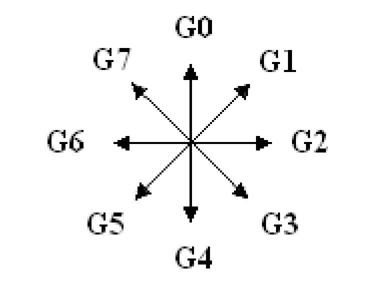


Figure 5.4 Eight Directions of Word

Figure 5.4 shows the 8 directions of words in counter-clockwise direction. Referring to Figure 5.4, the matrices of word directions can be defined as in Figure 5.5.

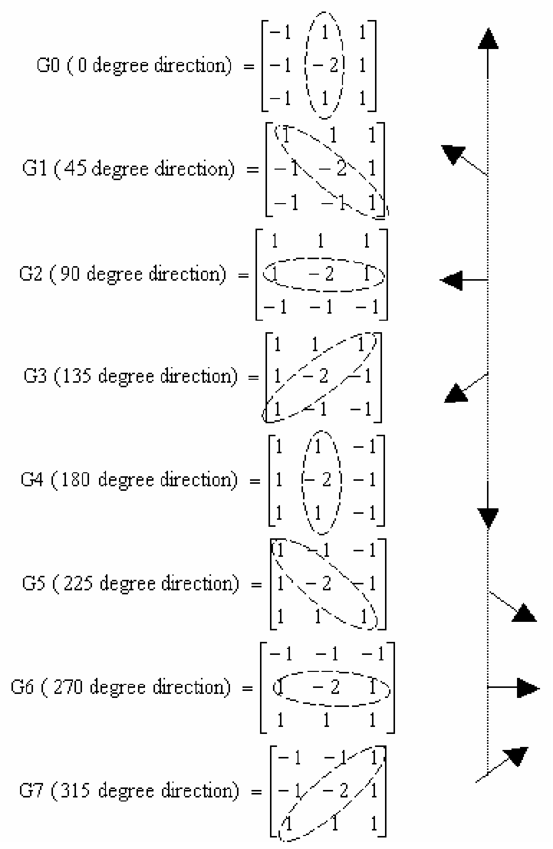

G0 (0 degree direction) = $\begin{bmatrix} -1 & 1 & 1 \\ -1 & -2 & 1 \\ -1 & 1 & 1 \end{bmatrix}$
G1 (45 degree direction) = $\begin{bmatrix} 1 & 1 & 1 \\ -1 & -2 & 1 \\ -1 & -1 & 1 \end{bmatrix}$
G2 (90 degree direction) = $\begin{bmatrix} 1 & 1 & 1 \\ 1 & -2 & 1 \\ -1 & -1 & -1 \end{bmatrix}$
G3 (135 degree direction) = $\begin{bmatrix} 1 & 1 & 1 \\ 1 & -2 & -1 \\ 1 & -1 & -1 \end{bmatrix}$
G4 (180 degree direction) = $\begin{bmatrix} 1 & 1 & -1 \\ 1 & -2 & -1 \\ 1 & 1 & -1 \end{bmatrix}$
G5 (225 degree direction) = $\begin{bmatrix} 1 & -1 & -1 \\ 1 & -2 & -1 \\ 1 & 1 & 1 \end{bmatrix}$
G6 (270 degree direction) = $\begin{bmatrix} -1 & -1 & -1 \\ 1 & -2 & 1 \\ 1 & 1 & 1 \end{bmatrix}$
G7 (315 degree direction) = $\begin{bmatrix} -1 & -1 & 1 \\ -1 & -2 & 1 \\ 1 & 1 & 1 \end{bmatrix}$

Figure 5.5 Direction Matrices

The stroke direction of the input handwritten document image is determined by using G0~ G7. The Word Direction-Based GLCM (WDGLCM) can be determined after the stroke direction has been calculated.

$$WDGLCM = occurrences\ of\ pixel\,(i, j)\,lying\ at\begin{bmatrix}(i, j-d) & \dots & (i, j) & \dots & (i, j+d)\end{bmatrix}, when\ G0;$$

$$WDGLCM = occurrences\ of\ pixel\,(i, j)\,lying\ at\begin{bmatrix} \dots & \dots & \dots & \dots & (i-d, j+d) \\ \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & (i, j) & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots \\ (i+d, j-d) & \dots & \dots & \dots & \dots \end{bmatrix}, when\ G1;$$

$$WDGLCM = occurrences\ of\ pixel\,(i, j)\,lying\ at\begin{bmatrix} (i-d, j) \\ \dots \\ (i, j) \\ \dots \\ (i+d, j) \end{bmatrix}, when\ G2;$$

$$WDGLCM = occurrences\ of\ pixel\,(i, j)\,lying\ at\begin{bmatrix} (i-d, j-d) & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & (i, j) & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & (i+d, j+d) \end{bmatrix}, when\ G3.$$

etc…..

where *d* is half of the typical stroke width of the word in the input document image.

The co-occurrence matrix WDGLCM has dimension n×n, where n is the number of grey levels in the image.

5.1.3.2 Word Direction Based Edge Strength (WDES)

Edge Strength can be defined based on the WD-GLCM vector as.

$$\begin{aligned} WDES &= \sqrt{\frac{1}{K^2}\sum_{i=1}^{K}\sum_{j=1}^{K}(i-j)^2 \times WDGLCM(i,j)}, \\ &= \sqrt{mean\left[(i-j)^2 \times WDGLCM(i,j)\right]} \end{aligned} \qquad \text{Eq (5-2)}$$

where *i* and *j* are the coordinates of WD-GLCM, and *K* is the number of grey levels in the input image.

Direction Based Edge Strength measures the grey level gradient differences in a certain direction determined by the conditions of the input image. It can provide more useful information for further analysis and works more effectively than simple edge strength based on GLCM.

5.1.3.3 Word Direction Based Variance (WDVAR)

Word direction based variance (WDVAR) measures the variability of grey value differences and hence coarseness of texture. A large value of variance indicates large local variation. Word Direction Based Variance is defined by WDGLCM as:

$$WDVAR^2 = \frac{1}{K-1}\sum_{i=1}^{K}\sum_{j=1}^{K}\left[WDGLCM(i,j)-\mu\right]^2 \qquad \text{Eq (5-3)}$$

where $\mu = \frac{1}{K^2}\sum_{i=1}^{K}\sum_{j=1}^{K} WDGLCM(i,j)$;

5.1.3.4 Mean-Gradient (MG)

Gradient is the change of image texture along some direction in the image, and the mean-gradient of the intensity image *I(x,y)* at location *(x,y)* which has been illustrated in Eq (4-2).

G_N is the mean-gradient value of the sub-block in direction N given by

$$G_N(x,y)=\sum_{x=0}^{i-1}\sum_{y=0}^{j-1}\frac{\left[\partial I(x,y)/\partial x_N,\ \partial I(x,y)/\partial y_N\right]}{x\times y}$$

Eq(5-4)

Where ∂x_N is the x distance on N direction, ∂y_N is the y distance on N direction.

Mean gradient (MG) is sensitive to small variance between strokes; it can be used to detect faint strokes between heavy strokes.

5.1.4 Sub-Block Classification

The three feature vectors described in Section 3.3.1 were used to test the local regions and classify them into three types: **background**, **heavy strokes** or **faint strokes**. Typical examples of these three types of region are shown in Figure 5.6.

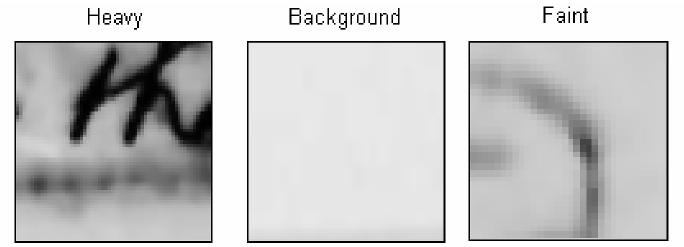


Figure 5.6 Examples of Sub-regions containing heavy strokes, background (no strokes) and faint strokes.

The background of a document does not contain any useful information content. A background area typically has lower values of edge strength and variance. A noise-free background also has a small mean-gradient value.

Heavy stroke areas have strong edge strength, variance and mean-gradient value. Faint stroke areas contain weak strokes, which are very difficult to detect from the background. This kind of area typically has a medium value of edge strength and mean-gradient but less variance.

5.1.5 Applying the Thresholding Method

Different threshold methods are applied for the above three classes of sub-images. The six methods described above (improved Niblack's method [110], Yanowitz & Bruckstein's method [103], Bernsen's method [4], ETM [25], Otsu's method [64] and QIR [83]) cannot provide ideal results for degraded historical handwritten images, especially for regions in the faint strokes class. Bernsen's method, the improved Niblack's algorithm and Yanowitz & Bruckstein's method are the best of the other existing well-known thresholding methods described in [91]. These three thresholding methods are chosen for the heavy and faint stroke classes.

For background areas, all pixels are simply set to white (greyscale value 255) because there is no useful information need to be thresholded.

5.1.5.1 Faint Strokes Class

The sub-image in the faint strokes class contains lower edge strength because the pen is lightly skating over the region so that less ink is deposited on the paper. The higher mean-gradient value will be detected for regions if there is more noise, thus noise removal is required. The variance will be low because the variation of the region is low.

One major problem for faint stroke detection is noise. Noise always affects the faint strokes so that the faint strokes are very difficult to detect. Another problem is faint strokes have very low variance which means many algorithms cannot work well on this kind of image area.

Noise removal and enhancement for the *faint stroke class* are needed before the threshold method is applied. A Wiener filter was first applied for noise removal. After that, the enhancement can be divided into two steps.

1). Use a 3×3 window to enhance the image by finding the maximum and minimum grey value in the window using Eq (5-5) and Eq (5-6):

$$Mini = \min \text{ (elements in the window)} \qquad \text{Eq (5-5)}$$

$$Maxi = \max \text{ (elements in the window)} \qquad \text{Eq (5-6)}$$

2). Compare the value of *Pixel* - *Mini* and *Maxi* – *Pixel*, where *Pixel* stands for pixel-value. If the former is larger, the *Pixel* is closer to the highest grey value than the lowest value in this window; hence the value of *Pixel* is set to the highest grey value (*Pixel* = *Maxi*). If the former is smaller, then the value of *Pixel* is set to the lowest grey value (*Pixel* = *Mini*).

Yanowitz & Bruckstein’s method works well on retaining detailed information of handwriting and hence it can retain more information of faint strokes. It was applied to the faint stroke class.

5.1.5.2 Heavy Strokes Class

There are two sub-classes in the *heavy stroke class*. One contains heavy strokes only; the other contains some faint connection strokes alongside heavy strokes.

Bernsen’s method works well because those regions have high contrast, which can work well for high contrast heavy strokes. It was applied to the heavy strokes region. The contrast threshold value for the experiment of Bernsen’s method was set to 180. Practically, the range of contrast value for heavy strokes regions is 190 ~ 220 (for a 8-bit greyscale image).

The improved Niblack method is sensitive to edge information, and is able to clearly maintain the faint strokes, which are connected to heavy strokes. It was applied to the heavy strokes region with some faint connection strokes.

5.2 Improved Decompose Algorithm

The quality of the thresholding result when separating foreground from background is decisive for subsequent analysis of the document content. It requires retention of the full information content on a clear white background. In some applications, such as forensic document analysis, or scholastic analysis of the writing style, we are interested in the detailed greyscale or colour variations of the pen strokes or printing.

Although the proposed decompose algorithm is effective on a wide range of degraded historical images, Yanowitz & Bruckstein's method fails to clearly keep faint loops inside the faint descender strokes. The proposed weighted mean-gradient thresholding method can provide better performance at keeping these faint loops, but it will over-threshold some strong stroke regions. This means it cannot work well for the various characteristic degraded images based on unchangeable weighted value. The improved decompose algorithm combines the decompose structure and the mean-gradient thresholding method with different weighted values for different sub-region classes.

The second version is an improved decompose algorithm which recursively decomposes a document image into sub-regions until appropriate weighted values can be selected to determine an appropriate single stage thresholding method for each region.

The single stage thresholding method, which is a novel mean-gradient based method, was described in Chapter 4.

The improved decompose algorithm is also a local adaptive analysis method, which uses local feature vectors to find the best approach for thresholding a local area. Compared to the original decompose algorithm, the appropriate weighted values are selected automatically for mean-gradient thresholding method (base on the specific types of document image regions under investigation) instead of choosing different thresholding techniques for different sub-regions. The original image is recursively broken down into sub-regions using quad-tree decomposition until an appropriate weighted mean-gradient thresholding method can be applied to each sub-region.

The new improved decompose algorithm analyzes the feature information of the local regions with different sizes, and applies a new mean-gradient based threshold method with **appropriate weighted values** to obtain the best result, as illustrated in Figure 5.6. The grey block of Figure 5.7 is the improved part (Appropriate Weighted Mean-gradient Threshold Method) in the Improved Decompose Algorithm.

5.2.1 Faint Handwritten Image

a). Enhancement. Enhancement method is the same as the one used in the Decompose Algorithm

b). Mean-Gradient Thresholding. A new weighted method based on mean-gradient direction is proposed for thresholding faint strokes.

Handwritten English or Western-style scripts normally contain strokes written in several directions. In this method a different matrix is used to detect different meangradients in eight different directions (G0 ~ G7) as described in Section 3.3 and shown in Figure 5.4.

The matrices G0 ~ G7 are convolved with the sub-block to discover the maximum mean-gradient value. The sub-block direction is the one in G0 ~ G7 that produces the largest mean-gradient value. For example, consider the sub-block shown in Figure 5.5. The largest mean gradient value exists at the convolution of G3 with the sub-block, indicating that the direction of strokes in Figure 5.6 is 135 degrees.

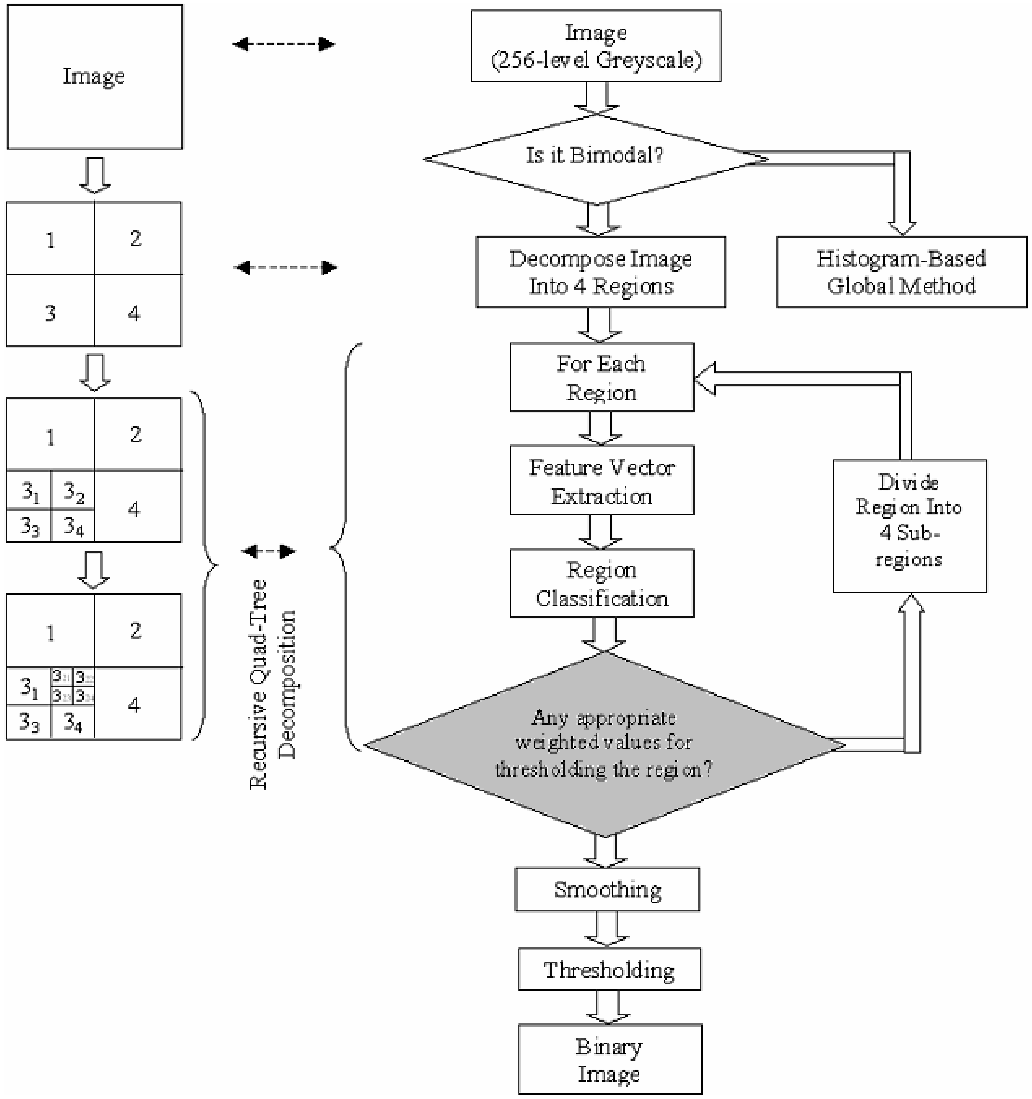


Image
1
2
3
4
1
2
3_1
3_2
3_3
3_4
4
1
2
3_1
3_3
3_4
4
Recursive Quad-Tree Decomposition
Image
(256-level Greyscale)
Is it Bimodal?
Decompose Image Into 4 Regions
Histogram-Based Global Method
For Each Region
Feature Vector Extraction
Region Classification
Divide Region Into 4 Sub-regions
Any appropriate weighted values for thresholding the region?
Smoothing
Thresholding
Binary Image

Figure 5.7 Flowchart of Improved Decompose Algorithm

The mean directions of the area's strokes are calculated as *N*, and then the mean-gradient method is applied using *N* to obtain the binary output.

$$T(x, y) = wM(x, y) + kG_N(x, y) \qquad \text{Eq (5-7)}$$

where *T* is the threshold value, *M* is the mean value of sub-block, *w* and *k* are weighted value, G_N is the mean-gradient value of the sub-block at direction *N* given by Eq (5-4).

5.2.2 Heavy Handwritten Image

There are two sub-classes in the heavy stroke class. One contains heavy strokes only; the other contains some faint connected strokes alongside heavy strokes.

The proposed weighted method is applied on these two sub-classes with different weighted value *w* and *k*. Practically, the formula for thresholding with a different weighted value for specific cases is:

$$T(x, y) = wM(x, y) + kG_N(x, y) \qquad \textit{where} \begin{cases} \text{if 'background'}, & w = 0, k = 0; \\ \text{if 'faint'}, & w = 0.5, k = -1.1; \\ \text{if 'heavyOnly'}, & w = 0.7, k = -0.8; \\ \text{if 'heavyWithFaint'}, & w = 0.7, k = -1.1. \end{cases}$$

Eq (5-8)

The values of *w* and *k* in Eq (5-8) are obtained from experimental results in the training of the Mean-Gradient threshold method.

To avoid blocking effects at the boundary of the sub regions, a smoothing matrix $A = \frac{1}{25}\begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix}$ was convolved with the area, which is a 5 by 5 window centred on the edges of each region before the mean-gradient based thresholding method is applied.

5.3 Experimental Results

The decompose algorithm was trained using 10 historical document images obtained from the Library of Congress, which contained considerable background noise or variation in contrast and illumination. (The images are shown in Appendix 3 and the attached CD).

The first version of the decompose structure is highly effective for images which contain different conditions at different locations; however, it fails on faint loops in some cases because there is no existing good threshold method for faint stroke detection. In order to improve the algorithm it is necessary to retain faint loops in descender strokes. The proposed weighted mean-gradient method, which can retain faint strokes by extracting the direction information from the image, is used in the Improved Decompose Algorithm. The new method has superior performance on all test images compared to the other six methods evaluated. Some results of the Improved Decompose Algorithm and the other six methods when applied to the original image of Figure 1.2 are shown in Figure 5.11. Other detailed images can be viewed on the attached CD.

5.3.1 Experimental Results of Decompose Algorithm

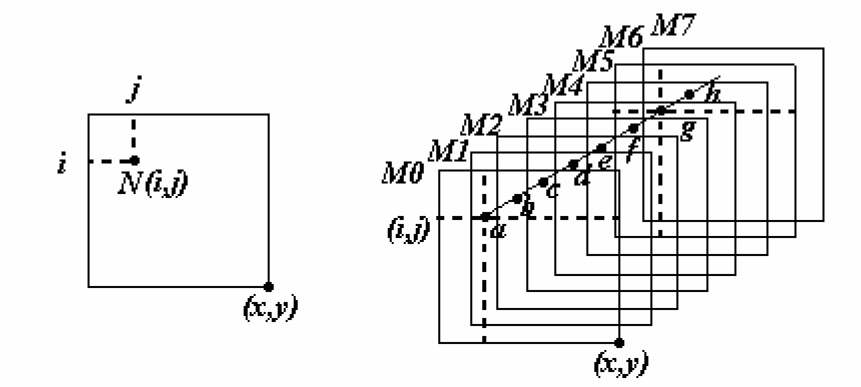
There are two important steps in the proposed Decompose Algorithm:

1. Determination of stroke direction

Direction of stroke slant in the words in the handwritten historical document images is an important factor to WDES and WDVAR features. The mean direction calculated from each region affects the value of WDES and WDVAR directly.

2. Classification of the local region

In order to apply the best appropriate threshold method on different regions of degraded historical image, the regions are classified into three classes: Background Region, Faint Region and Heavy Region.



```
for $i = 1:1:x$,
    for $j = 1:1:y$,
        $TempltMatrix(i,j) =$
            $Maximum\{a(i,j)\quad b(i,j)\quad c(i,j)\quad d(i,j)\quad e(i,j)\quad f(i,j)\quad g(i,j)\quad h(i,j)\}$;
        if $TempltMatrix(i,j) == a(i,j)$, then $N(i,j) = 0$;
        if $TempltMatrix(i,j) == b(i,j)$, then $N(i,j) = 1$;
        if $TempltMatrix(i,j) == c(i,j)$, then $N(i,j) = 2$;
        if $TempltMatrix(i,j) == d(i,j)$, then $N(i,j) = 3$;
        if $TempltMatrix(i,j) == e(i,j)$, then $N(i,j) = 4$;
        if $TempltMatrix(i,j) == f(i,j)$, then $N(i,j) = 5$;
        if $TempltMatrix(i,j) == g(i,j)$, then $N(i,j) = 6$;
        if $TempltMatrix(i,j) == h(i,j)$, then $N(i,j) = 7$;
    end
end
N is point direction matrix.
```

Figure 5.8 Tracing of Point Direction

5.3.1.1 Stroke Direction of Image

The Stroke Direction of an image is measured by convolving the eight Direction Matrices (shown in Figure 5.6: G0 ~ G7) with the local area of the image one-by-one to obtain matrices M0~M7 respectively. Matrices M0 ~M7 are the same size matrices and contain grey-scale values with direction information on each pixel (point). The tracing of the direction at each pixel (point) in pseudo code is shown in Figure 5.8.

Practically, in the experiment, 64×64 is the best size to measure the stroke direction. The result is not accurate if the size of the local area is too small, and too many calculations will be required to get an accurate result if the local area size is too big.

5.3.1.2 Local Region Classification

In order to apply the best appropriate threshold method to each region, classification of each local region is needed. Some experimental results from the 300 training images are shown in Table 5.1 and Figure 5.9. The example in Figure 5.9 is using region size 64×64. In practice, the local region can be larger than 64×64, it depends on the quad tree decomposition results. Here, 64×64 pixel local regions are used for easier visual comparison.

The left part of Figure 5.9 shows two background class images, which contain only noise and no stroke information. It is observed that background areas exhibit low edge strength and low mean-gradient value, but may have high variance, which is usually produced by noise.

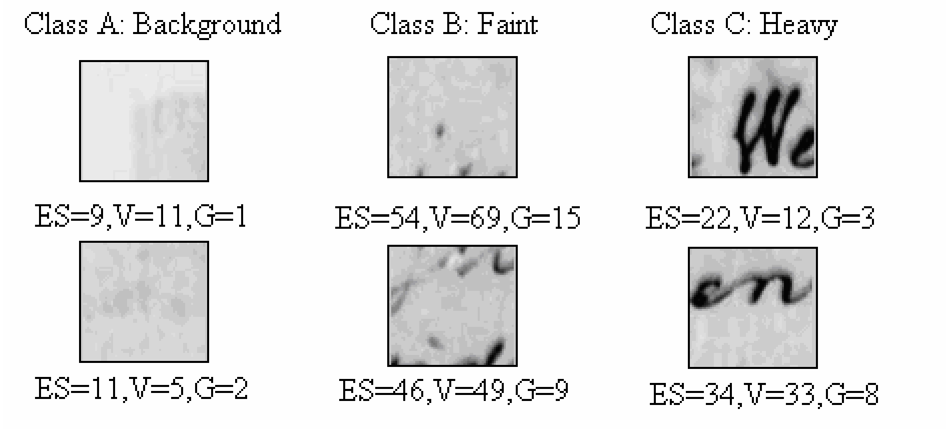


Figure 5.9 Local Region Classification & Feature Extraction

Table 5.1 Experimental Result of Area Classification from Training Image Group

Class Name / Feature Name	Background	Faint Strokes	Heavy Strokes	
			With some faint strokes	Only heavy strokes
Edge Strength	$1 \le ES \le 13$	$14 \le ES \le 40$	$ES \ge 41$	
Variance	$1 \le V \le 30$	$10 \le V \le 44$	$V \ge 45$	
Mean-Gradient	$1 \le G \le 2$	$3 \le G \le 10$	$3 \le G \le 10$	$G \ge 10$

Note: ES: Edge Strength, V: Variance; G: Mean-Gradient.

The middle part of Figure 5.9 shows two faint stroke areas, which include noise and faint handwritten strokes. This area contains stronger edges, variance and higher mean-gradient value than a background area.

The right part of Figure 5.9 shows two areas that contain heavy strokes. These regions have strong edges, strong variance and high mean-gradient value. The upper image in the heavy class contains only heavy strokes. Compared to the upper image, the lower image contains not only heavy strokes but also a light stroke connecting ‘e’ and ‘n’. This faint stroke results in a lower mean-gradient value.

5.3.1.3 Decompose Algorithm

From an aesthetic and subjective point of view, the decompose structure performs better than other single-stage local methods. It detects feature vectors of different areas and then applies appropriate methods to avoid losing important useful information.

The proposed method was evaluated using these six thresholding algorithms on 10 historical images selected from the Library of Congress where considerable background noise or variation in contrast and illumination exists with varying resolution, sizes, and contrast.

The standard measures, *recall* (**Definition 4**) and *precision* (**Definition 5)** were used to evaluate the performance of the proposed methods. For recall value calculation, the number of words, which are correctly separated from the background and accord with the following tight requirements, were counted. The correctly detected words were then divided by the total number of handwritten words in the original images. For precision calculation, the correctly detected words were divided by the total detected words in the original images.

Figure 5.10 shows the recall and precision values of the seven threshold methods, and the average Recall and precision values are presented in percentage of the content of each image. From the table, it is apparent that the proposed *Decompose Thresholding method* produced significantly better recall and precision results than the other individual six methods.

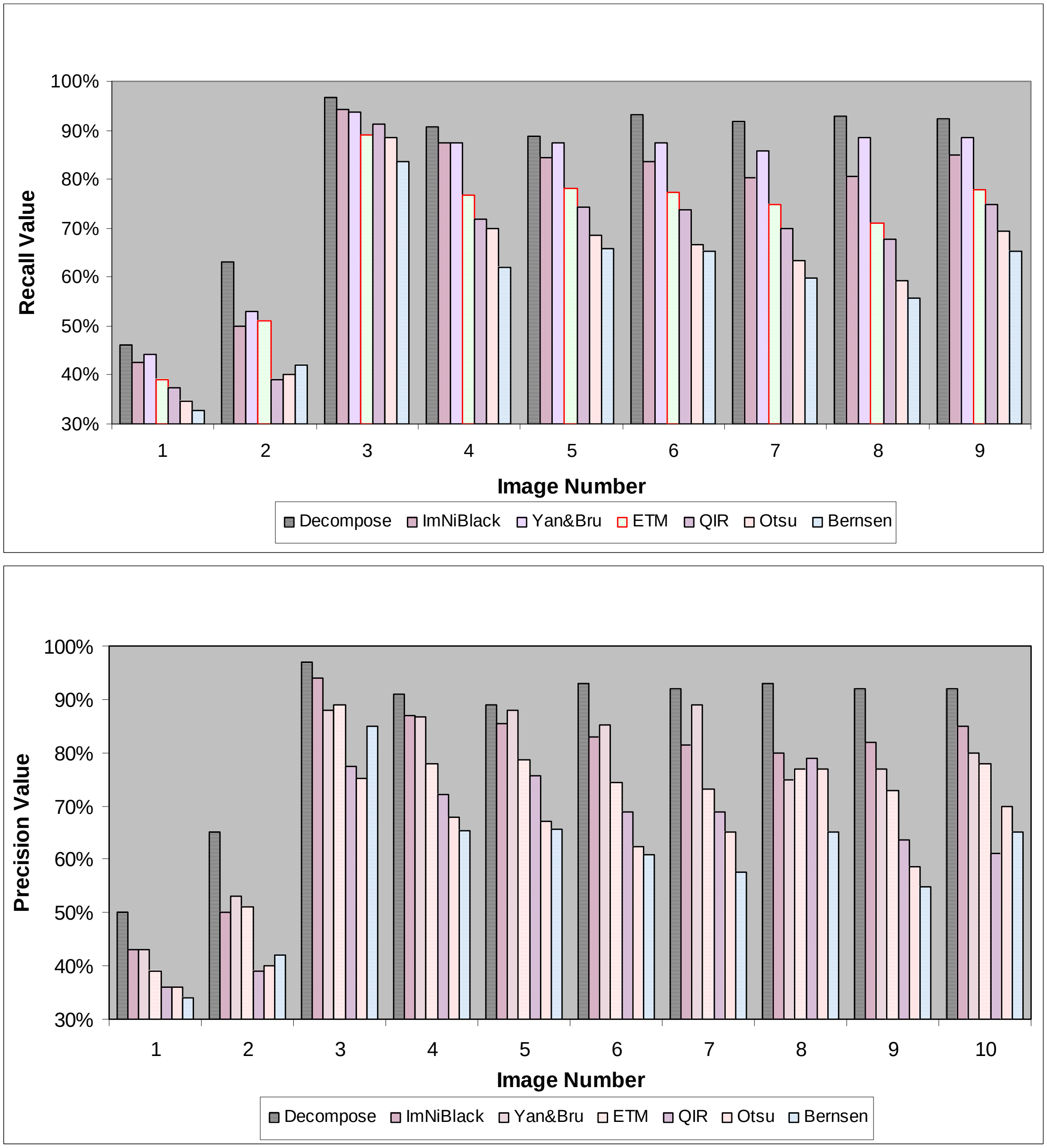


Figure 5.10 Evaluation of the 7 Algorithms by Recall and Precision Value

Decompose: Proposed Decompose Threhsolding Method
ImNiblack: Improved Niblack's Method
Yan&Bru: Yanowitz/Bruckstein's Algorithm
ETM: Eikvil/Taxt/Moen's Method
QIR: QIR Algorithm
Otsu: Otsu's Algorithm
Bernsen: Bernsen's Method

Of these six methods, Bernsen's method works well on clear background and high contrast historical images, but it is a contrast-based method so it is sensitive to the noise in the images. ETM's method uses a manual value to determine the difference between two windows. It works well for both faint and heavy handwriting, but failed when there was a noisy background. Yanowitz's method can retain very detailed strokes but still retains useless noise points. The improved Niblack technique can retain detailed stroke information but is sensitive to noise. QIR and Otsu's technique only work well on bimodal histogram images.

From an aesthetic and subjective point of view, the *Decompose-Threshold Approach* is better than other local threshold methods. The other local and global threshold methods apply the same process to the whole input images, but ignore the fact that degraded historical images always contain different characteristics at different regions. The Decompose-Threshold Approach detects feature vectors of different sub-areas and then applies appropriate methods on different local regions to avoid losing important useful information.

The decompose-threshold structure is highly effective for the images, which contain different conditions at different locations.

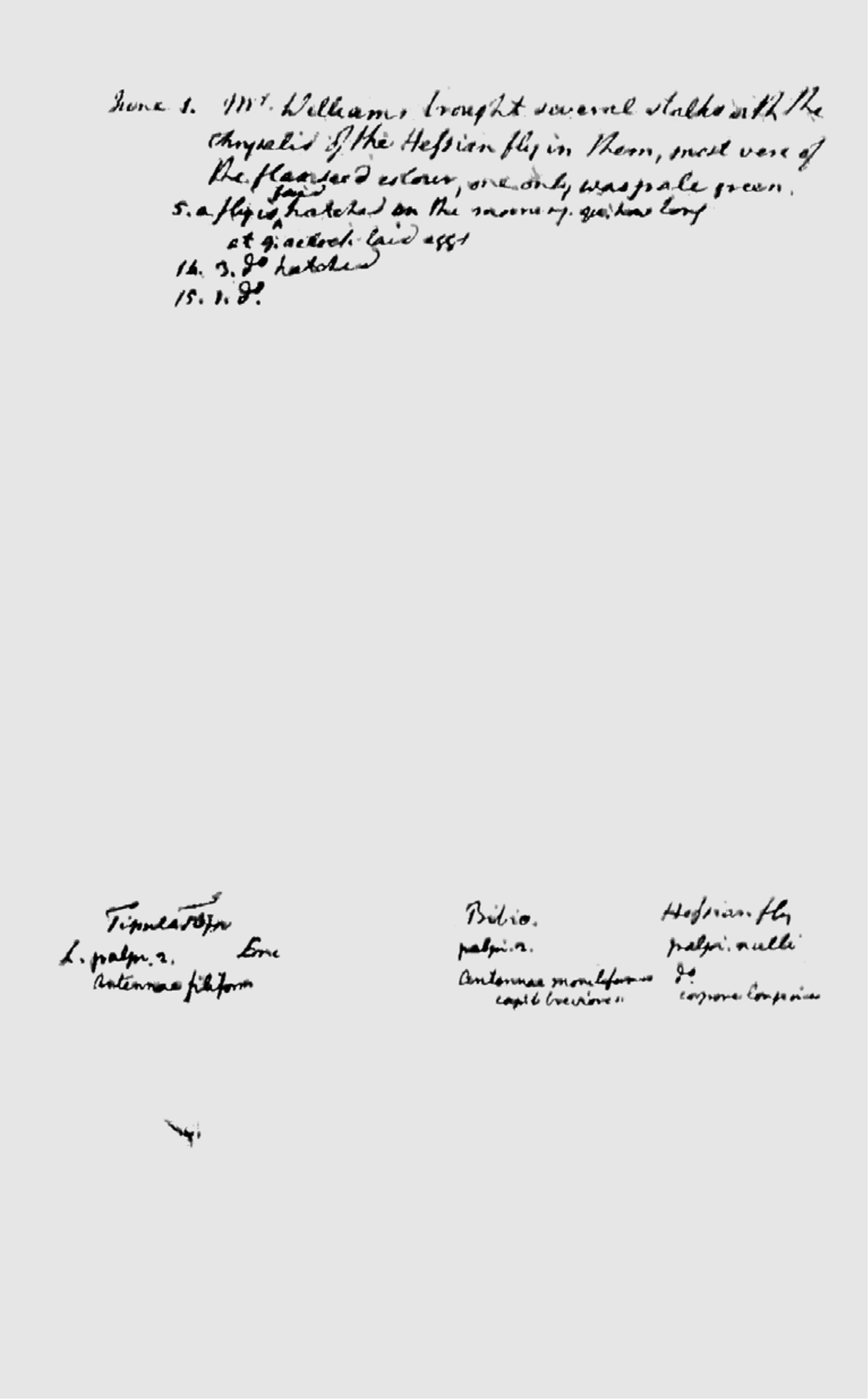
5.3.2 Experimental Results of the Improved Decompose Algorithm

Of these six methods, Bernsen's method works well on clear background and high contrast historical images, but it is a contrast-based method and so is sensitive to the noise in the images. ETM's method uses a fixed value to determine the difference between two windows. It works well for both faint and heavy handwriting, but fails when there is a noisy background. Yanowitz's method can retain very detailed strokes but still includes much useless noise. The improved Niblack technique can retain detail of strokes but is sensitive to noise. QIR and Otsu's technique only work well on bimodal histogram images and so do not perform well on these degraded document images.

The same 300 historical images were selected from the Library of Congress on-line database to train the algorithms. The images were chosen to have varying resolutions, sizes, and contrast to ensure correct comparison of performance between the algorithms. In these selected images, some historical images still have acceptable quality even through they were created many years ago, but many of them contain considerable background noise or variation in contrast and illumination. The scanned images were characterized by high resolution with varying contrast of the handwriting.

The Improved decompose algorithm was evaluated and compared with these six thresholding algorithms on a further 300 historical images selected from the Library of Congress. The standard measure, recall and precision [41], were again used to quantitatively compare the relative performance of the proposed methods at retaining the word information in the documents.

Table 5.2 shows the recall and precision values of the seven threshold methods. In the table, the 300 example historical images are classified into six groups, with 50 images per group. The grouping was random. Average recall and precision values are presented as mean values using 50 images per group for each of the seven thresholding methods. The table also shows the average recall value of the 300 example images. From Table 5.2 and Figure 5.12, it is apparent that the improved decompose algorithm produced significantly better recall and precision results than the other six methods.



June 1. Mr. Williams brought several stalks with the
chrysalis of the Hessian fly in them, most were of
the flaxseed colour, one only was pale green.
5. a fly is ^female^ hatched on the morning. [illegible] long
at 9 o'clock laid eggs
14. 3. do hatched
15. 1. do.

Tipula [illegible]
L. palpi 2. Linn
Antennae filiform

Bibio.
palpi 2.
Antennae moniliformes
capite breviores

Hessian fly
palpi nulli
do.
corpore compressa

Figure 5.11(a) Result of Improved Decompose Algorithm

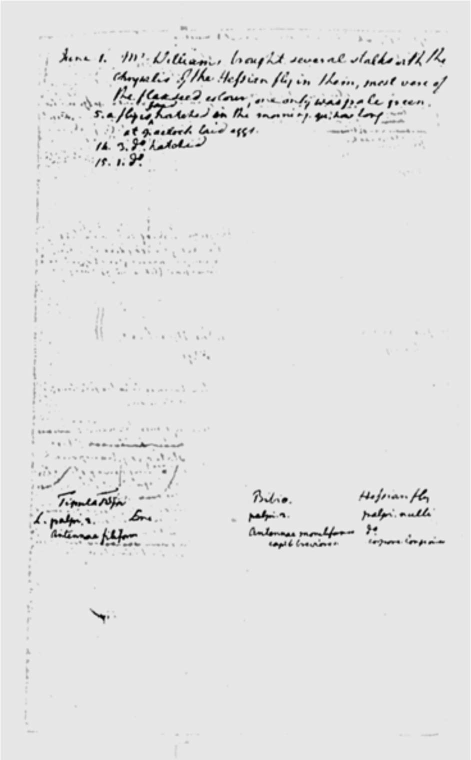


Figure 5.11(b). ETM's Algorithm

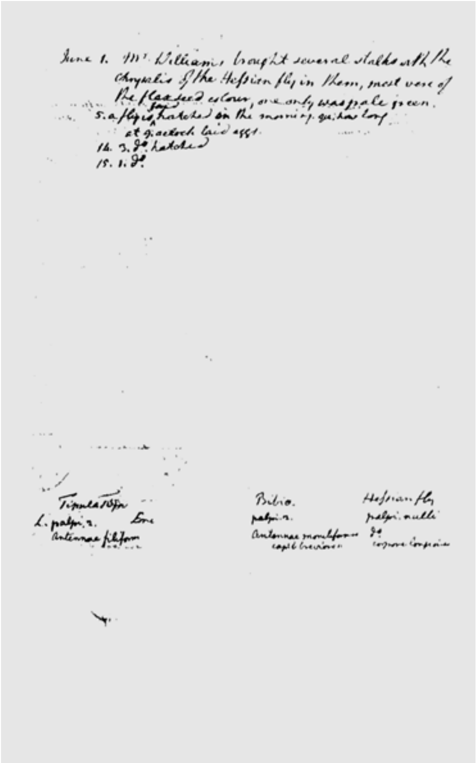


Figure 5.11(c) Improved Niblack's Technique

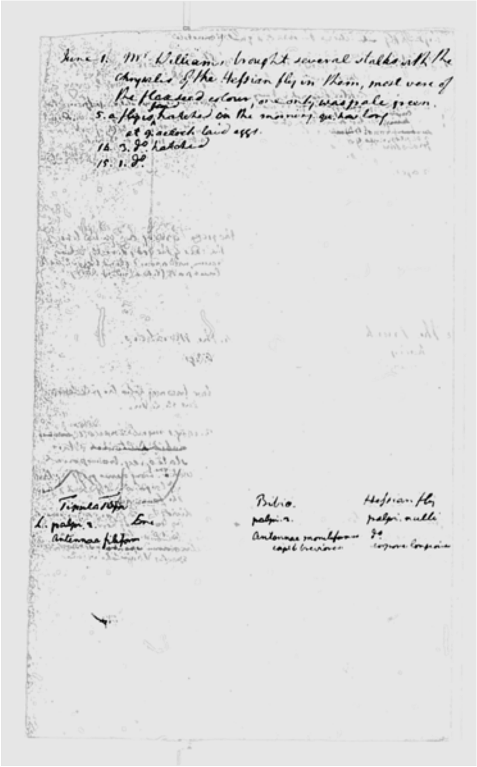


Figure 5.11(d). Bernsen's Method

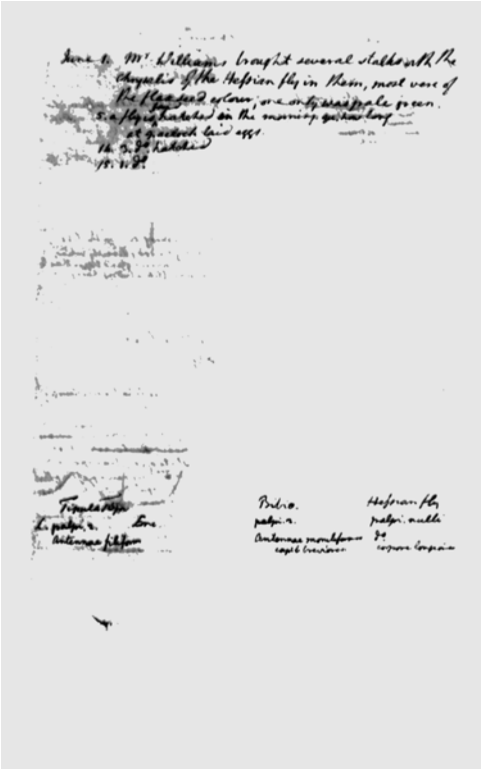


Figure 5.11(e) Otsu's Technique

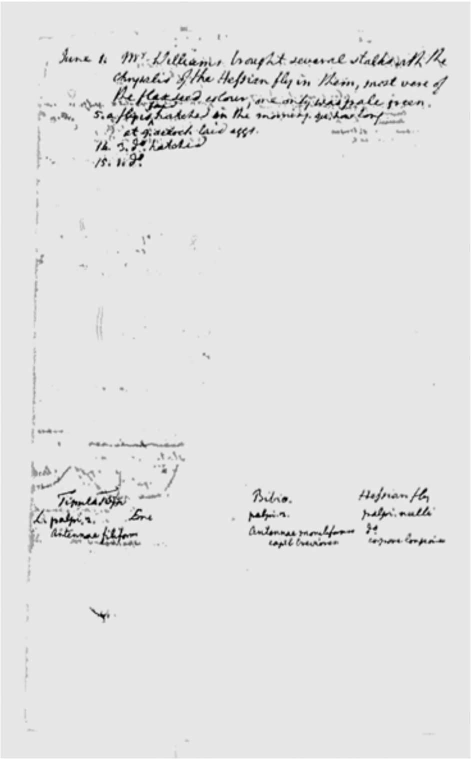


Figure 5.11(f) Yanowitz's Algorithm

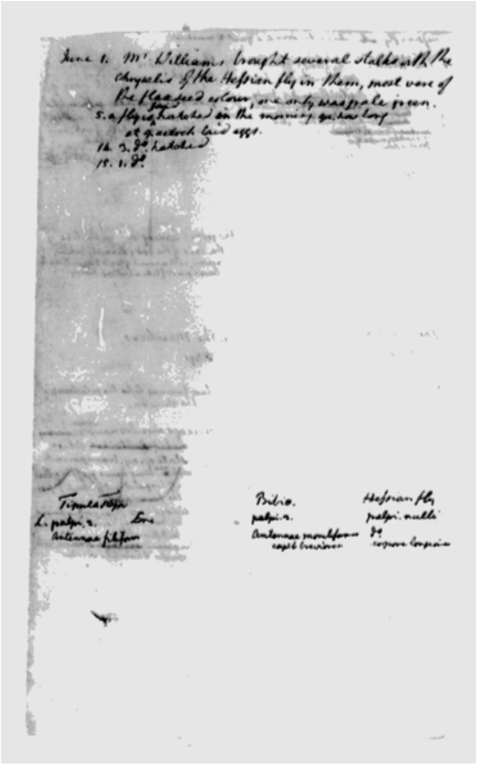
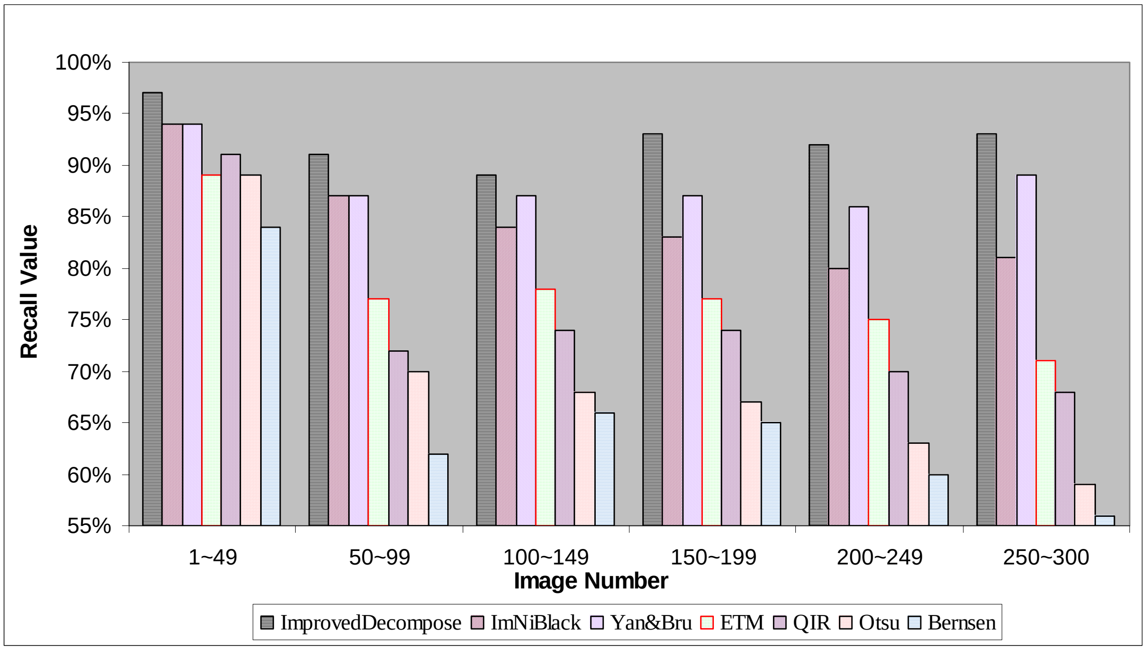


Figure 5.11(g) QIR Technique

Table 5.2 Comparison of Precision & Recall for the Seven Algorithms

Image No / Algorithm	0~49		50~99		100~149		150~199		200~249		250~299	
	P *(%)*	**R *(%)***	**P *(%)***	**R *(%)***	**P *(%)***	**R *(%)***	**P *(%)***	**R *(%)***	**P *(%)***	**R *(%)***	**P *(%)***	**R *(%)***
Bersen	86	84	63	62	66	66	65	65	61	60	56	56
ETM	89	89	79	77	78	78	77	77	75	75	71	71
Im-Niblack	94	94	87	87	84	84	83	83	80	80	81	81
Improved Decompose	97	97	91	91	89	89	93	93	92	92	93	93
Otsu	89	89	70	70	68	68	67	67	65	63	59	59
QIR	91	91	72	72	74	74	74	74	71	70	68	68
Yanowitz	94	94	87	87	87	87	87	87	87	86	89	89



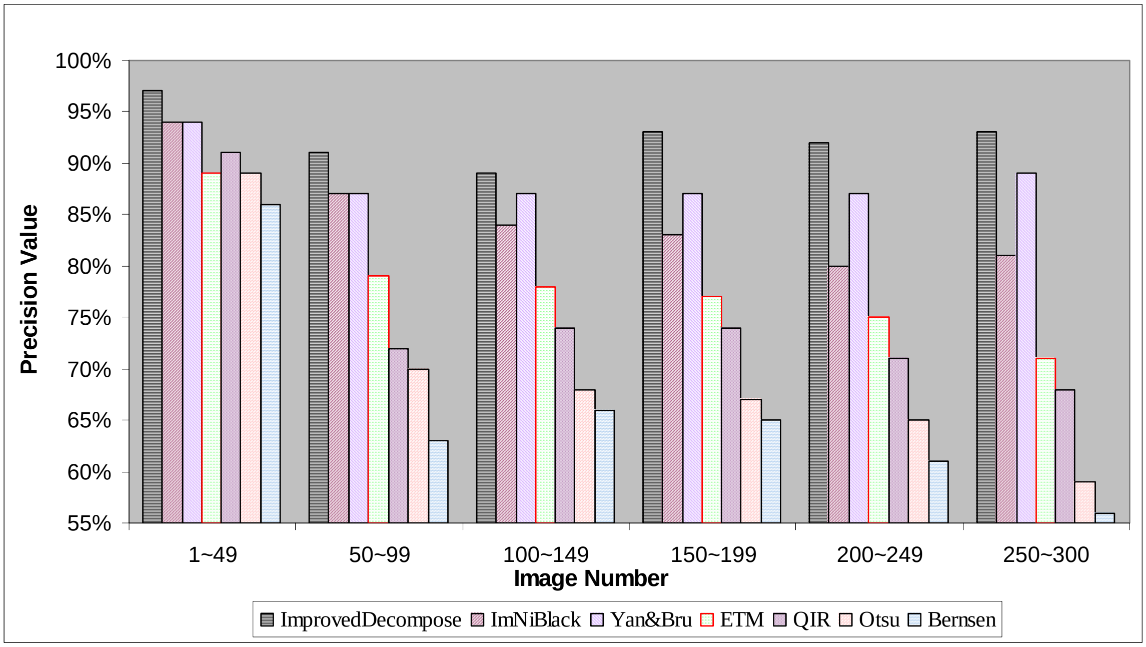


Figure 5.12 Evaluation of 7 Algorithms in 10 groups (30 images/group)

5.4 Summary

From oberservations during the experiments reported in this chapter, degraded historical handwritten images normally contain the following characteristic: 1) Varying contrast; 2) Varying stroke quality; 3) Many marks or blotches which do not contain any information. A number of techniques have previously been proposed for thresholding document images. However, none of them can provide ideal results for degraded historical handwritten document images. Because of varying contrast and noise conditions across the image, applying the same processing to the whole image is not flexible and will not produce good results.

The effectiveness of the two algorithms is assessed on 300 'difficult' document images extracted from the Library of Congress on-line database of historical documents. Six thresholding algorithms were used for comparative evaluation comprising four local thresholding methods: Improved Niblack's method [110], Yanowitz & Bruckstein's Method [103], Bernsen's method [4], ETM [25] and two global methods: Otsu's method [64] and QIR [83].

The improved Niblack method [110] works well on Heavy handwriting image blocks, because it is sensitive to edge information. Bernsen's method [4] also works well on Heavy handwriting image blocks because a heavy image has high contrast. Compared to the other five methods, it also works well on faint handwriting images. Yanowitz & Bruckstein's [103] method works well on various kinds of document image, but is not good when there is a noisy background, especially in images where there is double-sided noise (where the ink has seeped through the paper from the other side). ETM [25] works well for both faint and heavy handwriting images. Otsu's method [64] can achieve good performance with simple documents where the background and foreground are clearly distinct in the histogram. However, Otsu’s algorithm is very time-consuming for image binarization because of its inefficient formulation of the between-class variance, and the performance varies with data sets. QIR [83] works well for bimodal histogram images. Bersen’s, Yanowitz’s, Niblack’s, Otsu’s, QIR and the ETM thresholding methods have been evaluated and compared in [50] and [91].

Degraded historical images often exhibit varying image qualities. Satisfactory thresholding results can rarely be obtained if only a single global or local method is applied to the whole image.

The decompose algorithm recursively breaks down an image into sub-regions using quad-tree decomposition and extracts local features from each sub-region until an appropriate thresholding method can be applied to each sub-region.

The improved decompose algorithm is demonstrated as effective at improving the result. The three feature vectors proposed in this chapter can accurately classify the different regions so that the appropriate weighted value can be applied to the mean-gradient based threshold method. The new proposed weighted mean-gradient based threshold method is based on the local mean gradient value for the word direction. It can retain faint strokes by the direction information from the image.

Chapter 6

Independent Component Analysis based Segmentation

Many document images contain florid handwriting, which frequently exhibits extravagant loops in ascenders, descenders and upper case letters. These often result in touching or overlapping of words on adjacent lines. Separating the lines and words is difficult as the overlapping words on adjacent lines are often degraded to such an extent they are difficult for a human to decipher due to the damage caused by poor storage and handling over several hundred years. The segmentation of touching or overlapping words on adjacent lines is an important stage in the processing of historical cursively written documents.

In this chapter, an Independent Component Analysis (ICA)-based segmentation algorithm is proposed, which can be used effectively on degraded document images containing many different kinds of overlapping and touching words in adjacent lines. This chapter first describes Independent Component Analysis and then goes on to describe the ICA based Segmentation Algorithm Approach proposed in this work. Experimental results are presented at the end of the chapter along with a summary.

The mean-gradient thresholding method described in this chapter was published in *Proc. 8th Int. Conf. on Document Analysis and Recognition*, Seoul, Korea, Vol.2, pp 680-684, 2005 and is under review by the journal *Pattern Recognition Letters*.

6.1 Independent Component Analysis

Independent component analysis (ICA) is a statistical and computational technique for revealing hidden factors that underlie sets of random variables, measurements, or signals, and is widely used in image processing [39]. It is noted [39] that ICA defines a generative model for the observed multivariate data, which is typically given as a large database of samples. The data variables of the model are assumed to be linear or non-linear mixtures of some unknown latent variables, and the mixing system is also unknown. The latent variables are assumed to be non-Gaussian and mutually independent of each other and referred to as the independent components of the observed data. These independent components can be found by ICA.

6.1.1 Definition of Independent Component Analysis (ICA)

ICA can be seen as an extension to principal component analysis and factor analysis. ICA is a much more powerful technique, however, capable of finding the underlying factors or sources when the classic methods fail completely. The definition of ICA from Hyvarinen's survey [40] made in 1999 is as follows:

❖ **Definition 9** - ICA of a random vector x consists of estimating the following generative model for the data:

$$x=As$$

where the latent variables (components) s_i *in the vectors* $s=(s_1,\ldots,s_n)^T$ *are assumed independent. The matrix A is a constant* $m \times n$ *'mixing' matrix.*

This is the simplest and widest used definition in most research on ICA. There are also other ICA definitions which can be found in the literature [17], [42].

6.1.2 Identifiability of the ICA Model

The ICA was chosen in this algorithm based on the three identification points of the ICA model, as described by [40]:

1. All the independent components s_i, with the possible exception of one component, must be non-Gaussian;
2. The number of observed linear mixtures N must be at least as large as the number of independent components M, i.e., $N \geq M$;
3. The matrix A must be of full column rank.

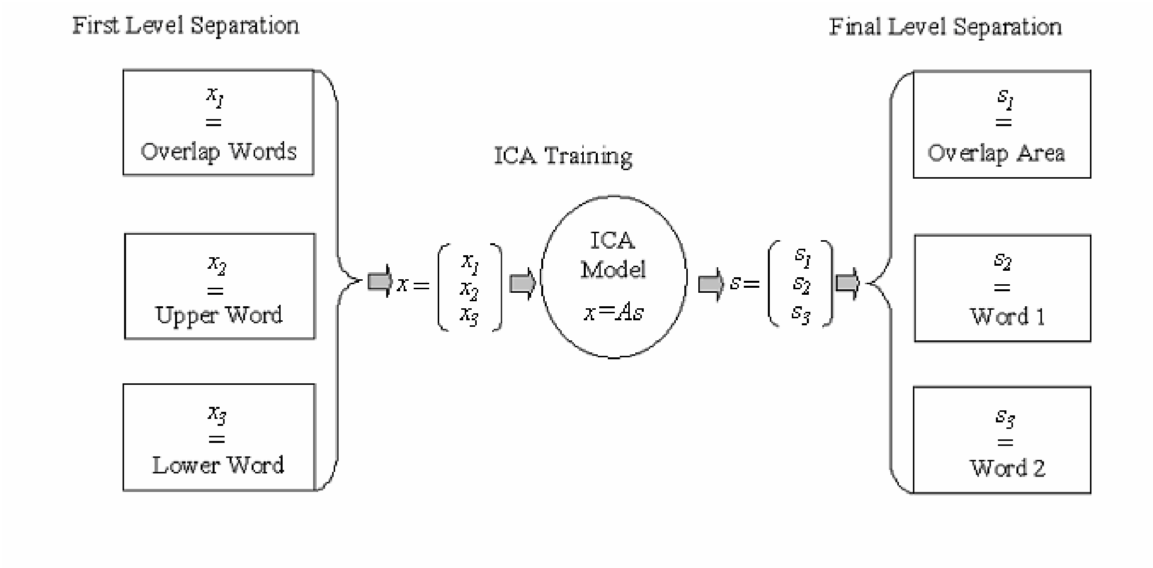


Figure 6.1 ICA-based segmentation approach

Imagine that Figure 6.1 is a perceptual system called the ICA-based segmentation approach, which is exposed to a series of small rectangular size image patches, which contain overlapping handwritten word components from a larger image.

Each image patch is represented by the vector $x_{1,}$ x_2 and x_3.

$x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$, where x_1 is the original overlapped word; x_2 is the upper class word; and x_3 is the lower class word.

The independent component matrix after applying the ICA-based segmentation approach is, $s = \begin{bmatrix} s_1 \\ s_2 \\ s_3 \end{bmatrix}$, where s_1 is the separated independent component one: overlapped words image; s_2 is the separated independent component two: upper class word image; s_3 is the separated independent component three: lower class word image.

The independent components are based on handwriting image components, which are represented by a non-linear matrix. This is consistent with point 1 above which states that all independent components must be "non-linear". Point 2 above declares that the number of observed linear mixtures N in x must be larger than the independent component M. From Figure 6.1, the number of independent components M is equal to three, and N, the column number of x is equal to the size of image patches containing overlapped words. The size of the image patches is much larger than three because the original image patch contains two words. Besides, the matrix A is full column rank and is consistent with point 3.

In conclusion, ICA can be used in overlapped word segmentation after studying typical historical document images that contain overlapped words.

6.2 ICA based Segmentation Algorithm Approach

Whilst many existing separation techniques have proven effective at segmenting words correctly if the handwritten text lines are not overlapping or touching, none has been shown able to produce consistently good results on the wide range of document images containing touching or overlapping handwritten strokes.

The new segmentation algorithm called the ICA (Independent Component Analysis) Segmentation Algorithm is described for segmenting the overlapping/touching loop strokes observed in historical document images. The ICA-based Segmentation Algorithm focuses on 8-bit grey-scale images of historical handwritten documents, which contain overlapping or touching words on adjacent lines (as shown in Figure 6.3 and Figure 6.6).

The first step of the novel ICA-based segmentation algorithm is the first level separation of the overlap words, which produces a blind source matrix (vector matrix) based on the original overlapping word components and the first level separated words.

In the second step, a fast ICA algorithm [112] is performed to calculate the weighted value matrix before the separated words are re-evaluated. Fast ICA is based on a fixed-point iteration scheme maximizing non-gaussianity as a measure of statistical independence. It is an efficient and popular algorithm for independent component analysis. Finally, the readjusted weighted value matrix is applied on the vector matrix to obtain the word components separately.

The flowchart of the whole algorithm is shown in Figure 6.2.

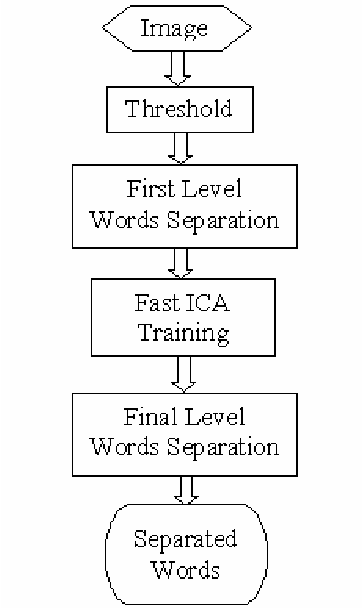


Figure 6.2 Flowchart of the ICA-based segmentation algorithm

6.2.1 Vector Matrix (VM) Building for ICA Model Training

The crucial part of the proposed algorithm is to build an effective Vector Matrix (VM), which is the source matrix for training the ICA model. The outline of building the Vector Matrix for the ICA-based segmentation algorithm is:

1. Pre-processing and Thresholding
2. Overlapping Word Component Detection
3. Overlapping Word Component Area Classification
4. Fuzzy Area Loops Classification
5. Classify Word Components and Restore Grey-Scale Value
6. Achieving Vector Matrix

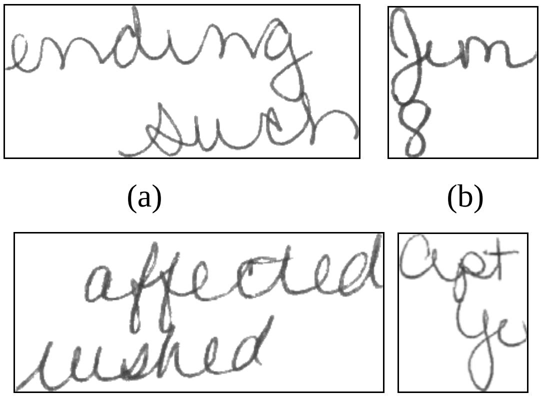
These stages are described in more detail below.

6.2.1.1 Pre-processing and Thresholding

The Decompose Thresholding Algorithm was described in Chapter 5 for thresholding degraded historical document images. The decompose algorithm has been demonstrated to be excellent for thresholding the degraded historical document image compared with other six algorithms.

An improved version of the Decompose Thresholding Algorithm, where the document image is recursively broken down into sub-regions using quad-tree decomposition until a suitable thresholding weighted value is chosen for mean-gradient thresholding method, which can be applied to each sub-region. Mean-gradient thresholding method [50] which is sensitive to edges so that it works well in keeping strokes' internal loop and faint tips of words from sub-regions. The subsequent step after obtaining a binary image is to segment the handwritten words.

6.2.1.2 Touching/Overlapping Word Components Detection



(a) (b)

(c) (d)

Figure 6.3 Illustration of loops touching and overlapping with characters on adjacent lines

The extravagant loops in ascenders, descenders and upper case letters encountered in many historical documents often result in touching or overlapping of words on adjacent lines. And in most cases, the touching and overlapping of strokes happen in the characters which include loop descenders, such as ‘f’, ‘g’, ‘j’, ‘y’ and so on as shown in Figure 6.3(a) ~ (d). Those touching words can be detected by component labeling function (bwlabel) from Matlab.

6.2.1.3 Touching/Overlapping Words Region Classification

In order to separate the overlapping and touching words on adjacent lines, area classification of the touching/overlapping region is needed for subsequent processing.

The detected overlapping word components are first separated into Top Area, Bottom Area and Fuzzy Area. This classification is dependent on a left-to-right mapping histogram.

1. **Top Area**: the pixels in the first peak range (as shown in Figure 6.8(a)) of the histogram;
2. **Bottom Area**: if the last part of the histogram is flat, then the area from the last peak of the histogram to the end of the histogram is classified as Bottom Area; if the last part of the histogram is a peak, then the whole range of the peak is classified as Bottom Area;
3. **Fuzzy Area:** the unassigned area between the Top Area and Bottom Area is defined as Fuzzy Area.

6.2.1.4 Fuzzy Area Loop Classification.

Firstly, the Laplacian of Gaussian method is used to find edges by looking for zero crossings after filtering the image with a Laplacian of Gaussian filter. This operator works well on most historical handwriting binary image.

Secondly, the closed edges in the Fuzzy Area are retained, because most overlap words contain closed loops in the strokes overlapped area. These closed edges are the edges of the loops inside the Fuzzy Area's strokes. Normally, the maximum loop is the major part of the overlapped stroke area.

Finally, the Center Point Distance, the distance between the center point of two closed loops, is measured for each pair of closed loops in the Fuzzy Area to determine the closest loops. According to the position of two loops, they can be classified into upper loop or lower loop. These two loops are the major roles in handwriting overlapped component elementary classification.

6.2.1.5 Word Component Grey-Scale Value Restore

Firstly, the upper loop is dilated **N** times, where **N** is equal to the width of the word strokes in pixels. The grey scale value of the connected words located in the dilated area is restored. The area higher and near the upper loop is simply separated from the overlapping words. Secondly, the lower loop can be restored using the same process to separate the lower word from the overlapping words. In order to increase the accuracy of the separation in the overlapped region, the separated words and overlapping word components are used as the source signals to train the ICA model.

6.2.1.6 Building the Vector Matrix

The major task in this stage is to produce an effective Vector Matrix (VM) x as a source signal matrix for training the ICA model to separate the overlapping words. The VM for ICA-based Segmentation Algorithm is a 3-by- ($i \times j$) matrix, where i and j are the row and column numbers of the whole overlap words component respectively. As shown in Figure 6.4(a), A is the overlap words component which has m×n images, B is the first level separated upper word, and C is the first level separated lower word.

The first row of the components in VM are the pixel intensity values of the overlapping words component image A, which have been converted from $i \times j$ (i is row numbers, j is column number) format to 1×($i \times j$) (1 row, $i \times j$ columns) as shown in Figure 6.4(b). Then the upper class component B is converted from $i \times j$ to 1×($i \times j$), as the second row of the VM shown in Figure 6.4(b), as well as the same conversion of the lower word image component C as the third row of the VM shown in Figure 6.4(b).

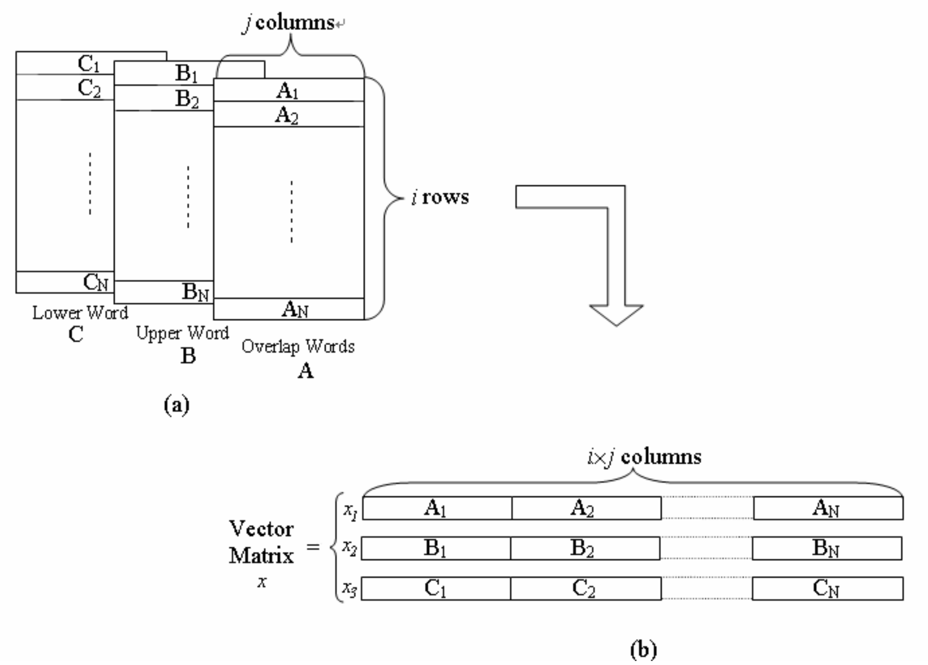


Figure 6.4 Conversion from Image to Vector Matrix

6.2.2 ICA Model Training

As shown in Figure 6.4(a) and Figure 6.4(b), the vector matrix x is produced after the first level separation of the overlapping word components. The elements x_1, x_2 and x_3 are the overlapping word components, the first level separated upper class word and the lower class word respectively. The vector matrix $x=[x_1\ x_2\ x_3]$ is the source signal of the ICA model.

The ICA model used in the proposed ICA-based Segmentation Algorithm is a Fast ICA. The Fast ICA algorithm is a computationally efficient method for performing the estimation of ICA. It finds the direction for the weight vector W maximizing the non-gaussianity of the projection $\mathrm{W}^{\mathrm{T}}x$ for the data x. It uses a fixed-point iteration scheme that has been found in independent experiments to be 10-100 times faster than conventional gradient descent methods [112]. Another advantage of the Fast ICA algorithm is that it can be used to perform projection pursuit as well, thus providing a general-purpose data analysis method that can be used both in an exploratory fashion and for estimation of independent components (or sources).

As shown in Figure 6.1, $x=\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$ is the input source signal of the Fast ICA model, where x_1 is the overlap words, x_2 is the initial separated upper class word, x_3 is the initial separated lower class word. The Fast ICA ($x= As$) is computed to find $\mathrm{W}=A^{-1}$, where A is a constant m × n ‘mixing’ matrix.

The weighting of this linear combination (which varies with each component) is given by a matrix vector, A. Each component of this vector has its own associated basis function, and represents an underlying 'cause' of the image. The vector *A* is determined by training the linear image synthesis ICA model.

From the ICA definition, ICA of the random vector *x* consists of finding a linear transform s=Wx so that the components s_i are as independent as possible, where W=A^{-1} [40]. The result of running the Fast ICA program W is the weighted value matrix, and it multiplies with the source signal matrix to obtain the independent component s_i.

$s = \begin{bmatrix} s_1 \\ s_2 \\ s_3 \end{bmatrix}$, where s_1 is the separated overlapped word 1; s_2 is the separated word 2; s_3 is the overlap area. s_1 and s_2 can be realigned to images which include the separated words respectively as shown in Figure 6.5.

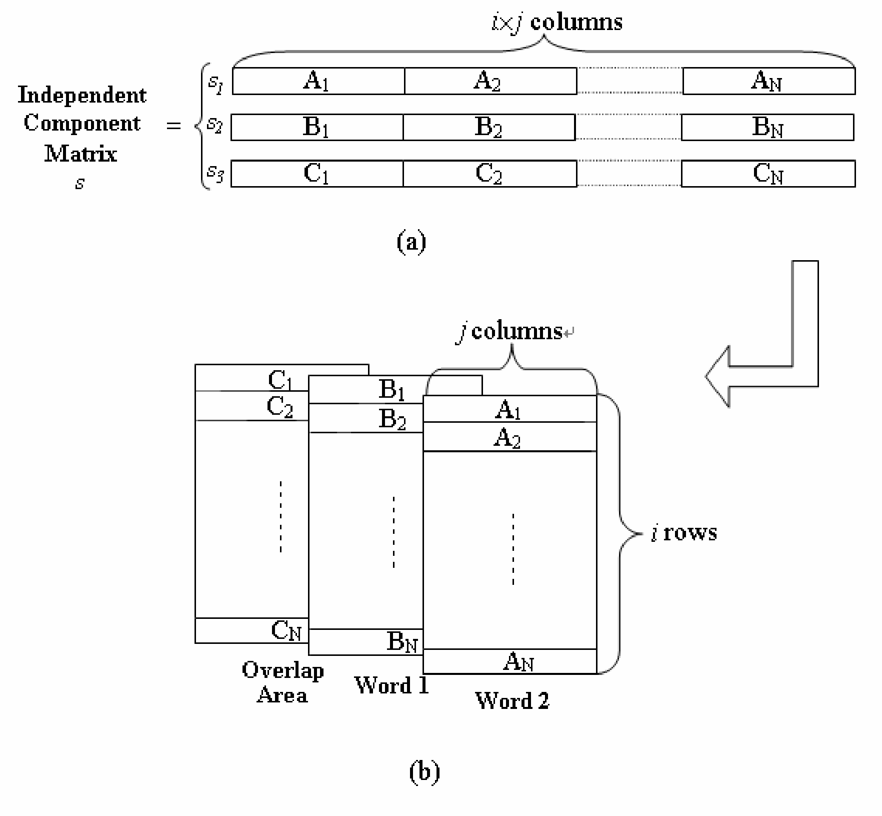
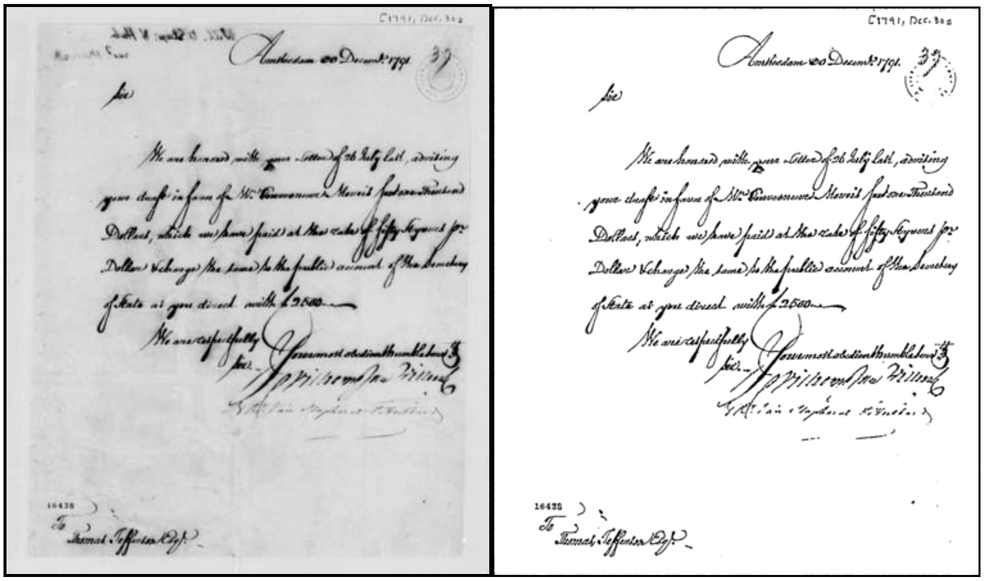


Figure 6.5 Conversion from Independent Components to Image

6.3 Experimental Results

6.3.1 Decompose Threshold

Figure 6.6(a) shows a typical degraded historical document image, and the binary result of the original image using the Improvement Decompose Threshold Algorithm described in Chapter 5 is shown in Figure 6.6(b).



(a) (b)

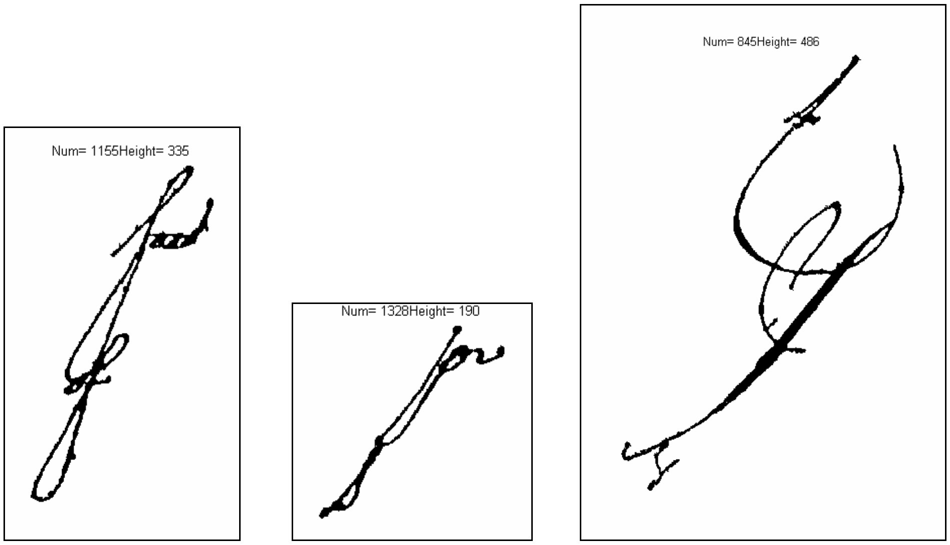
Figure 6.6 (a) Typical degraded historical document image; (b) Binary result by using the Improvement Decompose Threshold Algorithm

This algorithm has proven effective at dealing with degraded document images [11] and is used as the initial pre-processing in these experiments.

6.3.2 First Level Separation

6.3.2.1 Overlap Words Component Detection

As can be observed in Figure 6.6(a), there are some touching and overlapping words on adjacent lines, which are difficult to separate. In order to detach the overlapping words, all the connected character components in Figure 6.6(b) are first labeled. From observation, the size of touching and overlapping words on adjacent lines are always much bigger than the words in a single text line. All the connected character components on adjacent lines are separated out from Figure 6.6(b) as shown in Figure 6.7.



(a) (b) (c)

Figure 6.7 Connected character components of Figure 6.6

6.3.2.2 Region Classification of Touching/Overlapping Words Component

Figure 6.8(a) shows one of the detected overlapping components in Figure 6.6. Figure 6.8(b) is the histogram of the number of pixels in each row of Figure 6.8(a). The overlapping component is classified into three areas: Top Area, Bottom Area and Fuzzy Area according to analysis of the histogram in Figure 6.8(b). The result is shown in

Figure 6.9. The fuzzy area is the region between the upper line (area) and lower line (bottom area).

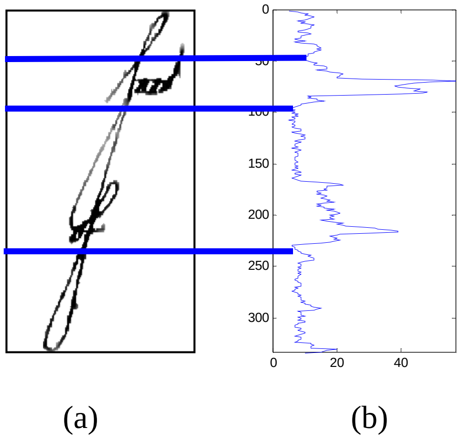
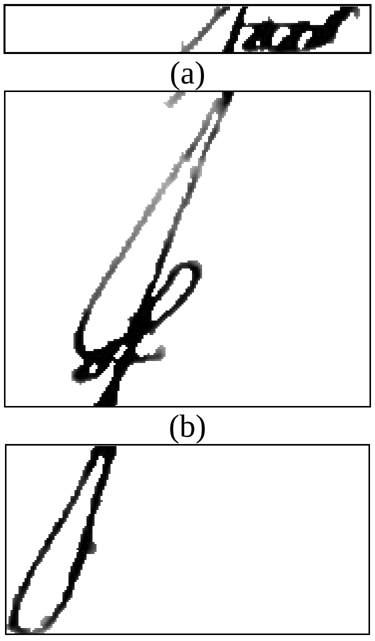


Figure 6.8 (a) Overlapping Component; (b) Histogram of the number of pixels in each row of (a)


(a)

(b)

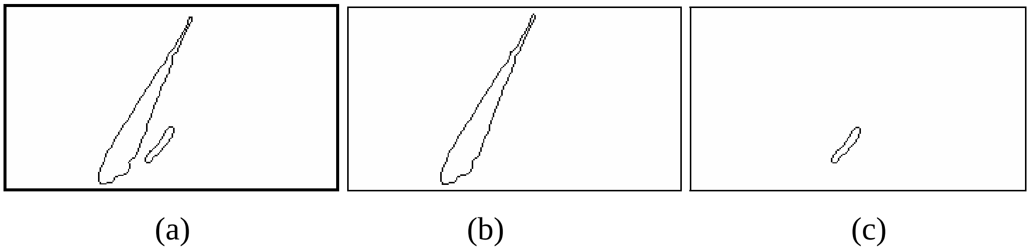
(c)

Figure 6.9 (a) Top Area; (b) Fuzzy Area; (c) Bottom Area

6.3.2.3 Fuzzy Area Loops Classification.

As shown in Figure 6.10(a), the two closed edges are the edges of the loops inside the Fuzzy Area's strokes. The largest loop can be detected as shown in Figure 6.10(b). It has been observed that normally the largest loop is the major part of the overlapped stroke area.

The center point distance of each of the two loops in the Fuzzy Area is measured to determine the closest loops as shown in Figure 6.10(a). According to the position of the two loops, they can be classified as upper loop or lower loop that is respectively shown in Figure 6.10(b) and Figure 6.10(c).



(a) (b) (c)

Figure 6.10 (a) Fuzzy Area's Loops; (b) Upper loop; (c) Lower Loop

6.3.2.4 Word Components Grey-Scale Value Restore

The dilated loop area is shown in Figure 6.11(a), where the grey scale value of the connected words region is restored in Figure 6.11(b). The area higher and near the upper loop is restored as shown in Figure 6.11(c). Hence, the first word is separated from the overlapping words.

The lower loop can be restored using the same process as shown in Figure 6.12 to separate the lower word from the overlapping words.

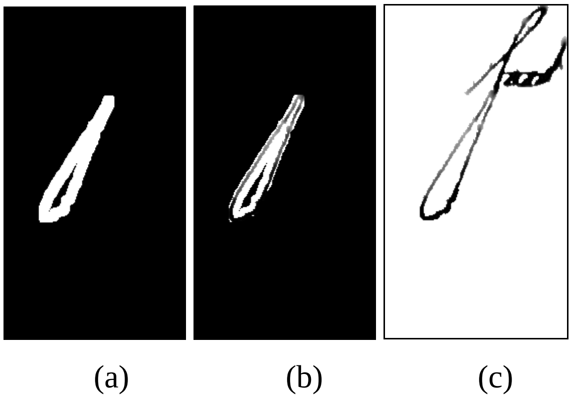


Figure 6.11 (a) Dilated Upper Loop Area; (b) Restored Upper Loop Area; (c) Restored Upper Words

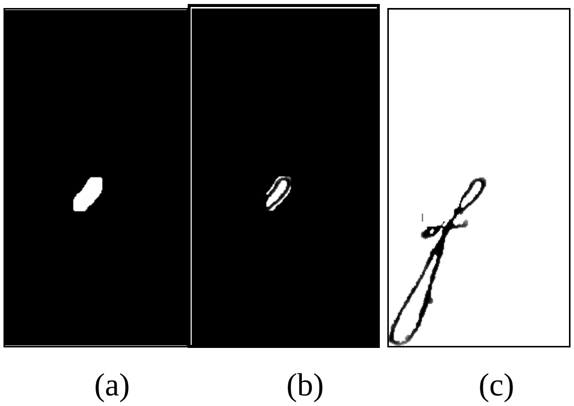


Figure 6.12 (a) Dilated Lower Loop Area; (b) Restored Lower Loop Area; (c) Restored Lower Words

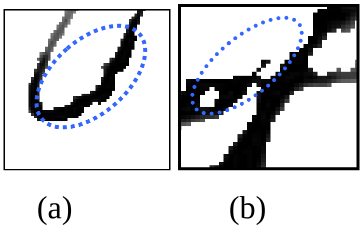


Figure 6.13 (a) Blown-up version of Figure 6.11c; (b) Blown-up version of Figure 6.12 (c)

The upper word in Figure 6.11(c) and the lower word in Figure 6.12(c) show the result of the first level separation. Figure 6.13 is the enlarged version of Figure 6.11(c) and Figure 6.12(c). It can be observed that the boundary of the separated stroke region shown in the round dotted circle is very coarse for recognition.

6.3.2.5 Vector Matrix

The Vector Matrix is built by the initially separated upper and lower words together with the overlapping word component.

As shown in Figure 6.4, $x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$, where x_1 is the overlapping words, x_2 is the initial separated upper class word, x_3 is the initial separated lower class word.

6.3.3 Second Level Separation based on Fast ICA Training

The linear transform equation of ICA training is

$$x=As \qquad \text{Eq(6-1)}$$

where x is the vector matrix built in Section 6.3.2, s_i in the vectors $s=(s_1,\ldots,s_n)^T$ are assumed independent. The matrix A is a constant m × n ‘mixing’ weighted value matrix.

Eq(6-1) can be converted to

$$s=\mathrm{W}x \qquad \text{Eq(6-2)}$$

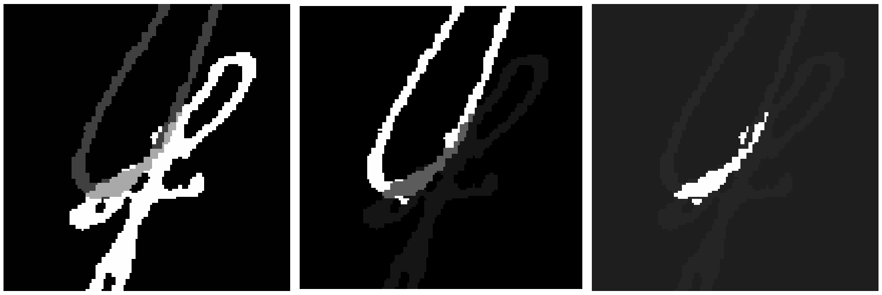
so that the components s_i are as independent as possible, where $\mathrm{W}=A^{-1}$.

The outcome of training Fast ICA is the weighted value matrix W as shown below:

$$W = \begin{bmatrix} 1.32 & -1.30 & 0.04 \\ 2.96 & -2.96 & -2.96 \\ -1.86 & -0.03 & 1.86 \end{bmatrix}$$

The weighted value matrix W is multiplied to the source signal matrix x to adjust the separation results. The enlarged version results are shown in Figure 6.14.

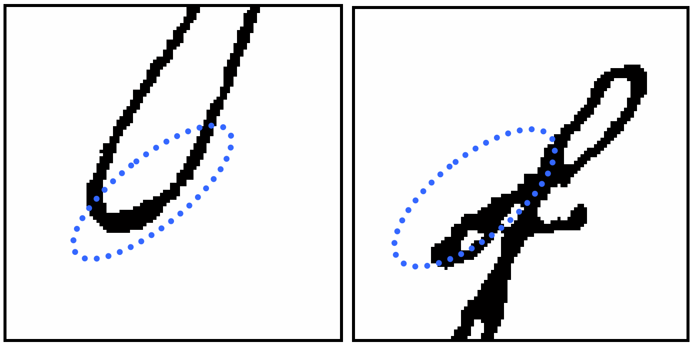
Figure 6.14(a) is part of the lower word of overlap words. Figure 6.14(b) shows the part of the upper separated word, and Figure 6.14(c) shows the overlapped area.



(a) (b) (c)

Figure 6.14 (a) Lower Word; (b) Upper Word; (c) Overlapping Area

Figure 6.15 shows the binary results of the final separated words. Within the dotted circle area, the much smoother boundary is shown on the overlapped area compared to the first level separated results in Figure 6.13.



(a) (b)

Figure 6.15 (a) Part of Separated Upper Word; (b) Part of Separated Lower Word

6.3.4 ICA-based Segmentation

From an aesthetic and subjective point of view, the proposed segmentation can solve the problem of overlapping and touching words on adjacent lines problem better than other segmentation method. It initially separates the overlapping words into upper and lower words, then uses the initially separated image as the source signal to train the Fast ICA model to obtain the adjusted weighted value matrix, and then adjusts the initial separated word to obtain the final separated words.

The ICA-based segmentation algorithm is highly effective for images which contain extravagant loops in ascenders, descenders and upper case letters.

30 historical images, which exhibit extravagant loops in ascenders, descenders and upper case letters, were selected from the Library of Congress on-line database to train the algorithms. The images were chosen to have varying resolutions, sizes, and contrast to ensure correct comparison of performance between the algorithms.

The ICA-based segmentation algorithm was further evaluated on another set of 30historical images, which exhibit extravagant loops in ascenders, descenders and upper case letters, also selected from the Library of Congress database.

From an aesthetic and subjective point of view, the ICA training structure performs better than other segmentation techniques. It re-evaluates the overlapping region and gives the best weighted matrix for the overlapping region again to obtain smooth separated strokes boundaries.

The images were, as with the training set, chosen to have varying resolutions, sizes, and contrast. In these selected images, some historical images still have acceptable quality even through they were created many years ago. The images were characterized by high resolution of the scanned images with varying contrast of the handwriting.

The standard measure, recall [108], was used to quantitatively show the relative performance of the proposed method at separating the overlapping words on adjacent lines. Table 6.1 shows the recall value of the ICA-based segmentation for the 30 example historical images.

Table 6.1 Recall value of the ICA-based segmentation

Image No.	1	2	3	4	5	6	7	8	9	10
Number of Correctly Separated Components	5	3	3	4	6	3	4	3	1	1
Number of Overlapping Components in the Image	5	3	4	4	6	3	4	3	2	1
Recall Value	100%	100%	75%	100%	100%	100%	100%	100%	50%	100%
Image No.	11	12	13	14	15	16	17	18	19	20
Number of Correctly Separated Components	2	4	2	3	5	4	2	2	4	3
Number of Overlapping Components in the Image	2	4	3	3	5	4	2	3	4	3
Recall Value	100%	100%	67%	100%	100%	100%	100%	67%	100%	100%
Image No.	21	22	23	24	25	26	27	28	29	30
Number of Correctly Separated Components	3	2	3	2	6	1	3	3	3	3
Number of Overlapping Components in the Image	5	2	3	2	6	1	3	3	3	4
Recall Value	60%	100%	100%	100%	100%	100%	100%	100%	100%	75%

The average recall value for the above 30 images is 93.9%, which shows that the ICA-based segmentation algorithm is an effective method for segmentation of overlap words on adjacent lines. More experimental results are shown in Appendix 4 and the attached CD.

6.4 Summary

A number of segmentation techniques have previously been shown to be effective at segmenting words correctly if the handwritten text lines are not overlapping or touching. However, none of them has been shown able to produce consistently good results on the wide range of document images containing touching or overlapping handwritten strokes. Many historical document images contain florid handwriting, which frequently exhibits extravagant loops in ascenders, descenders and upper case letters. These often result in touching or overlapping of words on adjacent lines.

The proposed ICA (Independent Component Analysis) Based Segmentation Algorithm works effectively for this kind of difficult task. The first step of the new ICA algorithm is to convert the original touching or overlapping word components into a vector source matrix. In the second step, a fast ICA model is run to calculate the weighted value matrix before the overlapped regions are re-evaluated. Finally, the readjusted weighted value matrix is applied on the vector matrix to obtain the word components separately.

The proposed algorithm has excellent performance on separating the overlapping words, which include loops in ascenders, descenders and upper case letters on adjacent lines. The method can be extended to separate other overlap patterns on adjacent lines.

Chapter 7

Conclusion and Future Research

The research in separating and segmenting text in degraded document images is a crucial step in a Document Management System. Subsequent character or word recognition can only be performed on the characters and words that have been extracted accurately from a noisy background.

An effective thresholding algorithm is the prerequisite for separating text from different types of scanned document for example, historical letters, forms, newspapers and cheque images' background. Compared to other types of images, degraded historical documents are ones which have become degraded due to age, handling or paper quality making the task more difficult. New thresholding algorithms are needed in order to separate text from the background in historical document images. Furthermore, the florid handwriting, frequently encountered in historical documents often exhibits extravagant loops in ascenders, descenders and upper case letters, which result in touching or overlapping of words on adjacent lines. To investigate techniques to separate the overlapping words on adjacent lines of historical document images resulting from florid handwriting in ascenders, descenders and upper case letters is another important step for subsequent character or word recognition in a Document Management System.

7.1 Achievements

There are several effective global and local thresholding algorithms that have been proposed in this thesis for different types of document images, especially the multistage approach for degraded historical document images, as well as the ICA-based segmentation for overlapping/touching words in adjacent lines.

In Chapter 3, an improved QIR (Quadratic Ratio Technique) algorithm is proposed for document images. The improved QIR algorithm enhances the image before applying the binarization step. In addition, the IR (Integral Ratio) algorithm is used for double-sided noise removal. The original QIR [83] has two main steps in the binarization stage, which are: fuzzy area determination and final thresholding value determination in fuzzy area according to the specific types of author's pen. However, it is very difficult to know what kind of pen the author had used. The determination method cannot provide an accurate cutting point at the second step. In the improved QIR algorithm, IR class is used to determine the final cutting point in the fuzzy area of the histogram, which is determined in the first step of the original QIR algorithm.

Based on research on histogram-based global thresholding, it had been found to be restricted to bimodal histogram document images. However, many document images do not exhibit a bimodal histogram. Furthermore, the global method cannot perform well on some document images which contain complex features. In Chapter 4, the research proceeded to derive the local adaptive mean-gradient method to separate text from different types of document images. It is a further improved variant of Niblack's local thresholding [60] approach, based on local mean and local mean-gradient values. Four categories of image were examined - historical documents, newspapers, forms, and cheques. They were used to test the algorithms' performance using varying resolutions, sizes, as well as contrast to ensure correct comparison of performance of the algorithms.

In Chapter 5, after proposing a local mean-gradient thresholding method for four different kinds of document images, a new structure, called the decompose thresholding approach is proposed for degraded historical images. This is the most challenging document type in the four types of document images considered. Although many thresholding algorithms have been proposed (see Chapter 2), there is a common disadvantage among them. Each of them can only work effectively with a limited range of image types. Sometimes, a technique will perform well in one type of document, but may totally fail in another document type. In a group of historical images, there are possibly all kinds of document degradations which cause noise and degradations in the document image (e.g. double-sided effect, low contrast, faint stroke, etc.); and it is difficult to decide which algorithm should be chosen for them on an individual basis. In this research, it was noted that there are some feature vectors that can be used to describe the characteristics of each image (sub-image) so that an appropriate algorithm or algorithms can be applied to that image or sub-image to produce high quality results. The local decompose thresholding is a multi-stage approach technique, which locally decides the best thresholding method for each block from the stage by stage analysis of the feature vectors of each image block. The improved decompose thresholding is based on analysing the information of sub-images and then choosing the most suitable weighted value for local mean-gradient thresholding method for them. Figures 5.10 and 5.12 and Table 5.2 summarize the recall values of the improved decompose thresholding algorithm and other six thresholding methods.

In Chapter 6, a new ICA (Independent Component Analysis) Segmentation Algorithm is investigated to separate touching/overlapping words in adjacent lines of degraded handwritten document images. The initial step of the algorithm is the first level separation of the overlapping words, which produces a blind source matrix (Vector Matrix) based on the original overlapping word components and the first level separated words. Secondly, a "fast" ICA model is performed to calculate the weighted value matrix before the separated words are re-evaluated. Finally, the readjusted weighted value matrix is applied to the Vector Matrix to obtain the separated word components. Table 6.1 summarized the recall value of the ICA-based segmentation.

7.2 Contributions

In this thesis, new algorithms have been described to separate and segment foreground words from degraded document images especially historical document images.

An improved QIR (Quadratic Integral Ratio) technique is proposed for extracting the handwritten text from noisy backgrounds. It has proven effective for document images compared to other existing global thresholding methods. It can retain more useful information and reduce noise around handwritten characters and has been proven to have better performance for bimodal histogram document images. More experimental results of QIR and Improved QIR are shown in Appendix 1 and the attached CD.

A mean-gradient technique was proposed and evaluated. This technique analyses the mean-gradient in local regions for different types document images. It can produce better results on the four different image types. In contrast, the other three techniques only focus on one or two document types. The comparison showed that the Mean-Gradient technique works well in keeping high contrast strokes, and retains the small holes in characters. It is particularly effective in removing blotches and smudges in images while maintaining the handwriting details, and performs well in removing patterned backgrounds. Tables 4.1 to 4.4 summarize the Precision and Recall values of the mean-gradient method and other three thresholding methods. More experimental results are shown in Appendix 2 and the attached CD.

A decompose algorithm was proposed and evaluated, which analyses the block information in local areas after which the most appropriate threshold method for that area is determined. The improved decompose algorithm is highly effective for the images, which contain different conditions at different locations. It can produce better performance on historical document images. Figure 5.10 summarizes the Recall values of the Decompose Threshold method, Figures 5.11 and Table 5.2 summarize the Recall values of the Improved Decompose Threshold method compared to other six thresholding methods. More experimental results are shown in the attached CD.

From an aesthetic and subjective point of view, the multi-stage decompose structure performs better than other single-stage local methods. A multi-stage structure is proposed which depends on feature extraction, for degraded historical document image thresholding. It detects feature vectors of different areas and then applies appropriate methods to avoid losing important useful information. These features can be usefully used in knowledge-based segmentation /separation.

An ICA-based segmentation algorithm for separating overlapping/touching words in the adjacent lines is proposed. It is demonstrated to be effective at resolving the problems encountered in various forms of overlapping/touching text lines. There are some experimental results shown in Chapter 6.

In the course of the research, 6 papers (2 journal papers and 4 conference papers) have been submitted for publication, 5 have been published and the remaining one is currently under review.

Journals:

- Y. Chen and C.G. Leedham, "Decompose Algorithm for Thresholding Degraded Historical Document Images", IEE Proceedings on Vision, Image and Signal Processing, Vol. 152, No. 6, pp 702 – 714, 2005.
- Y. Chen and C.G. Leedham, "Independent Component Analysis to Separate Overlapping Handwritten Strokes in Degraded Document Images", submitted and under review by Pattern Recognition Letters.

Refereed Conferences:

- C.G. Leedham, Y. Chen, K. Takru, J. Tan and M. Li, "Comparison of some thresholding algorithms for text/background segmentation in difficult document images", *Proc. 7th Int. Conf. on Document Analysis and Recognition*, Edinburgh, United Kingdom, Vol. 2, pp 859 -865, 2003.
- Y. Chen, C.G. Leedham, "The Multistage Approach to Information Extraction in Degraded Document Images", *Proc. 17th Int. Conf. on Pattern Recognition*, Cambridge, United Kingdom, Vol.1, pp 445-448, 2004.
- Y. Chen and C.G. Leedham, "Decompose-threshold approach to handwriting extraction in degraded historical document images", *Proc. 9th Int. Workshop on Frontiers in Handwriting Recognition*, Tokyo, Japan, pp 239-244, 2004.

- Y. Chen and C.G. Leedham, ”Independent Component Analysis Segmentation Algorithm”, *Proc. 8th Int. Con. on Document Analysis and Recognition*, Seoul, Korea, Vol.2, pp 680-684, 2005.

7.3 Discussion and Suggestions for Future Work

Extracting text from degraded document image is still a crucial step in a Document Management System, and many aspects remain which need to be researched. Future work should concentrate on defining feature vectors to accurately describe the local texture properties. The decompose algorithm works well on handwritten documents containing text but does not work well on document images with big patterns or pictures. For example, the pictures in the newspapers images which contain big illustrations were not binarised well. More complete evaluation of the decompose method could investigate its applicability to a wider range of difficult document image, such as bank cheques and newspaper images.

The ICA-based segmentation algorithm only can separate overlapping/touching words with loop strokes in two adjacent lines. But this is not enough for an effective Document Management System. Future work can be extended to separate overlapping/touching words with/without loop strokes crossing the three or even more lines of the whole document image.

The local multi-stage thresholding structure can be implemented on Field Programmable Gate Arrays (FPGAs), which allows the development of digital architectures without requiring the complex processes used in a VLSI chip fabrication. FPGAs can be easily embedded in traditional system design flows to perform prototyping and emulation tasks. Moreover, they are best suited to reconfigurable implementations in which the hardware must be dynamically adaptable to a specific problem. By exploiting reconfigurability, the image processing techniques can be rapidly modified to optimize for performance without the need to rely on alternative hardware solutions.

The Virtex family of FPGAs could fully evaluate the proposed hardware techniques as they redefine the future of programmable logic that break density and performance barriers while offering unprecedented system level integration. High-end Virtex series consists of up to 2,000,000 system gates at clock speeds up to 200 MHz, and include many new features that address system level design methodologies for rapid development will enable the migration of complex algorithms and dynamic structures to reconfigurable platforms in an efficient manner and pave the way for area-time optimal VLSI implementations.

Future research should concentrate on refining and developing robust algorithms that can be efficiently implemented in hardware to achieve real-time implementation of the multi-stage algorithm on large difficult images.

BIBLIOGRAPHY

[1] Abutaleb A.S., (1989), "Automatic thresholding of gray-level pictures using two-dimensional entropy", Computer Vision, Graphics and Image Processing, Vol. 47, Page(s): 22-32.

[2] Amara M., deBrucq D., Courtellemont P., Wallon P., Mesmin C., Lecourtier Y. (1996), "A recursive estimation of parameters of straight lines and circles: application to the segmentation of the Rey's Complex Figure", in Proceedings of the 13th International Conference on Pattern Recognition, Vienna, Austria, Vol. 2(2) , Page(s): 467-471.

[3] Beghdadi A., Negrate A. L., and Lesegno P. V. D. (1995), "Entropy thresholding using a block source model", Graphical Models and Image Processing, Vol. 57, Page(s): 197-205.

[4] Bernsen J. (1986), "Dynamic thresholding of grey-level images", in Proceedings of the 8th International Conference on Pattern Recognition, Paris, France, Vol. 2, Page(s): 1251-1255.

[5] Bishnu A., Chaudhuri B. B. (1999), "Segmentation of Bangla handwritten text into characters by recursive contour following", in Proceedings of the 5th International Conference on Document Analysis and Recognition, Bangalore, India, Page(s): 402-405.

[6] Boukharouba S., Rebordao J. M., Wendel P. L. (1985), " An amplitude segmentation method based on the distribution function of an image", Computer Vision Graphics and Image Processing, Vol. 29, Page(s): 47-59.

[7] Breuel T. M. (2001), "Segmentation of hand printed letter strings using a dynamic programming algorithm", in Proceedings of the 6th International Conference on Document Analysis and Recognition, Seattle, USA, Page(s): 821-826.

[8] Brink A. D. (1995), "Minimum spatial entropy threshold selection", IEE Proceeding on Vision, Image and Signal Processing, Vol. 142 (3), Page(s): 128-132.

[9] Chang C. I., Chen K., Wang J. and Althouse M. L. G. (1994), "A relative entropy-based approach to image thresholding," Pattern Recognition, Vol. 27, No. 9, Page(s): 1275-1289.

[10] Cheriet M., Huang Y. S. and Suen C. Y. (1992), "Background region-based algorithm for the segmentation of connected digits", in Proceedings of the 11th IAPR International Conference on Pattern Recognition, Vol. 2, Conference B: Pattern Recognition Methodology and Systems, Hague, Netherlands, Page(s): 619-622.

[11] Chen T. and Takagi M. (1993), "Run length coding based new approach to automatic image thresholding", IEEE International Symposium on Circuits and Systems, Chicago, Illinois, USA, Page(s): 555-558.

[12] Cheriet M. (1998), "Extraction of handwritten data from noisy grey-level images using a multi-scale approach," in Proceedings of Vision Interface, Vol. 1, Vancouver, BC, Canada, Page(s): 389 - 396.

[13] Chen Y. K., Wang J. F. (2000), "Segmentation of handwritten connected numeral string using background and foreground analysis", in Proceedings of the

15th International Conference on Pattern Recognition, Vol. 2(2), Barcelona, Spain, Page(s): 598-601.

[14] Cheung K. W., Yeung D. Y. and Chin R. T. (2002), "Bidirectional deformable matching with application to handwritten character extraction", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 24(8), Page(s): 1133-1139.

[15] Chen Q. S., Zhen L. X. (2002), "Word segmentation in handwritten Chinese text image based on component clustering techniques", in IEEE Region 10 Technical Conference on Computers, Communications, Control and Power Engineering, Beijing, China, Vol. 1(1), Page(s): 435-440.

[16] Chigusa Y., Hattori T., Ikegami M. and Tanaka M. (1992), "An image binarization system for composite pictures", in Proceedings of IEEE International Symposium on Circuits and Systems, San Diego, CA, USA, Vol. 5, Page(s): 2292 -2295.

[17] Comon P. (1994), "Independent component analysis - a new concept", Signal Processing, Vol. 36, Page(s): 287-314

[18] Congedo G., Dimauro G., Impedovo S. and Pirlo G. (1995), "Segmentation of numeric strings", in Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, Canada, Vol. 2(2), Page(s): 1038-1041.

[19] Daekeun Y., Gyeonghwan K. (2003), "An approach for locating segmentation points of handwritten digit strings using a neural network", in Proceedings of the 7th International Conference on Document Analysis and Recognition, Edinburgh, Scotland, UK, Vol. 1, Page(s): 142-146.

[20] Djeziri S., Nouboud F., and Plamondon R. (1998), "Extraction of signatures from check background based on a filiformity criterion", IEEE Transactions on Image Processing, Vol. **7**, Page(s): 1425-1438.

[21] Deravi F., Pal S. K. (1983), "Grey-level thresholding using second-order statistics", Pattern Recognition Letter, Vol. 1, Page(s): 417-422.

[22] Don H. S. (1995), "A noise attribute thresholding method for document image binarization", in Proceedings of 3rd International conference on Document Analysis and Recognition, Montreal, Canada, Page(s): 231-234.

[23] Guo J. K., Ma M. Y. (2001), "Separating handwritten material from machine printed text using hidden Markov models", in Proceedings of the 6th International Conference on Document Analysis and Recognition, Seattle, USA, Page(s): 439-443.

[24] Eastwood B., Jennings A., Harvey A. (1997), "Neural network based segmentation of handwritten words", in Proceedings of the 6th International Conference on Image Processing and Its Applications, Dublin, Ireland, Vol. 2(2), Page(s): 750-755

[25] Eikvil L., Taxt T., and Moen K. (1991), "A fast adaptive method for binarization of document images", in Proceedings of the 1st International Conference on Document Analysis and Recognition, ICDAR, St. Malo, France, Page(s): 435-443.

[26] Ejiri M. (1989), "Knowledge-based approaches to practical image processing", in International Workshop on Industrial Applications of Machine Intelligence and Vision, Tokyo, Japan, Page(s): 1-8.

[27] Esposito F., Malerba D. and Semeraro G. (1993), "Automated acquisition of rules for document understanding", in Proceedings of the 2nd International Conference on Document Analysis and Recognition, Tsukuba City, Japan, Page(s): 650-654.

[28] Fan L. Y., Fan L. X. and Tan C. L. (2001), "Binarizing document image using coplanar prefilter", in Proceedings of the 6th International Conference on Document Analysis and Recognition, Seattle, USA, Page(s): 34-38.

[29] Feldbach M., Tonnies K. D. (2003), "Word segmentation of handwritten dates in historical documents by combining semantic a-priori-knowledge with local features", in Proceedings of the 7th International Conference on Document Analysis and Recognition, Edinburgh, Scottland, UK, Vol. 1, Page(s): 333-337.

[30] Fujisawa H., Nakano Y. and Kurino K. (1992), "Segmentation methods for character recognition: from segmentation to document structure analysis", Proceedings of the IEEE , Vol. 80 (7) , Page(s): 1079-1092.

[31] Fontanot P., Ramponi G. (1993), "A polynomial filter for the pre-processing of mail address images", in IEEE Winter Workshop on Nonlinear Digital Signal Processing, Tampere, Finland, Page(s): 2.1_6.1-2.1_6.6.

[32] Giuliano E., Paitra O., and Stringa L. (1977), "Electronic Character Reading System", United States Patent, Application Number: US1976000738456.

[33] Gonzalez R. C., Woods R. E. (1993), "Digital Image Processing", Addison-Wesley Publishing Company, Page(s): 443 - 456.

[34] Gorman L. O. (1994), "Binarization and multi-thresholding of document images using connectivity", CVGIP: Graphical Models Image Process, Vol. 56(6), Page(s): 494 - 506.

[35] Gu L., Kaneko T., Tanaka N. and Haralick R. M. (1998), "Robust extraction of characters from color scene image using mathematical morphology," in Proceedings of 14th International Conference on Pattern Recognition, Brisbane, Australia, Page(s): 1002 - 1004.

[36] Johannsen G. and Bille J. (1982), "A threshold selection method using information measures," in Proceedings of 6th International Conference on Pattern Recognition, Munich, Germany, Page(s): 140 - 143.

[37] Hamid A., Haraty R. (2001), "A neuro-heuristic approach for segmenting handwritten Arabic text", IEEE International Conference on Computer Systems and Applications, Beirut, Lebanon, Page(s): 110 - 113

[38] Hertz L. and Schafer R. W. (1988), "Multilevel thresholding using edge matching", Computer Vision, Graphics, and Image Processing, Vol. 44, Page(s): 279 - 295.

[39] Hurri J. (1997), "Independent Component Analysis of Image Data", Master thesis, Helsinki University of Technology, Helsinki, Finland.

[40] Hyvarinen A. (1999), "Survey on Independent Component Analysis", Helsinki University of Technology, Finland.

[41] Junker M. and Hoch R. (1999), "On the Evaluation of Document Analysis Components by Recall, Precision, and Accuracy", in Proceeding of 5th International Conference on Document Analysis and Recognition, Bangalore, India, Page(s): 713 - 716.

[42] Jutten C., Herault J. (1991), "Blind separation of sources", Signal Processing, Part I: An adaptive algorithm based on neuromimetic architecture. Vol. 24, Page(s): 1-10

[43] Kapur J. N., Sahoo P. K., and Wong A. K. C. (1985), "A new method for grey-level picture thresholding using the entropy of the histogram," Computer Vision, Graphics, and Image Processing, Vol. 29, Page(s): 273 - 285.

[44] Kamel M. and Zhao A., (1993), "Extraction of binary character/graphics images from grey-scale document images", CVGIP: Graphical Models and Image Process, Vol. 55, No. 3, Page(s): 203 - 217.

[45] Kim S. H., Jeong S., Lee G. S., Suen C. Y. (2001), "Gap metrics for handwritten Korean word segmentation", Electronics Letters, Vol. 37(14), Page(s): 892 – 893

[46] Kittler J., Illingworth J. (1985), "Threshold selection based on a simple image statistic", CVGIP: Graphical Models and Image Process, Vol. 30, Page(s): 125-147.

[47] Kittle J., Illingworth J., (1986), "Minimum error thresholding", Pattern Recognition, Vol. 19, No.1, Page(s): 41- 47.

[48] Kohler R., (1981), "A segmentation system based on thresholding", Computer Vision, Graphics and Image Processing, Vol. 15, Page(s): 319 - 338.

[49] Lam S.W. (1996), "Texture feature extraction using grey level gradient based co-occurrence matrices", in Proceedings of IEEE International Conference on Systems, Man, and Cybernetics, Beijing, China, Vol. 1, Page(s): 267-271.

[50] Leedham C. G., Chen Y., Takru K., Tan J. and Li M. (2003), "Comparison of some thresholding algorithms for text/background segmentation in difficult document images", in Proceedings of the 7th International Conference on Document Analysis and Recognition, Scotland, UK, Vol. 2, Page(s): 859-865.

[51] Leung C. K. and Lam F. K., (1996), "Maximum segmented-scene spatial entropy thresholding", in Proceedings of the IEEE International Conference on Image Processing, Lausanne, Switzerland, Page(s): 963-966

[52] Liang S., Ahmadi M. (1994), "A morphological approach to text string extraction from regular periodic overlapping text/background images", CVGIP: Graphical Models and Image Processing, Vol.56, No.5, Page(s): 402-413.

[53] Liao P. S., Chen T. S. and Chung P. C. (2001), "A Fast Algorithm for Multilevel Thresholding", Journal of Information Sciense and Engineering, Vol. 17, Page(s):713-727

[54] Liu Y., Srihari S. N. (1997), "Document image binarization based on texture features", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 19, No. 5, Page(s): 540-544.

[55] Liu K., Suen C. Y., Cheriet M., Said J. N., Nadal C., and Tang Y. Y. (1997), "Automatic extraction of baselines and data from check images," International Journal of Pattern Recognition and Artificial Intelligence, Vol. 11, No. 4, Page(s): 675-697.

[56] Machii K., Fukushima H. and Nakagawa M. (1993), "On-line text/drawings segmentation of handwritten patterns", in Proceedings of the 2nd International Conference on Document Analysis and Recognition, Tsukuba City, Japan, Page(s): 710-713.

[57] Manjunath B. S., Chellappa R. (1991), "A computational approach to boundary detection", IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Maui, HI, USA, Page(s): 358-363.

[58] Mestetskii L. M., Reyer I. A. and Sederberg T. W. (2002), “Continuous approach to segmentation of handwritten text”, in Proceedings of the 8th International Workshop on Frontiers in Handwriting Recognition, Ontario, Canada, Page(s):440-445

[59] Naoi S., Hotta Y., Yabuki M. and Asakawa A. (1994), “Global interpolation in the segmentation of handwritten characters overlapping a border”, in Proceedings of the IEEE International Conference on Image Processing, Austin, Taxas, USA, Vol. 1(1), Page(s): 149-153

[60] Niblack W. (1986), “An Introduction to Digital Image Processing”, Prentice Hall, Page(s): 115 - 116.

[61] Negishi H., Kato J., Hase H. and Watanabe T. (1999), “Character extraction from noisy background for an automatic reference system”, in Proceedings of 5th International Conference on Document Analysis and Recognition, Bangalore, India, Page(s): 143-146

[62] Ohya J., Shio A., and Akamatsu S. (1994), “Recognizing characters in scene images”, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 16, Page(s): 214-220.

[63] Oh I. S. (1995), “Document image binarization preserving stroke connectivity,” Pattern Recognition Letters, Vol. 16, Page(s): 743-748.

[64] Otsu N. (1979), “A threshold selection method from grey level histogram”, IEEE Transactions on Systems, Man and Cybernetics, Vol. 9, No. 1, Page(s): 62-66.

[65] Palumbo P. W., Swaminathan P. and Srihari S. N. (1986), “Document image binarization: Evaluation of algorithms,” SPIE Applications of Digital Image Processing IX, Vol. 697, Page(s): 278-285.

[66] Pal N. R.; Pal S. K. (1988), “Object extraction from image using higher order entropy”, in Proceedings of 9th International Conference on Pattern Recognition, Rome, Italy, Vol.1, Page(s): 348 - 350.

[67] Papamarkos N., Gatos B. (1994), “A new approach for multi-level threshold selection”, CVGIP: Graphical Models and Image Processing, Vol. 56(5), Page(s): 357-370.

[68] Parker J. (1991), “Gray level thresholding on badly illuminated images”, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 13, Page(s): 813-819.

[69] Plamondon R., Privitera C. M. (1999), “The segmentation of cursive handwriting: an approach based on off-line recovery of the motor-temporal information”, IEEE Transactions on Image Processing, Vol. 8(1), Page(s): 80-91

[70] Pun T. (1981), “Entropic thresholding: A new approach”, Computer Vision, Graphics, and Image Processing, Vol. 16, Page(s): 210 - 239.

[71] Qian W., Chew L. T. (2001), “Matching of Double-Sided Document Images to Remove Interference”, IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Kauai Marriott, Hawaii, USA, Vol. 1, Page(s): 1084-1089.

[72] Rangsanseri Y., Rodtook S. (2001), “Comparative study of thresholding techniques for grey-level document image binarization”, Proceedings of the

IEEE Region 10 International Conference on Electrical and Electronic Technology, Vol. 1, Page(s): 152-155.

[73] Sabourin R., Plamondon R. (1988), "Segmentation of handwritten signature images using the statistics of directional data", in Proceedings of the 9th International Conference on Pattern Recognition, Rome, Italy, Vol. 1, Page(s): 282-285

[74] Sadri J., Suen C. Y., Bui T. D. (2004), "Automatic Segmentation of Unconstrained Handwritten Numeral Strings", in Proceedings of the 9th International Workshop on Frontiers in Handwriting Recognition, Tokyo, Japan, Page(s): 317-322

[75] Sahoo P. K., Soltani S., and Wong A. K. C. (1988), "A survey of thresholding techniques", Computer Vision, Graphics, and Image Processing, Vol. 41, Page(s): 233 -260.

[76] Said J.N., Cheriet M., and Suen C. Y. (1996), "Processing of business forms: system overview," in Proceeding of the 13th International Conference on Pattern Recognition, Vienna, Austria, Page(s): 84 - 92.

[77] Salton, G. (1989), "Automatic text processing: the transformation, analysis, and retrieval of information by Computer", Addison-Wesley Series in Computer Science, Page(s): 530.

[78] Sezgin M., Sankur B. (2004), "Survey over image thresholding techniques and quantitative performance evaluation", Journal of Electronic Imaging, Vol. 13(1) Page(s): 146-165.

[79] Sezan M. I. (1985), "A peak detection algorithm and its application to histogram-based image data reduction," CVGIP: Graphical Model and Image Processing, Vol. 29, Page(s): 47-59.

[80] Sharma .G (2001), "Show-through cancellation in scans of duplex printed documents", IEEE Transactions on Image Processing, Vol. 10, No. 5, Page(s): 736-754.

[81] Shapiro L. G., Stockman G.C. (2001), "Computer Vision", Prentice Hall, ISBN 0-13-030796, Page(s): 387.

[82] Shi Z., Srihari S. N., Shiu Y. C. and Ramanaprasad V. (1997), "A system for segmentation and recognition of totally unconstrained handwritten numeral strings", in Proceedings of the 4th International Conference on Document Analysis and Recognition, Ulm, Germany, Vol. 2(2), Page(s): 455 – 458

[83] Solihin Y., Leedham G. C. (1999), "Integral Ratio: A New class of Global Thresholding Techniques for Handwriting Images", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.21, No. 8, Page(s): 761-768.

[84] Suen C. Y., Tang Y. Y., Yu C. L. (1993), "Document architecture language (DAL) approach to document processing", in Proceedings of the 2nd International Conference on Document Analysis and Recognition, Tsukuba City, Japan, Page(s): 103-106

[85] Suwa M., Naoi S. (2004), "Segmentation of Handwritten Numerals by Graph Representation", in Proceedings of the 9th International Workshop on Frontiers in Handwriting Recognition, Tokyo, Japan, Page(s): 334-339

[86] Su L., Ahmadi M., Shridhar M. (1997), "Segmentation of handwritten interference marks using multiple directional stroke planes and reformalized

morphological approach", IEEE Transactions on Image Processing, Vol. 6(8), Page(s): 1195-1202

[87] Strathy N. W., Suen C. Y., Krzyzak A. (1993), "Segmentation of handwritten digits using contour features", Proceedings of the 2nd International Conference on Document Analysis and Recognition, Tsukuba City, Japan, Page(s): 577 – 580

[88] Taxt T., Flynn P. J., Jain A. K. (1989), "Segmentation of Document image", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 11 (12), Page(s): 1322 - 1329.

[89] Trier O. D., Taxt T. (1994), "Evaluation of binarization methods for utility map images", in Proceedings of IEEE International Conference on Image Processing, Austin, Texas, USA, Vol.2, Page(s): 1046 – 1050.

[90] Trier O. D., Jain A. K. (1995), "Goal-directed evaluation of binarization methods", IEEE Transactions on Pattern Analysis Machine Intelligence, Vol. 17(12), Page(s): 1191-1201.

[91] Trier O. D., Taxt T. (1995), "Evaluation of binarization methods for document images", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 17 (3), Page(s): 312 – 315

[92] Trier ϕ.D., Taxt T. (1995), "Improvement of 'Integrated Function Algorithm' for Binarization of Document Images", Pattern Recognition Letters, Vol. 16, Page(s): 277 -283.

[93] Tripathy N., Pal U. (2004), "Handwriting Segmentation of Unconstrained Oriya Text", in Proceedings of the 9th International Workshop on Frontiers in Handwriting Recognition, Tokyo, Japan, Page(s): 306-311.

[94] Tsai W. (1985), "Moment-Preserving Thresholding: a new approach", Computer Vision, Graphics, and Image Processing, Vol. 29, Page(s): 377-393.

[95] Veloso L. R., Sousa R. P., Carvalho J. M. (2000), "A new morphological method for cursive word segmentation", in Proceedings of the 7th International Conference on Image Processing, Vancouver, BC, Canada, Vol. 2(2), Page(s): 704-707.

[96] Wang S., Haralick R. M. (1984), "Automatic multi-threshold selection", Computer Vision, Graphics and Image Processing, Vol. 25, Page(s): 46-67.

[97] Wang L., Pavlidis T. (1993), "Direct grey-scale extraction of features for character recognition," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 15, Page(s): 1053-1067.

[98] Wang J. W., Yingzi E., Du C. C., Thouin P.D.; (2002), "Relative entropy-based methods for image thresholding", IEEE International Symposium on Circuits and Systems, Scottsdale, Arizona, USA, Vol. 2, Page(s): 265-268.

[99] Weszka J. S. (1978), "A survey of threshold selection techniques", Computer Vision, Graphics and Image Processing, Vol. 7, Page(s): 259-265.

[100] Weszka J. S., Rosenfeld A. (1978), "Threshold Evaluation techniques", IEEE Transactions on Systems, Man, and Cybernetics, Vol. 8, Page(s): 622-629.

[101] White J. M. and Rohrer G. D. (1983), "Image thresholding for Optical Character Recognition and other Applications requiring Character Image

Extraction", IBM Journal of Research and Developments, Vol. 27, No. 4, Page(s): 400-410.

[102] Yamamoto H., Sakaue S., Maruno S., Shimeki Y. (1993), "Segmentation of handwritten Japanese character strings with Hopfield type neural networks", in Proceedings of 1993 International Joint Conference on Neural Networks, Nagoya, Japan, Vol. 3(3), Page(s): 2073-2076

[103] Yanowitz S. D. and Bruckstein A. M. (1989), "A new method for image segmentation", Computer Vision, Graphics and Image Processing, Vol. 46, No. 1, Page(s): 82 - 95.

[104] Yang Y. B., Yan H. (2000), "An Adaptive Logical Method For Binarization of Degraded Document Images", Pattern Recognition, Vol. 33, No. 5, Page(s): 787-807.

[105] Yasuda Y., Dubois M., Huang T. S. (1980), "Data compression for check processing machine", Proceeding of IEEE, Vol. 68 (7), Page(s): 874 - 885.

[106] Ye X., Cheriet M., Suen C. Y., and Liu K. (1999), "Extraction of bank check items by mathematical morphology," International Journal of Document Analysis and Recognition, Vol. 2, Page(s): 53-66.

[107] Ye X., Cheriet M., and Suen C. Y. (1999), "Model-based character extraction from complex backgrounds", in Proceeding of 5th International Conference on Document Analysis Recognition, Bangalore, India, Page(s): 511-514.

[108] Zhang Y. J., (1996), "A Survey on Evaluation methods for image segmentation", Pattern Recognition, Vol. 29, no.8, Page(s): 1335-1346.

[109] Zhao M., Yan H. (1999), "Adaptive thresholding method for binarization blueprint images", in Proceedings of the 5th International Symposium Signal Processing and Its Applications, Vol. 2, Page(s): 931-934.

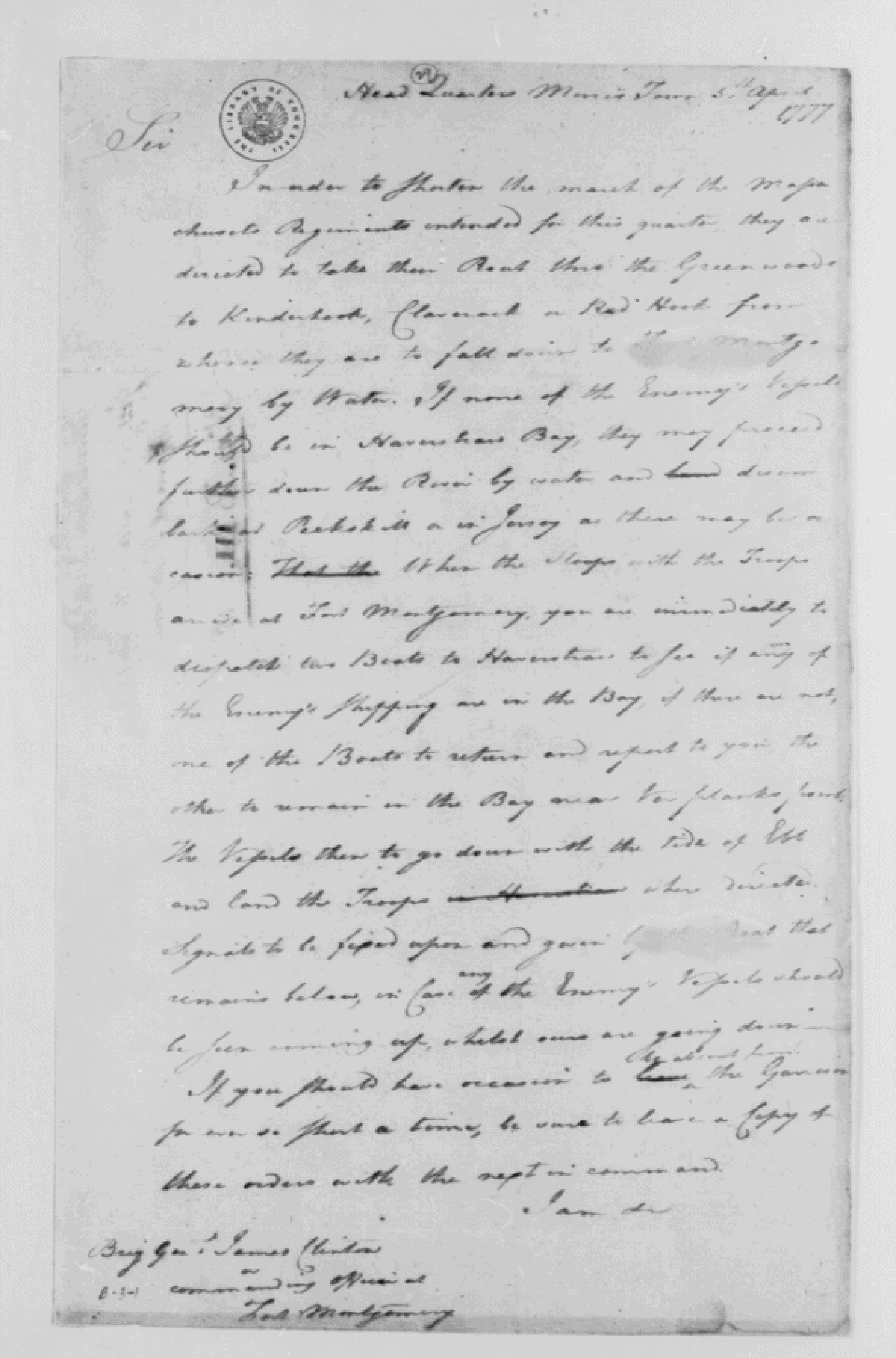
[110] Zhang Z. and Tan C. L. (2001), "Restoration of images scanned from thick bound documents", in Proceeding of 8th International conference on Image Processing, Thessaloniki, Greece, Vol. 1, Page(s): 1074-1077.

[111] Zhao S. Y., Chi Z. R., Shi P. F. and Wang Q. (2001), "Handwritten Chinese character segmentation using a two-stage approach", in Proceedings of the 6th International Conference on Document Analysis and Recognition, Seattle, WA, USA, Page(s): 179-183.

[112] Available at: http://www.cis.hut.fi/projects/ica/fastica/fp.shtml

[113] Available at: http://algorithm.diy.myrice.com/algorithm/complexity/index.htm

APPENDIX 1: Experimental Results of Original QIR and Improved QIR



Head Quarters Morris Town 5th April
1777

Sir

In order to shorten the march of the Massachusetts Regiments intended for this quarter, they are directed to take their Rout thro' the Green woods to Kinderhook, Claverack or Red Hook from whence they are to fall down to [illegible] Montgomery by Water. If none of the Enemy's Vessels should be in Haverstraw Bay, they may proceed further down the River by water and ~~land~~ disembark at Peekskill or in Jersey as there may be occasion. ~~That the~~ When the Sloops with the Troops arrive at Fort Montgomery, you are immediately to dispatch two Boats to Haverstraw to see if any of the Enemy's Shipping are in the Bay, if there are not, the one of the Boats to return and report to you, the other to remain in the Bay near Verplanks Point. The Vessels then to go down with the tide of Ebb and land the Troops ~~at Haverstraw~~ where directed. Signals to be fixed upon and given by [illegible] Boat that remains below, in Case of any the Enemy's Vessels should be seen coming up, whilst ours are going down—

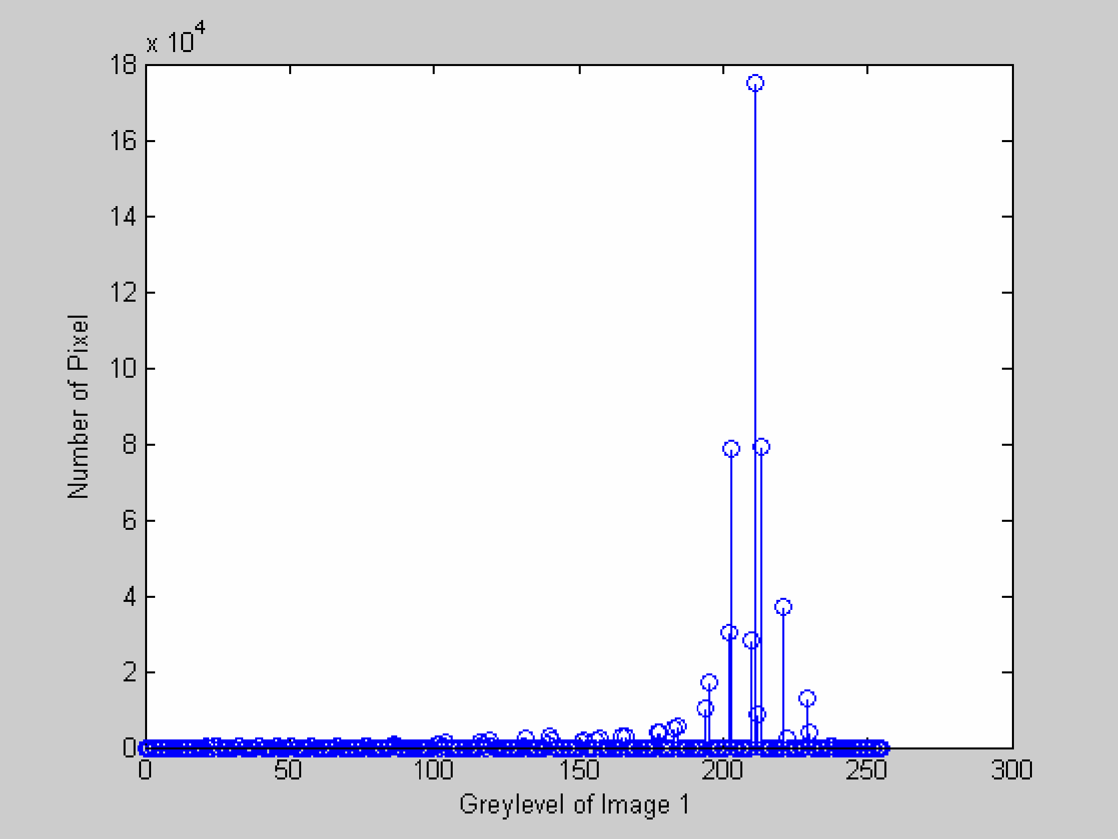
If you should have occasion to ~~leave~~ be absent from the Garrison for ever so short a time, be sure to leave a Copy of these orders with the next in command.

I am &c

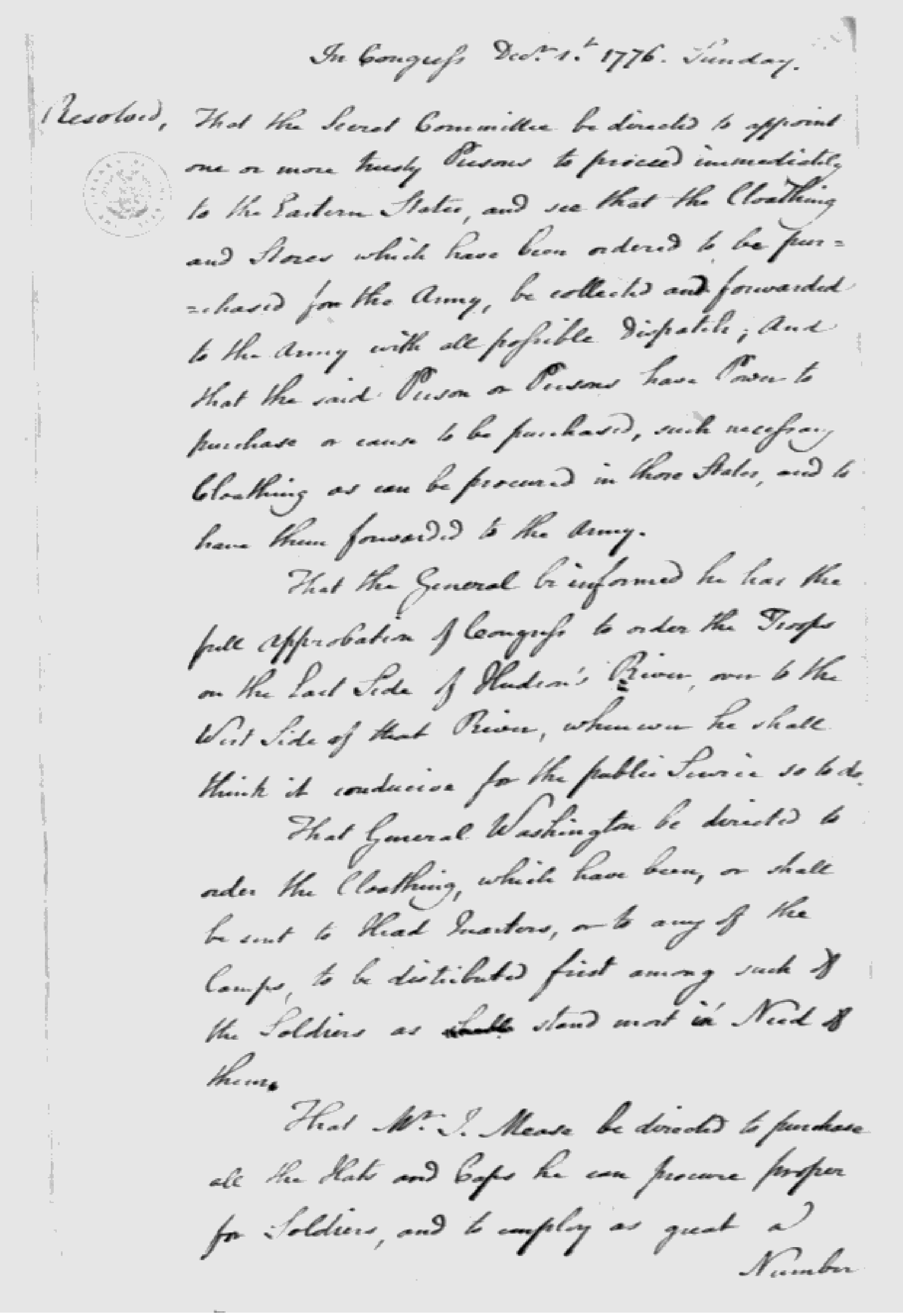
Brig Genl James Clinton
or commanding Officer at
Fort Montgomery

B-3-1

Image 1



Histogram of Image 1



In Congreſs Dec^{r} 1st 1776. Sunday.

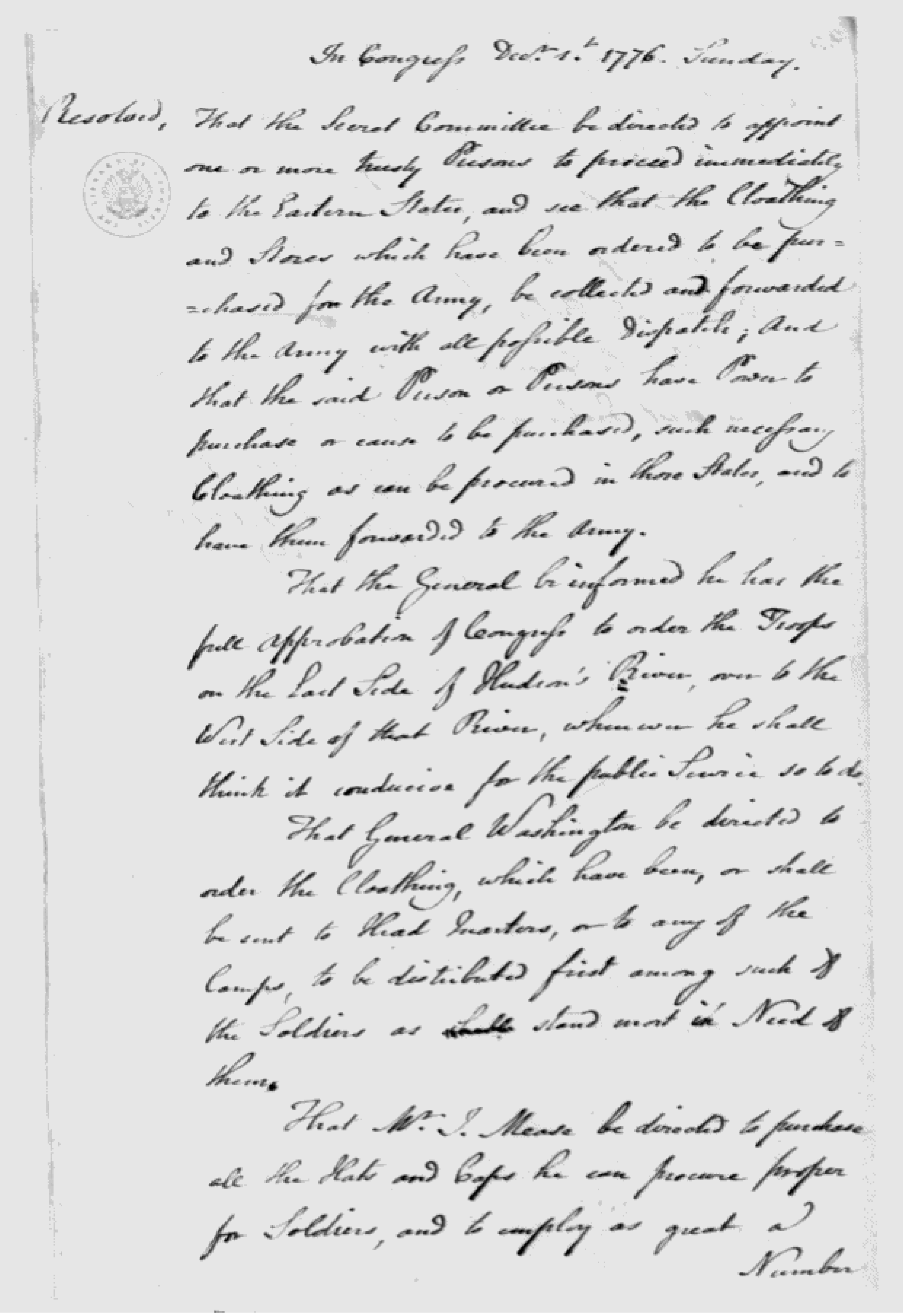
Resolved, That the Secret Committee be directed to appoint one or more trusty Persons to proceed immediately to the Eastern States, and see that the Cloathing and Stores which have been ordered to be purchased for the Army, be collected and forwarded to the Army with all poſsible dispatch; And that the said Person or Persons have Power to purchase or cause to be purchased, such neceſsary Cloathing as can be procured in those States, and to have them forwarded to the Army.

That the General be informed he has the full approbation of Congreſs to order the Troops on the East Side of Hudson's River, over to the West Side of that River, whenever he shall think it conducive for the public Service so to do.

That General Washington be directed to order the Cloathing, which have been, or shall be sent to Head Quarters, or to any of the Camps, to be distributed first among such of the Soldiers as ~~shall~~ stand most in Need of them.

That M^{r}. J. Mease be directed to purchase all the Hats and Caps he can procure proper for Soldiers, and to employ as great a Number

Original QIR



In Congreſs Decr. 1st 1776. Sunday.

Resolved, That the Secret Committee be directed to appoint one or more trusty Persons to proceed immediately to the Eastern States, and see that the Cloathing and Stores which have been ordered to be purchased for the Army, be collected and forwarded to the Army with all poſsible dispatch; And that the said Person or Persons have Power to purchase or cause to be purchased, such neceſsary Cloathing as can be procured in those States, and to have them forwarded to the Army.

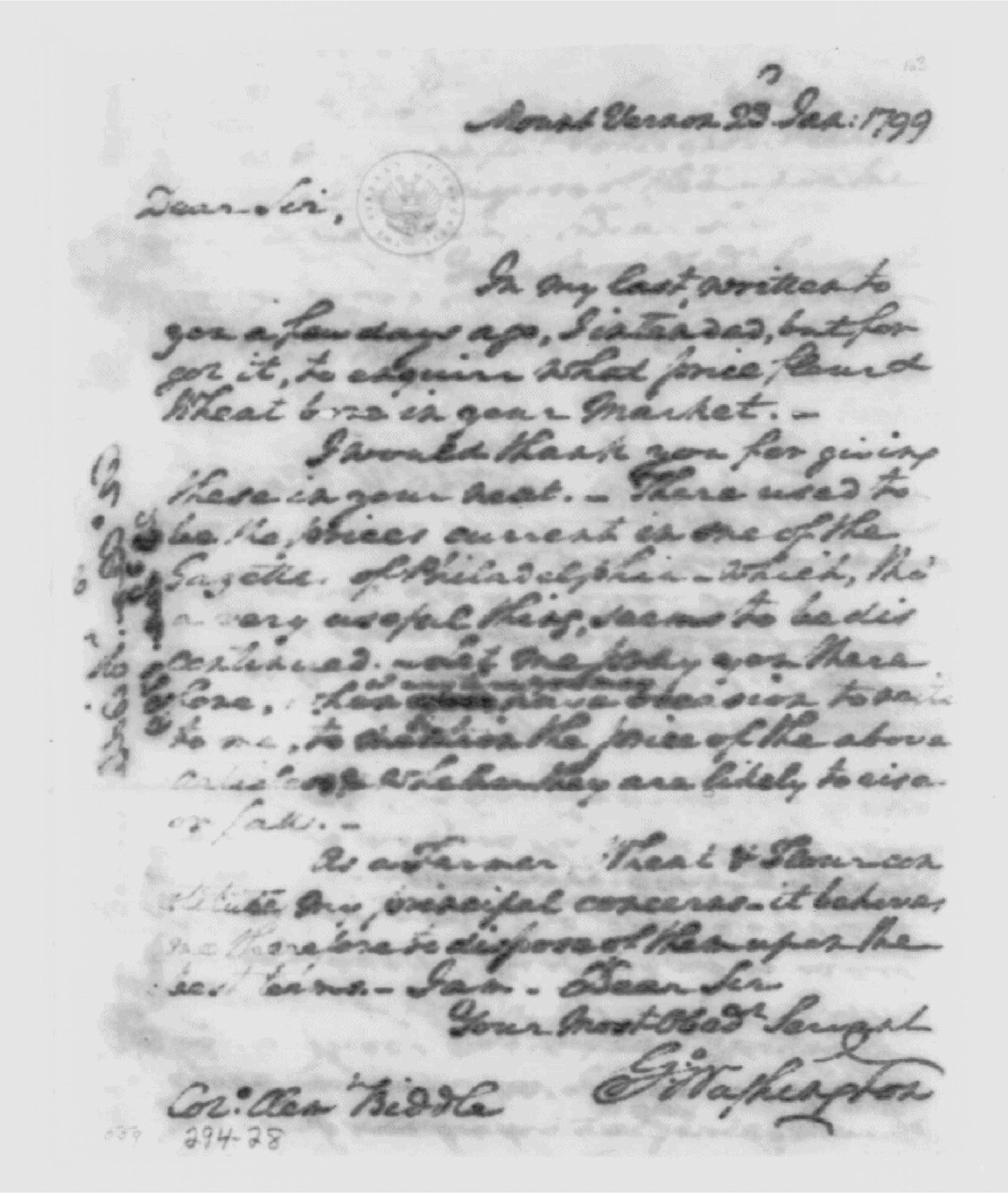
That the General be informed he has the full approbation of Congreſs to order the Troops on the East Side of Hudson's River, over to the West Side of that River, whenever he shall think it conducive for the public Service so to do.

That General Washington be directed to order the Cloathing, which have been, or shall be sent to Head Quarters, or to any of the Camps, to be distributed first among such of the Soldiers as ~~shall~~ stand most in Need of them.

That Mr. J. Mease be directed to purchase all the Hats and Caps he can procure proper for Soldiers, and to employ as great a

Number

Improved QIR



Mount Vernon 23 Jan: 1799

Dear Sir,

In my last, written to
you a few days ago, I intended, but for-
got it, to enquire what price Flour &
Wheat bore in your Market. —

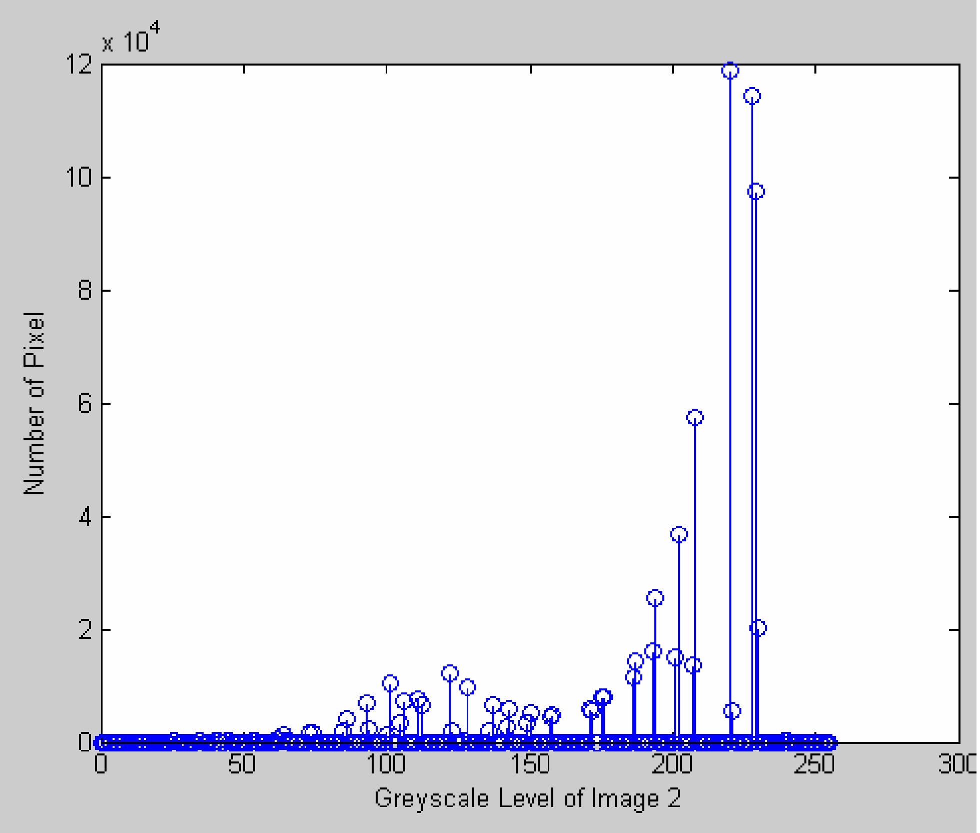
I would thank you for giving
these in your next. — There used to
be the prices current in one of the
Gazettes of Philadelphia — which, tho'
a very useful thing, seems to be dis-
continued. — Let me pray you there-
fore, when ~~you~~ have other occasion to write
to me, to mention the price of the above
articles, & whether they are likely to rise
or fall. —

As a Farmer Wheat & Flour con-
stitute my principal concerns — it behoves
me therefore to dispose of them upon the
best terms. — I am — Dear Sir
Your Most Obed. Servant
Go: Washington

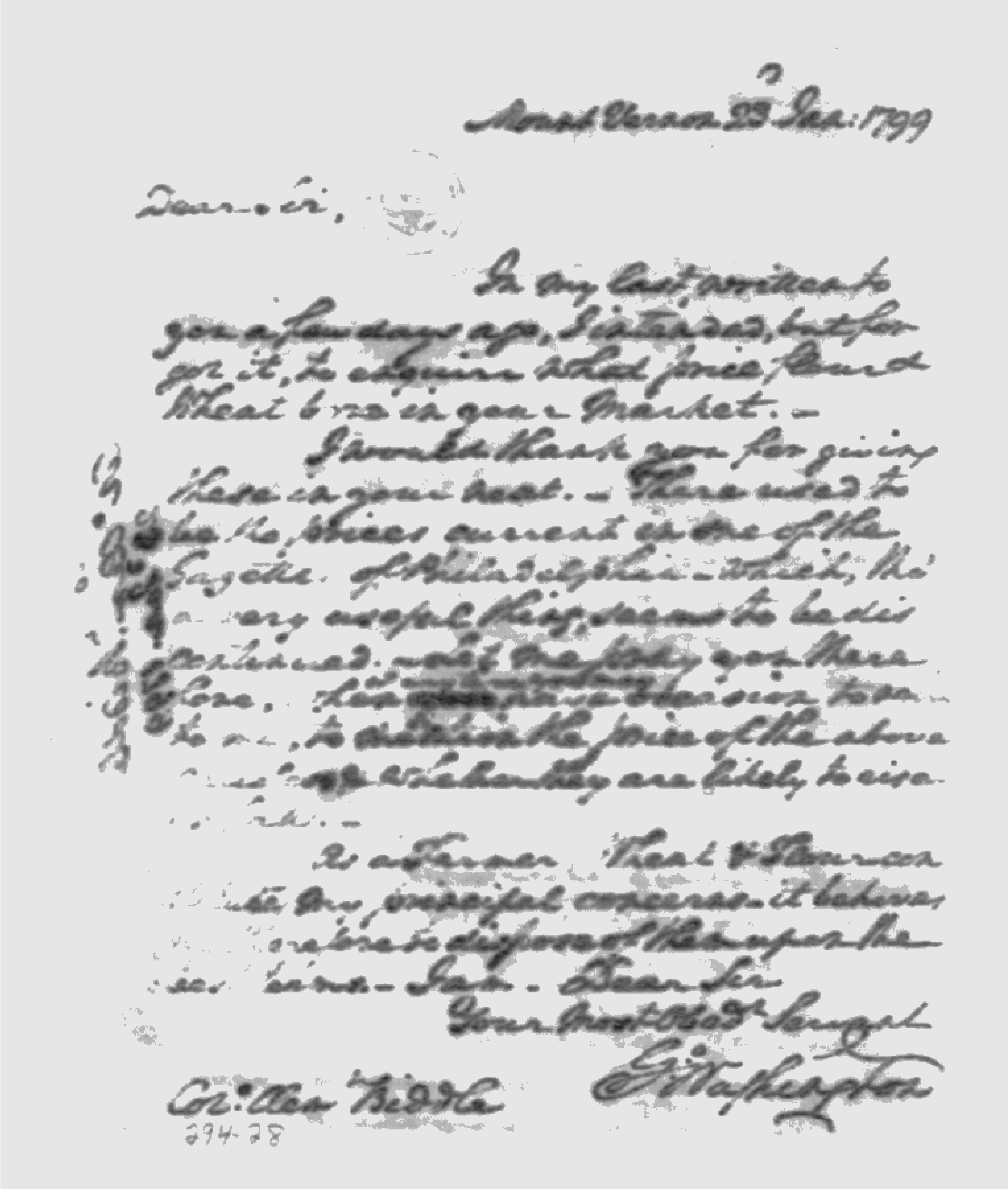
Col: Clem Biddle

294-28

Image 2



Histogram of Image 2



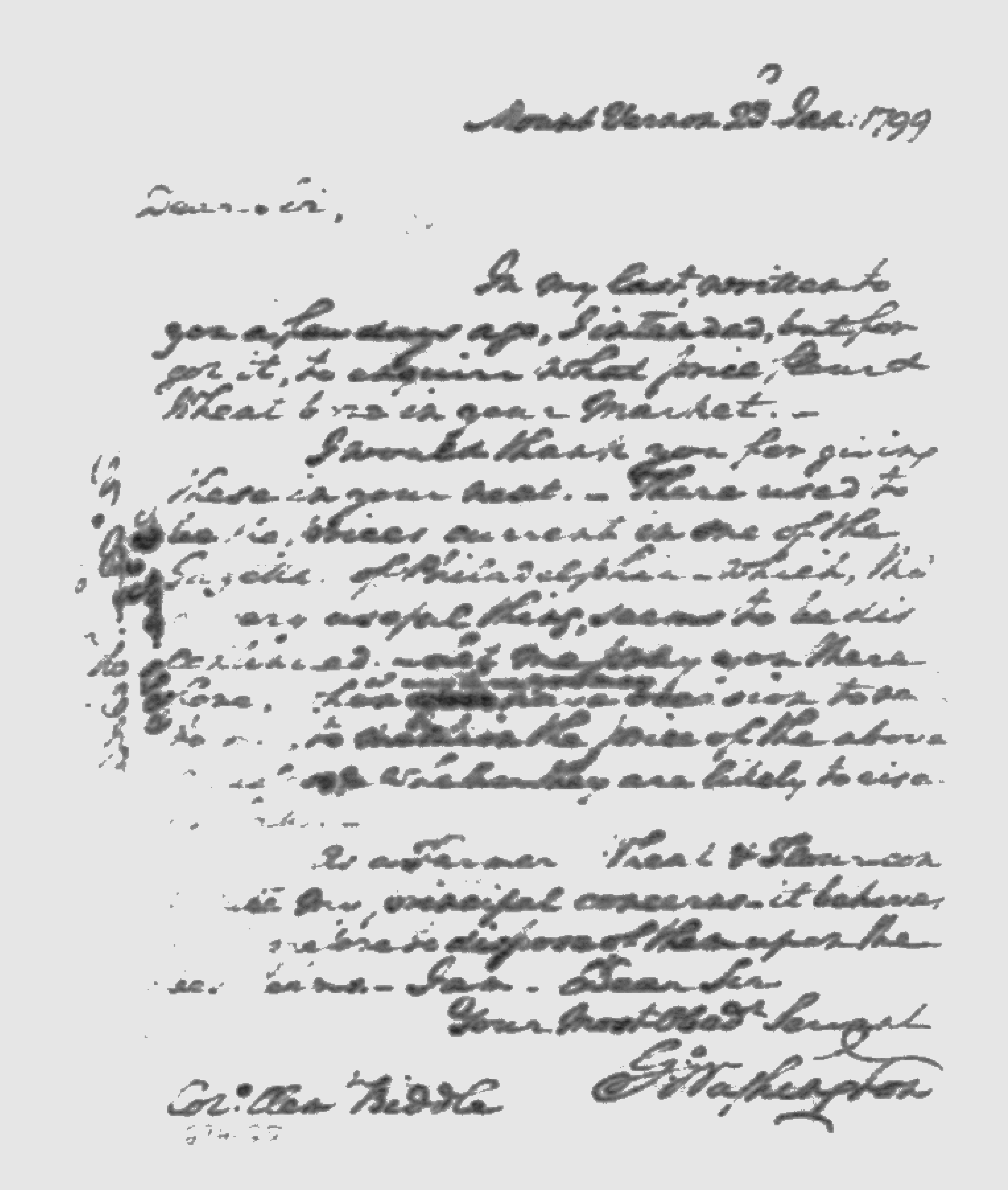
Mount Vernon 23^{d} Dec: 1799

Dear Sir,

In my last, written to
you a few days ago, I intended, but for
got it, to enquire what price Flour &
Wheat bore in your Market. —
I would thank you for giving
these in your next. — There used to
be the prices current in one of the
Gazettes of Philadelphia — which, tho'
a very useful thing, seems to be dis
continued. — and if the thing you there
before, [illegible] have been seen to
to me, to [illegible] the price of the above
[illegible] when they are likely to rise
[illegible]. —
As a Farmer Wheat & Flour con
stitute my principal concerns — it behoves
me to be enabled to dispose of them upon the
best terms. — I am — Dear Sir
Your most Obed^{t} Serv^{t}
G^{o} Washington

Col^{o} Clem Biddle
294-28

Original QIR



Mount Vernon 23 Jan: 1799

Dear Sir,

In my last written to you a few days ago, I intended, but forgot it, to enquire what price, Flour & Wheat bore in your Market. —

I would thank you for giving these in your next. — There used to be Tables, prices current in one of the Gazettes of Philadelphia — which, tho' a very useful thing, seems to be discontinued. — If the Tables were there before, [illegible] occasion to [illegible] to ascertain the price of the above [illegible] and whether they are likely to rise [illegible]

As a Farmer Wheat & Flour are my principal concerns — it behoves me to dispose of them upon the best terms — I am — Dear Sir

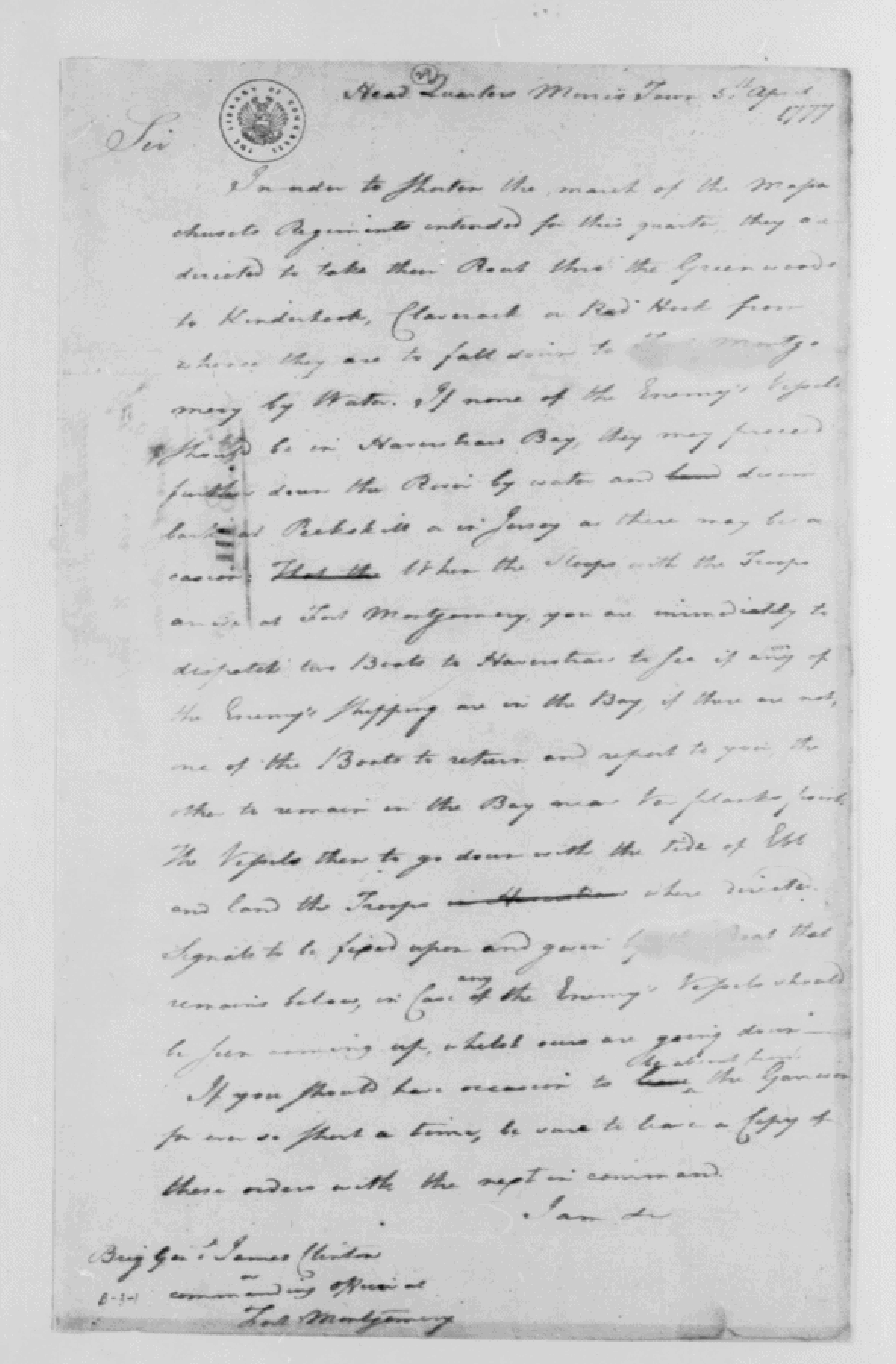
Your most Obed^t Servant

G^o Washington

Col^o Clem^t Biddle

Improved QIR

APPENDIX 2: Experimental Results of Mean-Gradient Method and other 3 Thresholding Methods on 4 Different Document Types



(9)

Head Quarters Morris Town 5th April
1777

Sir

In order to shorten the march of the Massachusets Regiments intended for this quarter, they are directed to take their Rout thro' the Green woods to Kinderhook, Claverack or Red Hook from whence they are to fall down to Fort Montgomery by Water. If none of the Enemy's Vessels should be in Haverstraw Bay, they may proceed further down the River by water and ~~land~~ disembark at Peekskill or in Jersey as there may be occasion. ~~That the~~ When the Sloops with the Troops arrive at Fort Montgomery, you are immediately to dispatch two Boats to Haverstraw to see if any of the Enemy's Shipping are in the Bay, if there are not, one of the Boats to return and report to you, the other to remain in the Bay near Verplanks Point. The Vessels then to go down with the tide of Ebb and land the Troops ~~at Haverstraw~~ where directed. Signals to be fixed upon and given [illegible] Boat that remains below, in Case any of the Enemy's Vessels should be seen coming up, whilst ours are going down.

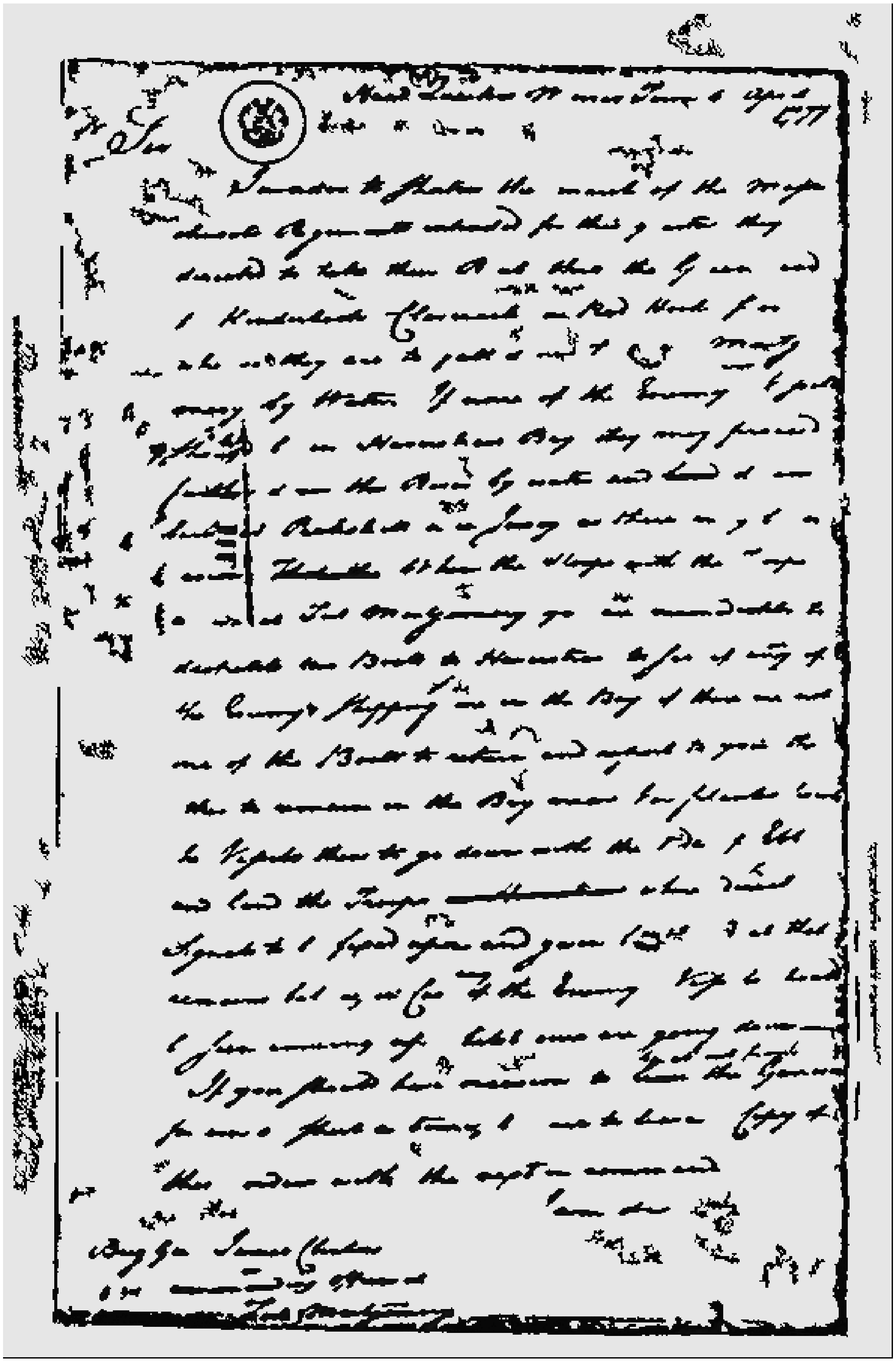
If you should have occasion to ~~leave~~ be absent from the Garrison for ever so short a time, be sure to leave a Copy of these orders with the next in command.

I am &c

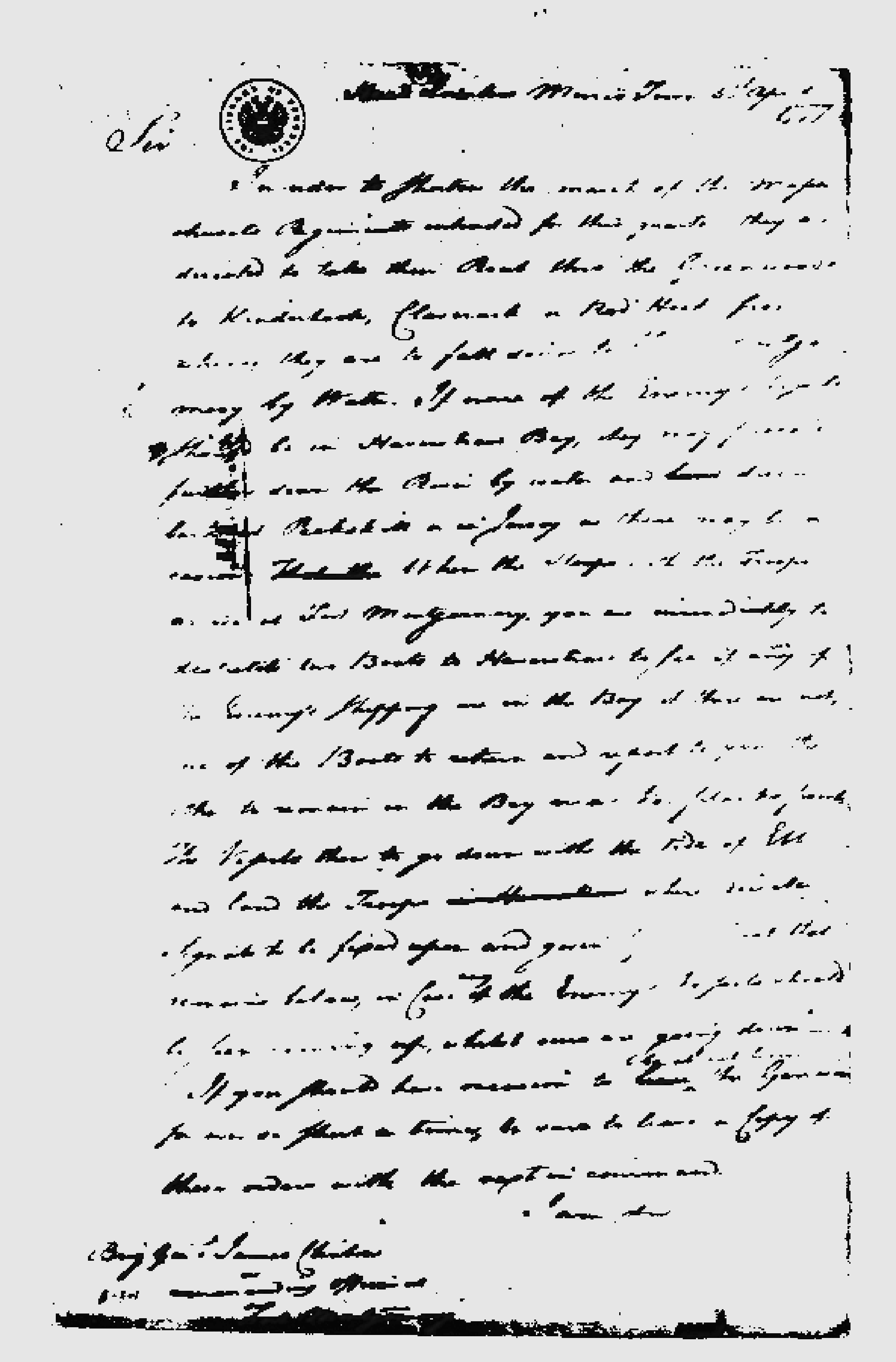
Brig. Genl. James Clinton
commanding Officer at
Fort Montgomery

B-3-1

Original Historical Image 1



Result of Improved Niblack's Algorithm



Head Quarters Morris Town 5th Apr 1777

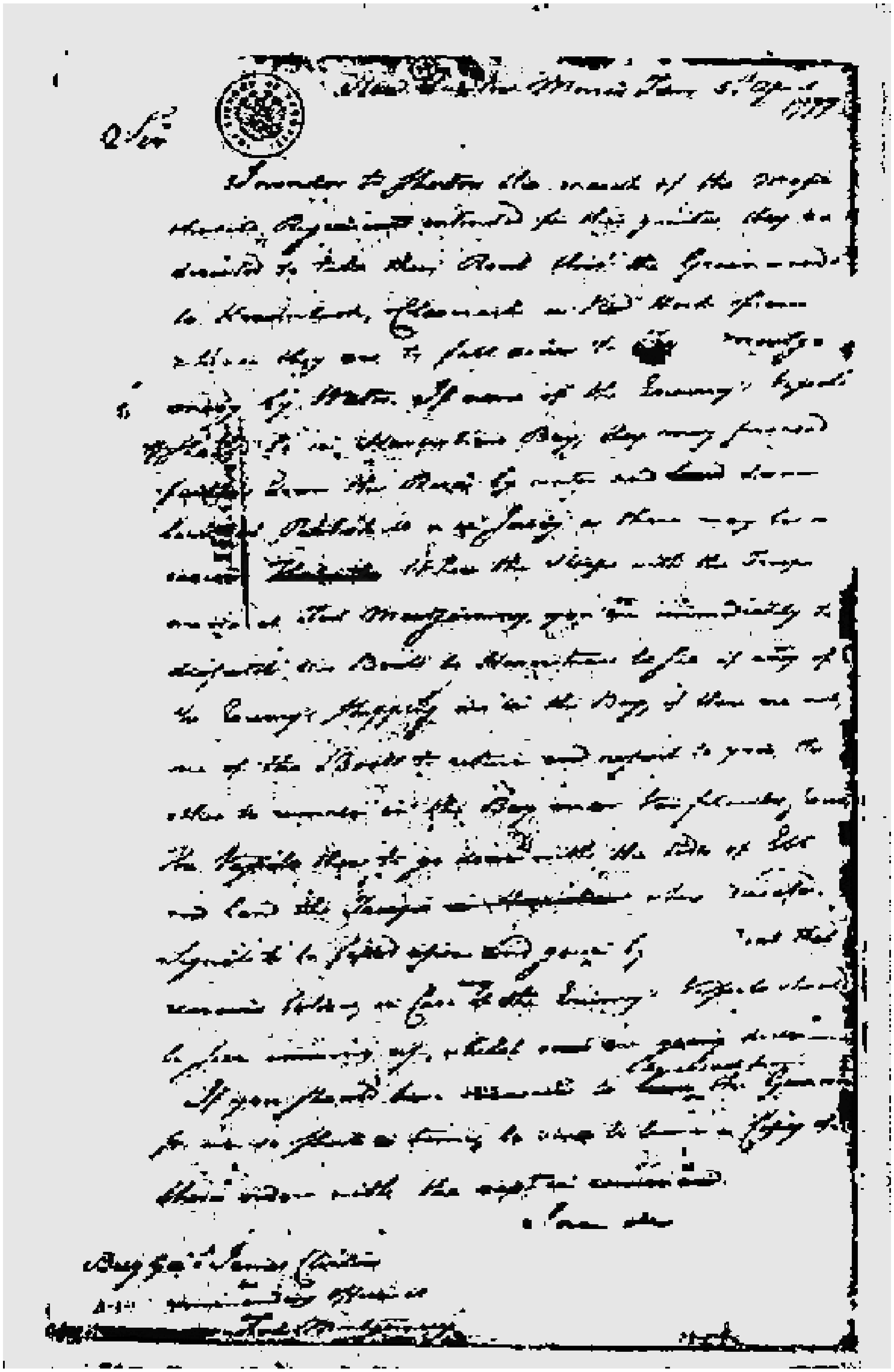
Sir

In order to shorten the march of the Massa-
chusets Regiments intended for this quarter they are
directed to take their Rout thro' the Greenwoods
to Kinderhook, Claverack or Red Hook from
whence they are to fall down to [illegible]
away by Water. If none of the Enemy's Ships
be in Haverstraw Bay, they may [illegible]
further down the River by water and land [illegible]
[illegible] Peekskill or wherever there may be a
[illegible] When the Sloops with the Troops
arrive at Fort Montgomery, you are immediately to
dispatch two Boats to Haverstraw to see if any of
the Enemy's Shipping are in the Bay. If there are not,
one of the Boats to return and report to you, the
other to remain in the Bay [illegible]
The Vessels then to go down with the Tide of Ebb
and land the Troops [illegible] where [illegible]
Signals to be fixed upon and given [illegible] that
remain below, in case of the Enemy's Vessels should
be seen coming up, whilst ours are going down [illegible]
If you should have occasion to [illegible] the Govern[illegible]
for men or should a [illegible] be sent to leave a Copy of
these orders with the next in command.

I am Sir

Brig. Genl. James Clinton
commanding Officer at

Result of Original QIR's Algorithm



Head Quarters Morris Town 5th April 1777

Sir

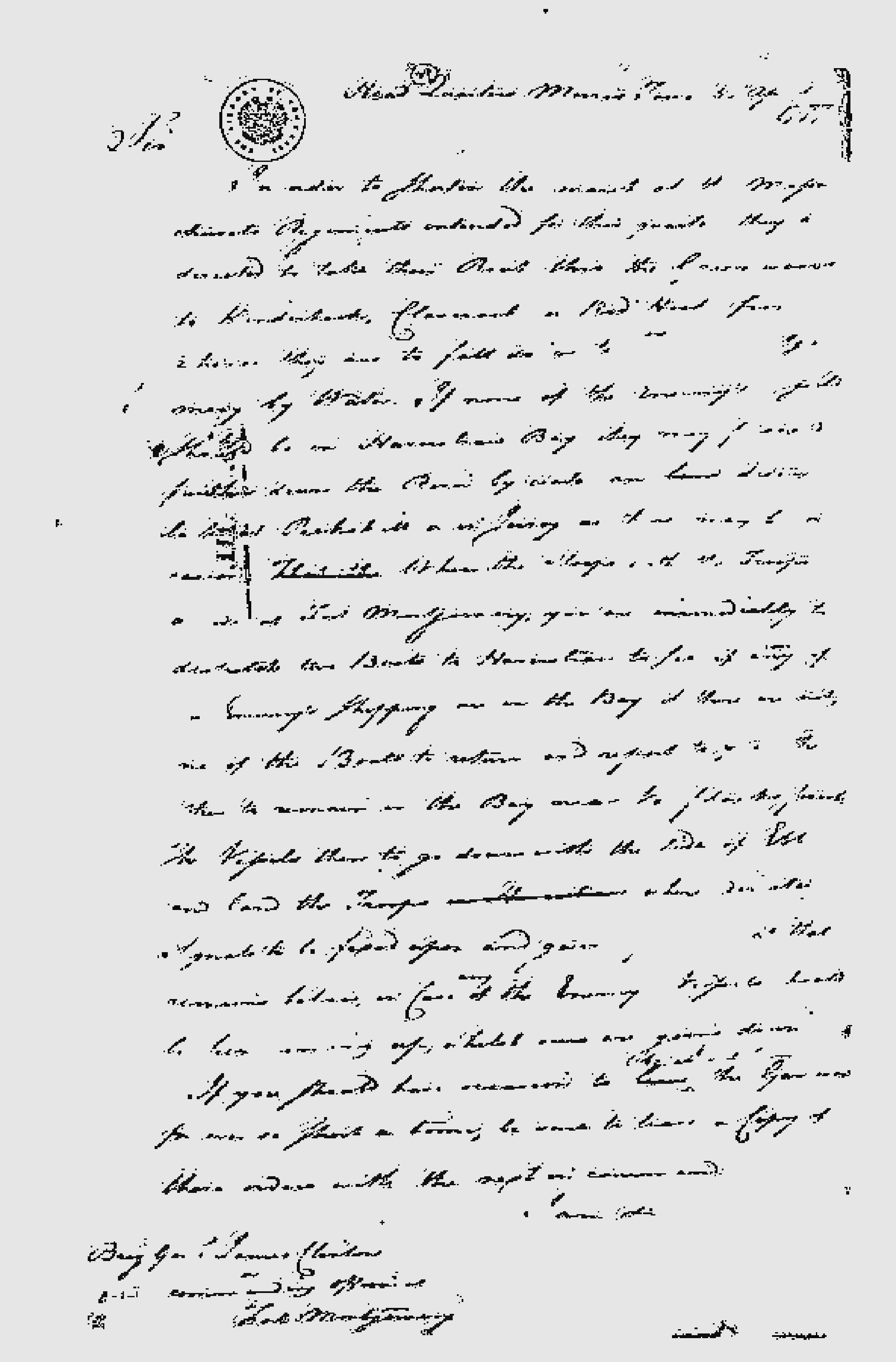
In order to shorten the march of the troops [illegible] Regiments ordered for this quarter, they are directed to take their Road thro' the Green woods to [illegible], [illegible] [illegible] [illegible] where they are to [illegible] [illegible] [illegible] by Water. If none of the Enemy's [illegible] [illegible] in [illegible] Bay, they may proceed [illegible] from the [illegible] by water and land [illegible] [illegible] is necessary as there may be [illegible] [illegible] the [illegible] with the Troops arrive at Fort Montgomery you are immediately to dispatch two Boats to [illegible] to see if any of the Enemy's Shipping are in the Bay, if there are not, one of the Boats to return and report to you, the other to remain in the Bay near [illegible] [illegible] the [illegible] then to go down with the tide of [illegible] and land the Troops ~~[illegible]~~ [illegible] [illegible] [illegible] to be fixed upon and given by [illegible] [illegible] in case the Enemy's [illegible] should be seen coming up, which would [illegible] [illegible]

If you should have occasion to [illegible] the [illegible] [illegible] [illegible] [illegible] [illegible] their orders with the utmost caution.

I am Sir

Brig Genl James Clinton
Commanding Officer at
Fort Montgomery

Result of Yanowitz's Algorithm



Head Quarters Morris Town 3d Feb 1777

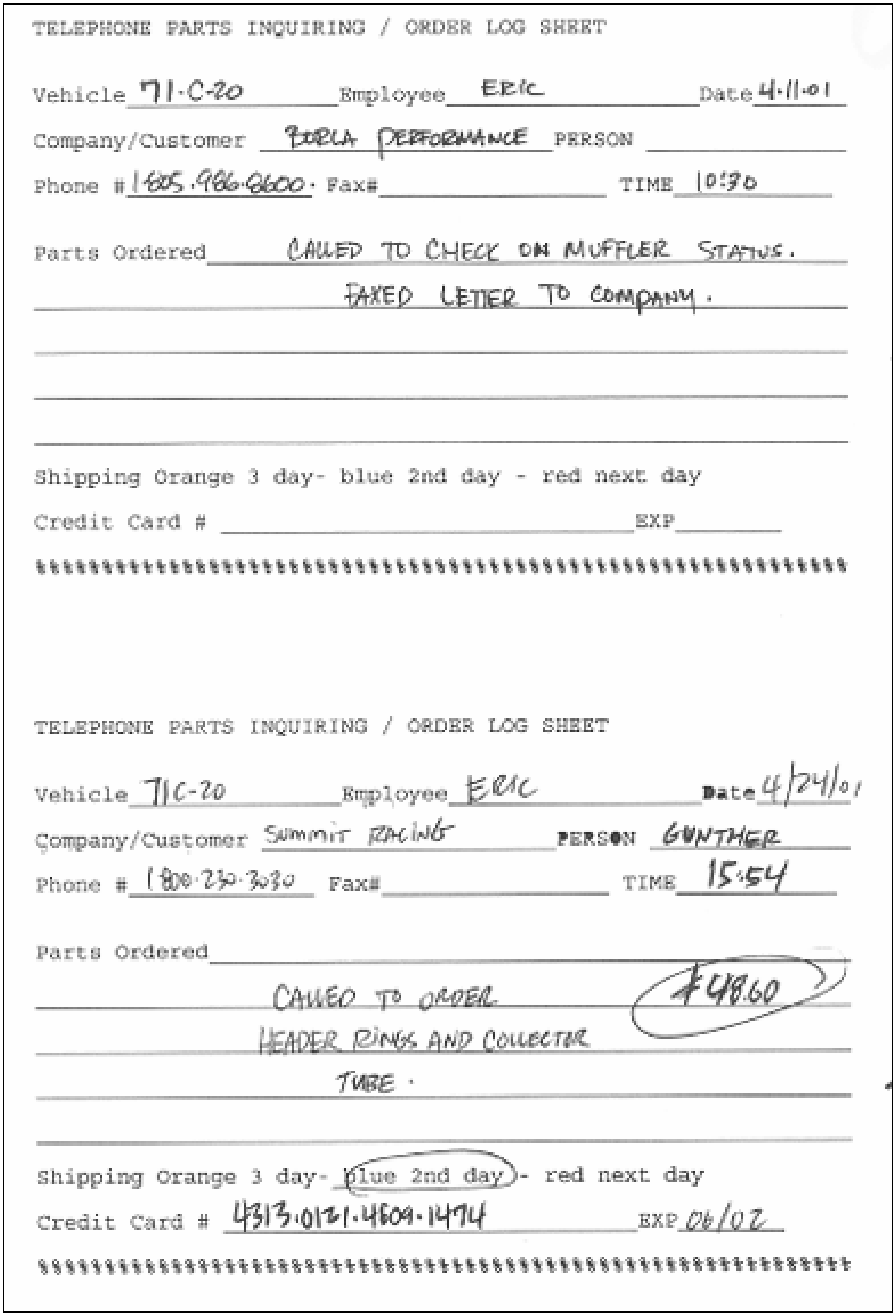
Sir

In order to shorten the march of the Massachusetts Regiments intended for this quarter they are directed to take their Rout thro' the [illegible] to Kinderhook, Claverack or Red Hook from whence they are to fall down to [illegible] by Water. If none of the Enemy's Ships be in Haverstraw Bay they may proceed further down the River by water or land [illegible] Peekskill is a Ferrying place [illegible] Thither. When the Troops [illegible] at Fort Montgomery, you are immediately to dispatch two Boats to Haverstraw to see if any of the Enemy's Shipping are in the Bay. If there are not, one of the Boats to return and report to you, the other to remain in the Bay near the [illegible]. The Vessels then to go down with the Tide of Ebb and land the Troops ~~at Haverstraw~~ where [illegible] that Signals be fixed upon and given [illegible] remain below in case of the Enemy's [illegible] be coming up, whilst ours are going down. If you should have occasion to [illegible] the Troops [illegible] for one or more a Convoy be sure to have a Copy of their orders with the [illegible] command.

I am Sir

Brig Genl James Clinton
commanding officer at
Fort Montgomery

Result of Mean-Gradient Algorithm



TELEPHONE PARTS INQUIRING / ORDER LOG SHEET

Vehicle 71-C-20 Employee ERIC Date 4-11-01

Company/Customer BORLA PERFORMANCE PERSON

Phone # 1605.986.8600. Fax# TIME 10:30

Parts Ordered CALLED TO CHECK ON MUFFLER STATUS.

FAXED LETTER TO COMPANY.

Shipping Orange 3 day- blue 2nd day - red next day

Credit Card # EXP

%%

TELEPHONE PARTS INQUIRING / ORDER LOG SHEET

Vehicle 71C-20 Employee ERIC Date 4/24/01

Company/Customer SUMMIT RACING PERSON GUNTHER

Phone # 1800.230.3030 Fax# TIME 15:54

Parts Ordered

CALLED TO ORDER $48.60

HEADER RINGS AND COLLECTOR

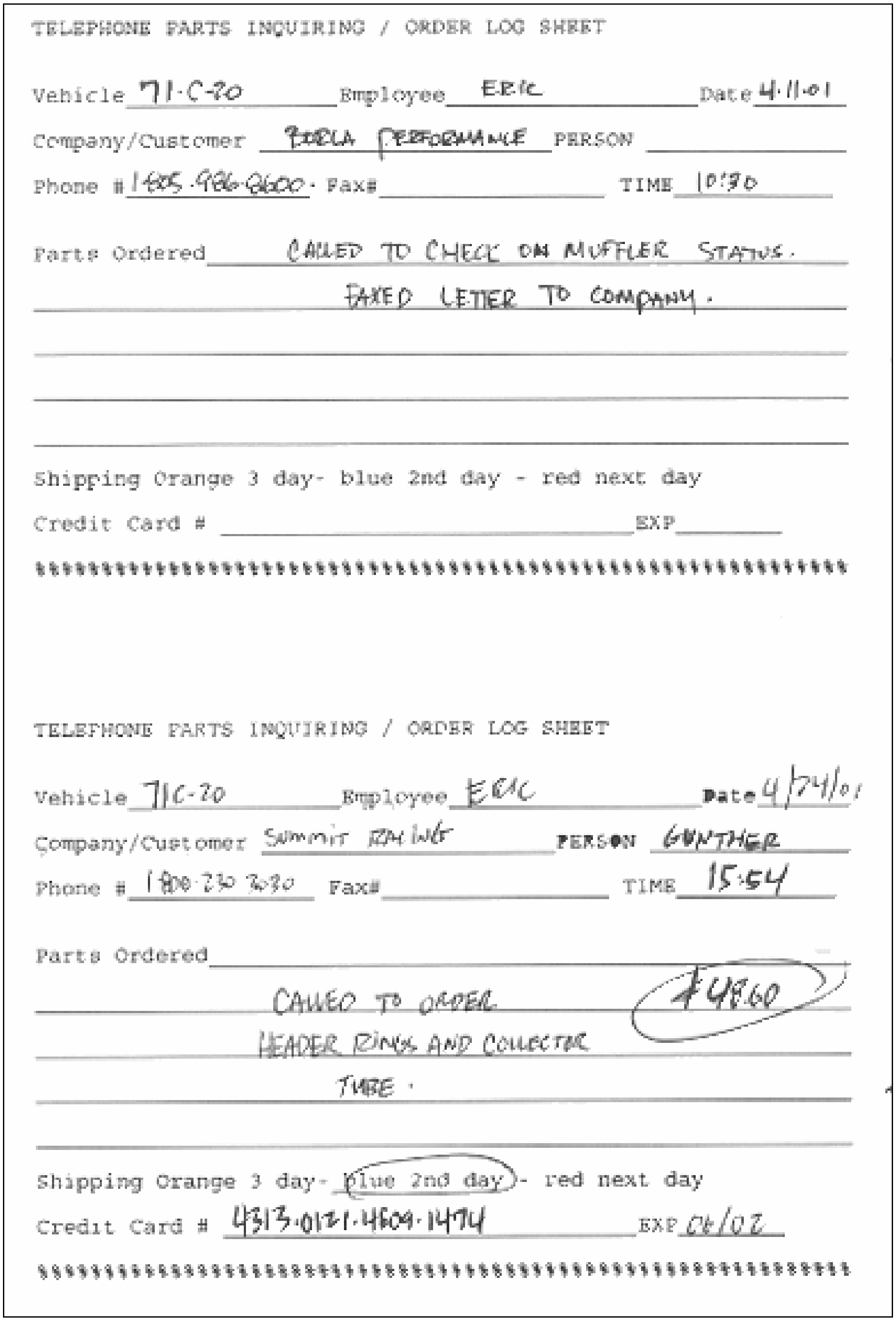
TUBE.

Shipping Orange 3 day- [✓ blue 2nd day] - red next day

Credit Card # 4313.0121.4609.1474 EXP 06/02

%%

Original Form Image 1



TELEPHONE PARTS INQUIRING / ORDER LOG SHEET

Vehicle 71·C-20 Employee ERIC Date 4·11·01

Company/Customer BORLA PERFORMANCE PERSON

Phone # 1·805·986·8600· Fax# TIME 10:30

Parts Ordered CALLED TO CHECK ON MUFFLER STATUS.

FAXED LETTER TO COMPANY.

Shipping Orange 3 day- blue 2nd day - red next day

Credit Card # EXP

$$

TELEPHONE PARTS INQUIRING / ORDER LOG SHEET

Vehicle 71C-20 Employee ERIC Date 4/24/01

Company/Customer SUMMIT RACING PERSON GUNTHER

Phone # 1800·230 3030 Fax# TIME 15:54

Parts Ordered

CALLED TO ORDER $48.60

HEADER RINGS AND COLLECTOR

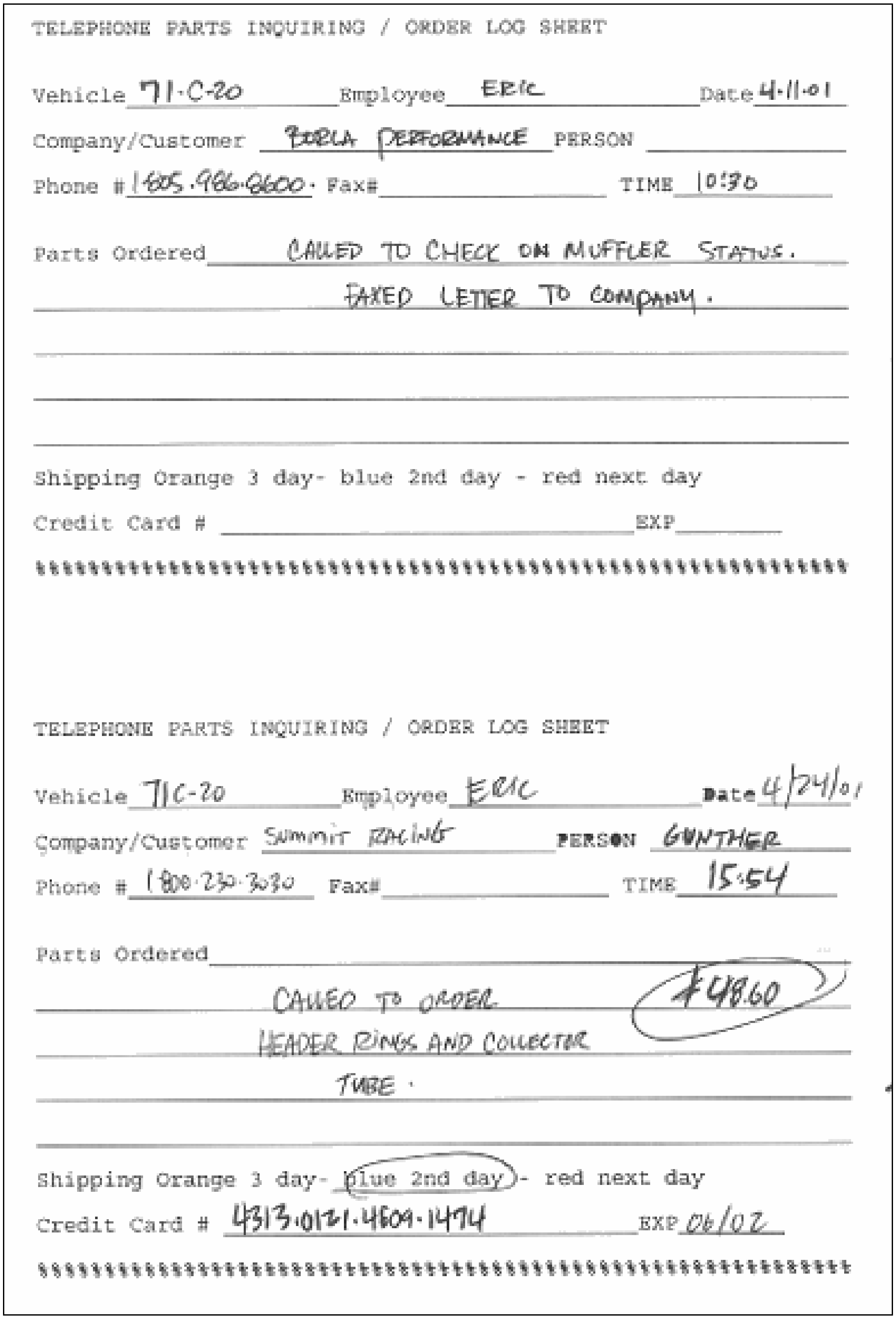
TUBE.

Shipping Orange 3 day- [✓ blue 2nd day] - red next day

Credit Card # 4313·0121·4609·1474 EXP 06/02

$$

Result of Improved Niblack's Algorithm



TELEPHONE PARTS INQUIRING / ORDER LOG SHEET

Vehicle 71·C-20 Employee ERIC Date 4·11·01

Company/Customer BORLA PERFORMANCE PERSON

Phone # 1·805·986·8600· Fax# TIME 10:30

Parts Ordered CALLED TO CHECK ON MUFFLER STATUS.

FAXED LETTER TO COMPANY.

Shipping Orange 3 day- blue 2nd day - red next day

Credit Card # EXP

%%%

TELEPHONE PARTS INQUIRING / ORDER LOG SHEET

Vehicle 71C-20 Employee ERIC Date 4/24/01

Company/Customer SUMMIT RACING PERSON GUNTHER

Phone # 1800·230·3030 Fax# TIME 15:54

Parts Ordered

CALLED TO ORDER $48.60

HEADER RINGS AND COLLECTOR

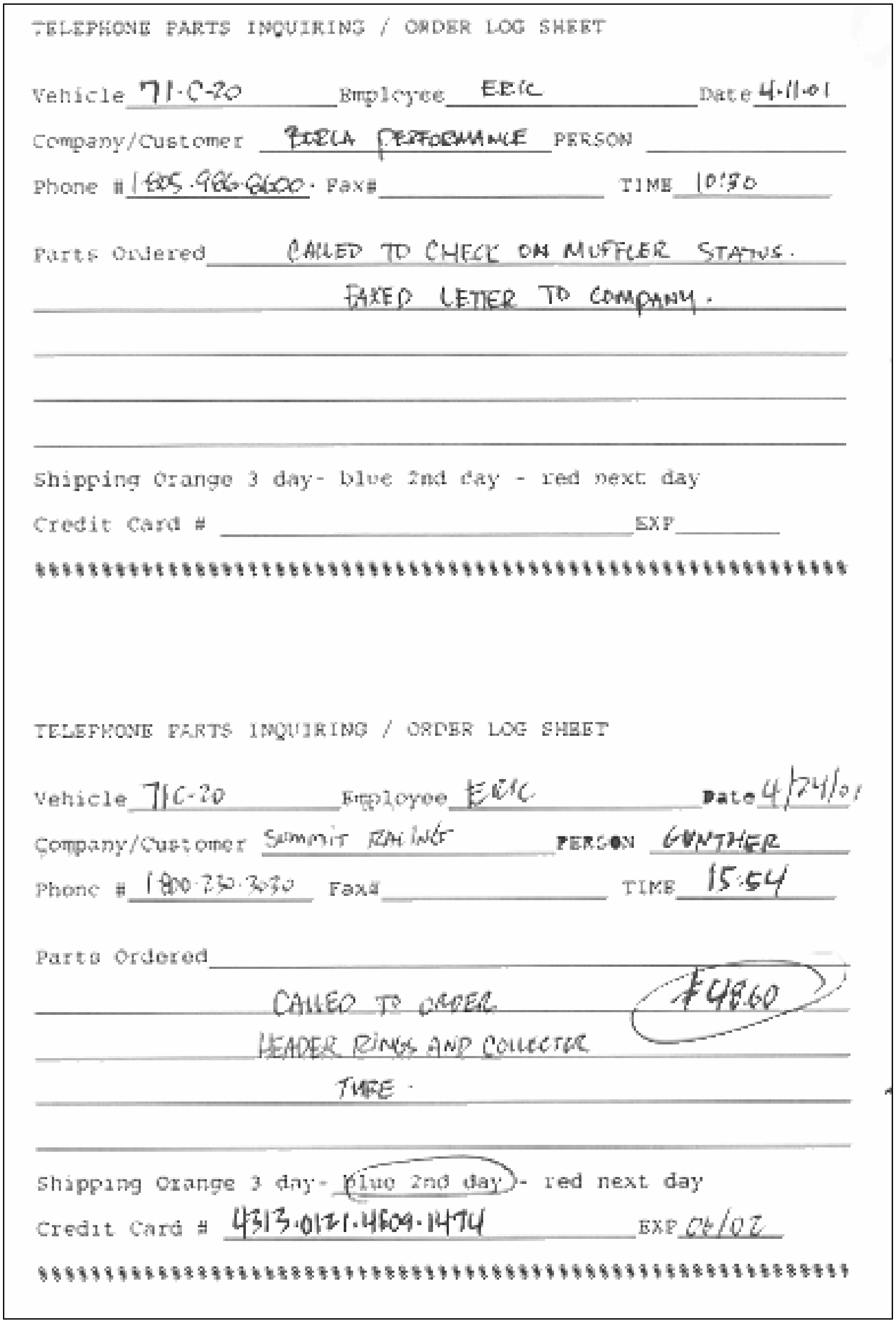
TUBE.

Shipping Orange 3 day- [✓ blue 2nd day]- red next day

Credit Card # 4313·0121·4609·1474 EXP 06/02

%%%

Result of Original QIR's Algorithm



TELEPHONE PARTS INQUIRING / ORDER LOG SHEET

Vehicle 71-C-20 Employee ERIC Date 4-11-01

Company/Customer BIRLA PERFORMANCE PERSON

Phone # 1 805.986.9600. Fax# TIME 10:30

Parts Ordered CALLED TO CHECK ON MUFFLER STATUS.

FAXED LETTER TO COMPANY.

Shipping Orange 3 day- blue 2nd day - red next day

Credit Card # EXP

$$$

TELEPHONE PARTS INQUIRING / ORDER LOG SHEET

Vehicle 71C-20 Employee ERIC Date 4/24/01

Company/Customer Summit RACING PERSON GUNTHER

Phone # 1 800.230.3030 Fax# TIME 15:54

Parts Ordered

CALLED TO ORDER $48.60

HEADER RINGS AND COLLECTOR

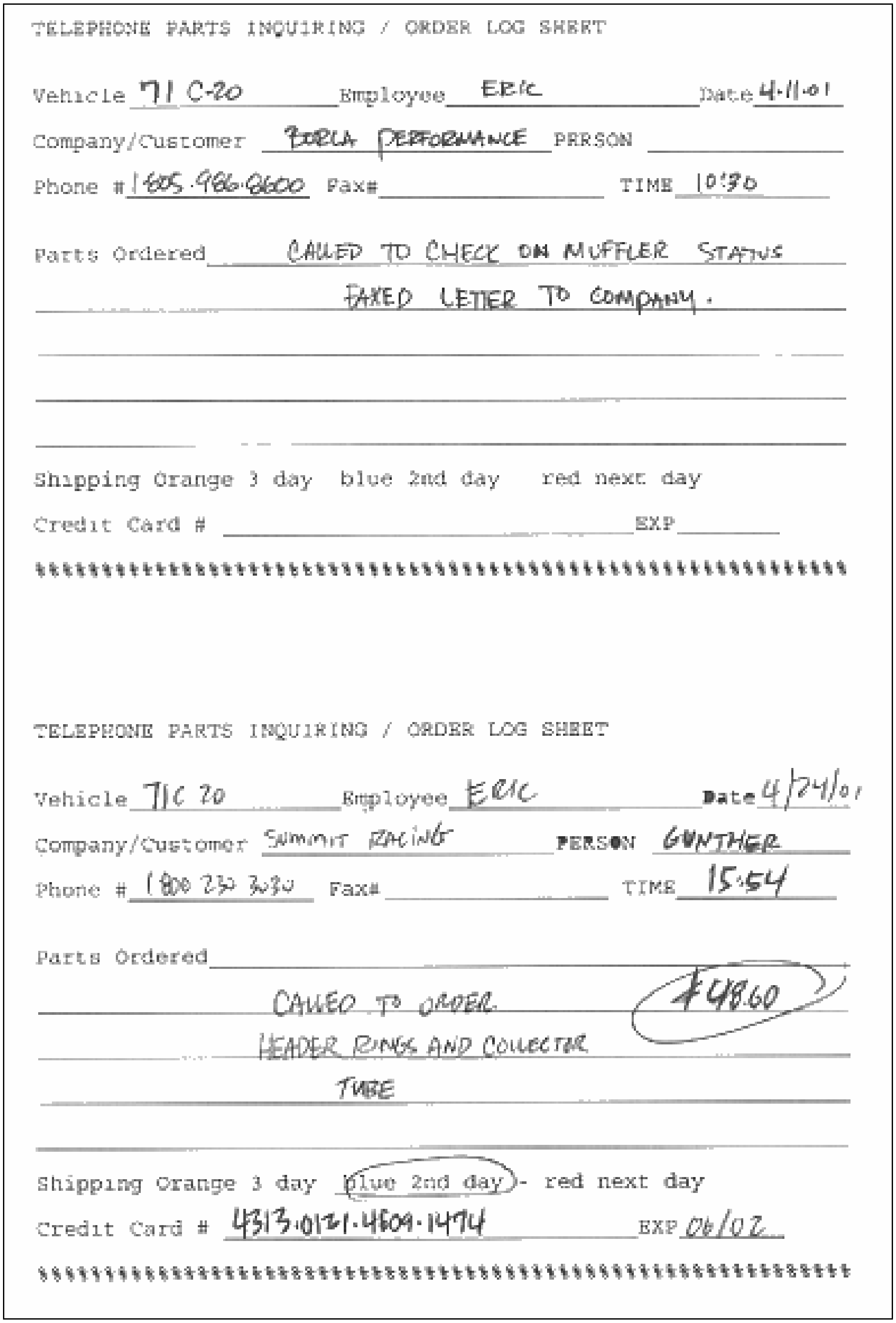
TUBE.

Shipping Orange 3 day- [✓ blue 2nd day] - red next day

Credit Card # 4313.0121.4609.1474 EXP 06/02

$$$

Result of Yanowitz's Algorithm



TELEPHONE PARTS INQUIRING / ORDER LOG SHEET

Vehicle 71 C-20 Employee ERIC Date 4-11-01

Company/Customer BORLA PERFORMANCE PERSON

Phone # 1 805 986 8600 Fax# TIME 10:30

Parts Ordered CALLED TO CHECK ON MUFFLER STATUS

FAXED LETTER TO COMPANY.

Shipping Orange 3 day blue 2nd day red next day

Credit Card # EXP

```
$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$
```

TELEPHONE PARTS INQUIRING / ORDER LOG SHEET

Vehicle 71C 20 Employee ERIC Date 4/24/01

Company/Customer SUMMIT RACING PERSON GUNTHER

Phone # 1 800 230 3030 Fax# TIME 15:54

Parts Ordered

CALLED TO ORDER $48.60

HEADER RINGS AND COLLECTOR

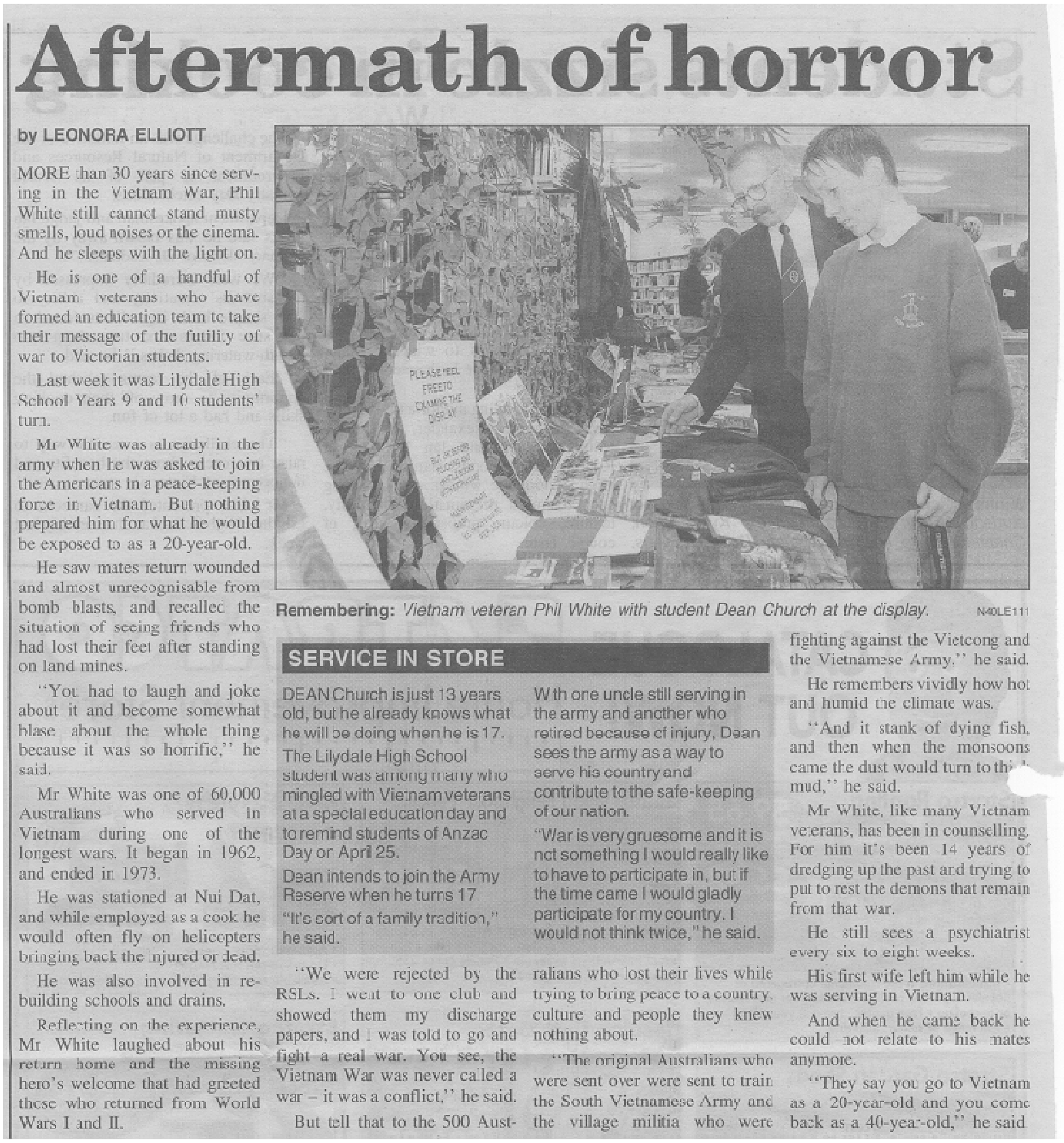
TUBE

Shipping Orange 3 day [✓ blue 2nd day] - red next day

Credit Card # 4313-0121-4609-1474 EXP 06/02

```
$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$
```

Result of Mean-Gradient Algorithm


PLEASE FEEL FREE TO EXAMINE THE DISPLAY

Aftermath of horror

by LEONORA ELLIOTT

Remembering: *Vietnam veteran Phil White with student Dean Church at the display.* N40LE111

MORE than 30 years since serving in the Vietnam War, Phil White still cannot stand musty smells, loud noises or the cinema. And he sleeps with the light on.

He is one of a handful of Vietnam veterans who have formed an education team to take their message of the futility of war to Victorian students.

Last week it was Lilydale High School Years 9 and 10 students' turn.

Mr White was already in the army when he was asked to join the Americans in a peace-keeping force in Vietnam. But nothing prepared him for what he would be exposed to as a 20-year-old.

He saw mates return wounded and almost unrecognisable from bomb blasts, and recalled the situation of seeing friends who had lost their feet after standing on land mines.

"You had to laugh and joke about it and become somewhat blase about the whole thing because it was so horrific," he said.

Mr White was one of 60,000 Australians who served in Vietnam during one of the longest wars. It began in 1962, and ended in 1973.

He was stationed at Nui Dat, and while employed as a cook he would often fly on helicopters bringing back the injured or dead.

He was also involved in rebuilding schools and drains.

Reflecting on the experience, Mr White laughed about his return home and the missing hero's welcome that had greeted those who returned from World Wars I and II.

"We were rejected by the RSLs. I went to one club and showed them my discharge papers, and I was told to go and fight a real war. You see, the Vietnam War was never called a war – it was a conflict," he said.

But tell that to the 500 Australians who lost their lives while trying to bring peace to a country, culture and people they knew nothing about.

"The original Australians who were sent over were sent to train the South Vietnamese Army and the village militia who were fighting against the Vietcong and the Vietnamese Army," he said.

He remembers vividly how hot and humid the climate was.

"And it stank of dying fish, and then when the monsoons came the dust would turn to thi[illegible] mud," he said.

Mr White, like many Vietnam veterans, has been in counselling. For him it's been 14 years of dredging up the past and trying to put to rest the demons that remain from that war.

He still sees a psychiatrist every six to eight weeks.

His first wife left him while he was serving in Vietnam.

And when he came back he could not relate to his mates anymore.

"They say you go to Vietnam as a 20-year-old and you come back as a 40-year-old," he said

SERVICE IN STORE

DEAN Church is just 13 years old, but he already knows what he will be doing when he is 17.

The Lilydale High School student was among many who mingled with Vietnam veterans at a special education day and to remind students of Anzac Day or April 25.

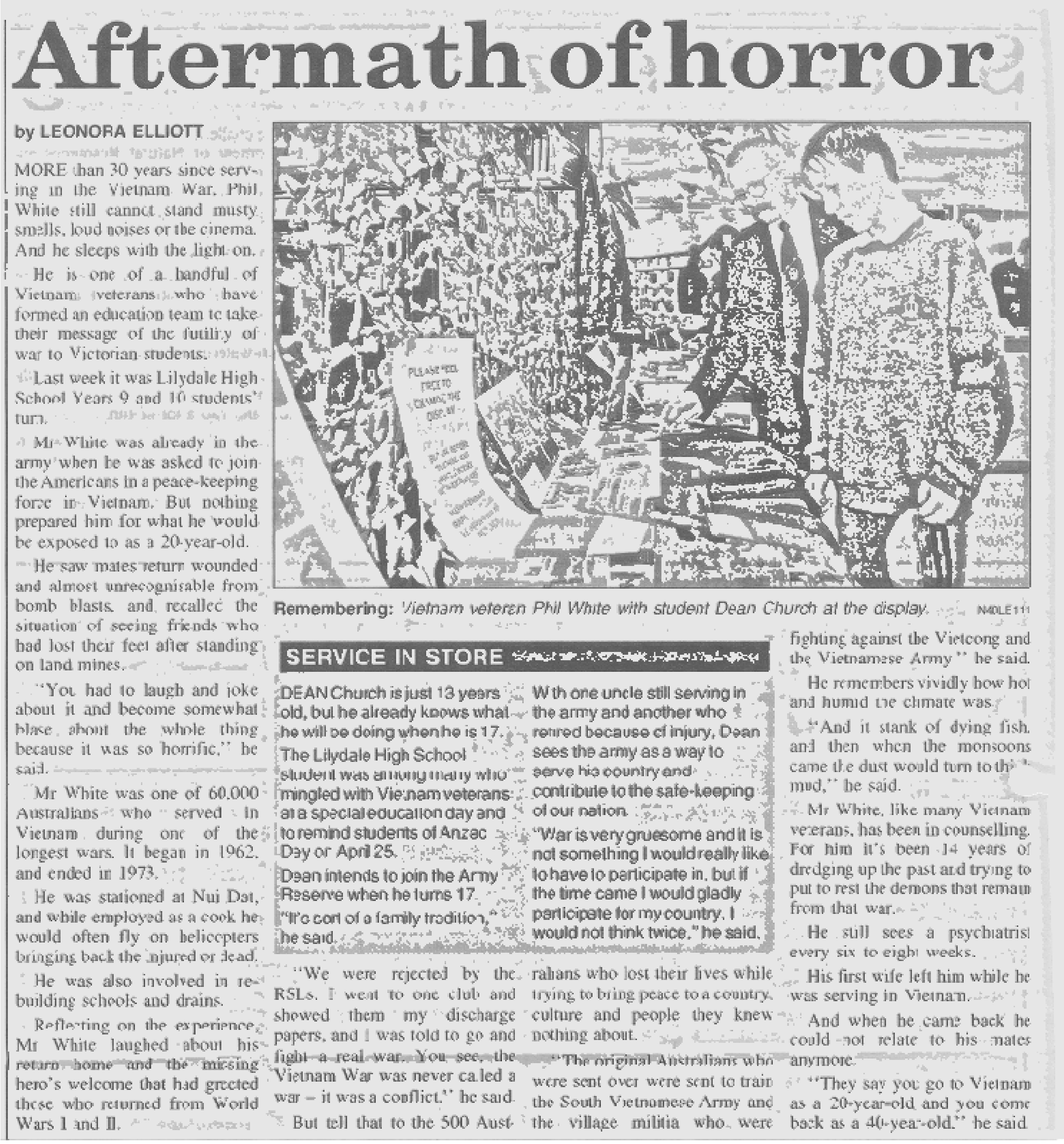
Dean intends to join the Army Reserve when he turns 17

"It's sort of a family tradition," he said.

With one uncle still serving in the army and another who retired because of injury, Dean sees the army as a way to serve his country and contribute to the safe-keeping of our nation.

"War is very gruesome and it is not something I would really like to have to participate in, but if the time came I would gladly participate for my country. I would not think twice," he said.

Original Newspaper Image 1



Aftermath of horror

by LEONORA ELLIOTT

MORE than 30 years since serving in the Vietnam War, Phil White still cannot stand musty smells, loud noises or the cinema. And he sleeps with the light on.

He is one of a handful of Vietnam veterans who have formed an education team to take their message of the futility of war to Victorian students.

Last week it was Lilydale High School Years 9 and 10 students' turn.

Mr White was already in the army when he was asked to join the Americans in a peace-keeping force in Vietnam. But nothing prepared him for what he would be exposed to as a 20-year-old.

He saw mates return wounded and almost unrecognisable from bomb blasts, and recalled the situation of seeing friends who had lost their feet after standing on land mines.

"You had to laugh and joke about it and become somewhat blase about the whole thing because it was so horrific," he said.

Mr White was one of 60,000 Australians who served in Vietnam during one of the longest wars. It began in 1962, and ended in 1973.

He was stationed at Nui Dat, and while employed as a cook he would often fly on helicopters bringing back the injured or dead.

He was also involved in rebuilding schools and drains.

Reflecting on the experience, Mr White laughed about his return home and the missing hero's welcome that had greeted those who returned from World Wars I and II.

Remembering: *Vietnam veteran Phil White with student Dean Church at the display.* N40LE111

SERVICE IN STORE

DEAN Church is just 13 years old, but he already knows what he will be doing when he is 17.

The Lilydale High School student was among many who mingled with Vietnam veterans at a special education day and to remind students of Anzac Day or April 25.

Dean intends to join the Army Reserve when he turns 17.

"It's sort of a family tradition," he said.

With one uncle still serving in the army and another who retired because of injury, Dean sees the army as a way to serve his country and contribute to the safe-keeping of our nation.

"War is very gruesome and it is not something I would really like to have to participate in, but if the time came I would gladly participate for my country. I would not think twice," he said.

"We were rejected by the RSLs. I went to one club and showed them my discharge papers, and I was told to go and fight a real war. You see, the Vietnam War was never called a war – it was a conflict," he said.

But tell that to the 500 Australians who lost their lives while trying to bring peace to a country, culture and people they knew nothing about.

"The original Australians who were sent over were sent to train the South Vietnamese Army and the village militia who were fighting against the Vietcong and the Vietnamese Army," he said.

He remembers vividly how hot and humid the climate was.

"And it stank of dying fish, and then when the monsoons came the dust would turn to thick mud," he said.

Mr White, like many Vietnam veterans, has been in counselling. For him it's been 14 years of dredging up the past and trying to put to rest the demons that remain from that war.

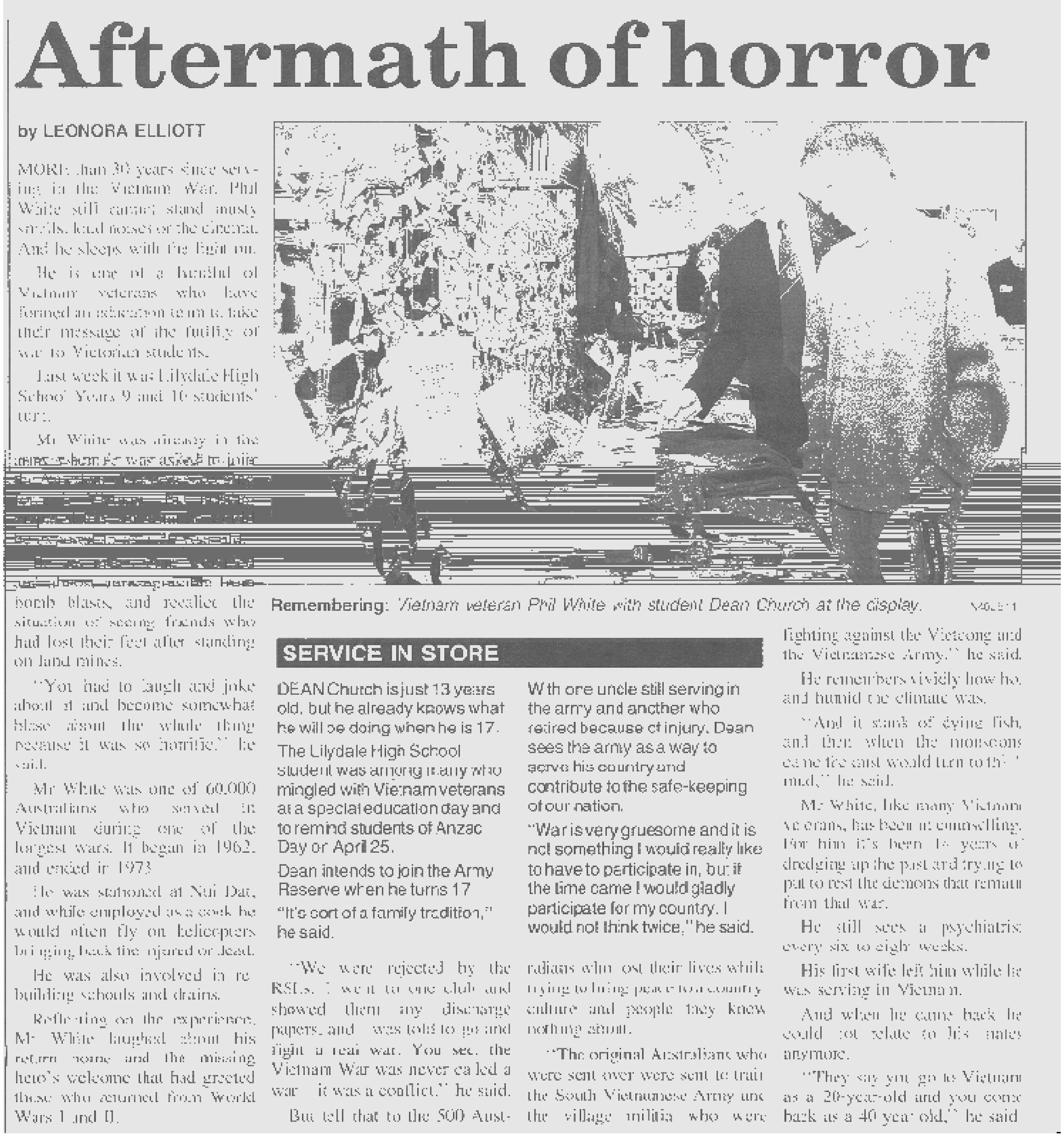
He still sees a psychiatrist every six to eight weeks.

His first wife left him while he was serving in Vietnam.

And when he came back he could not relate to his mates anymore.

"They say you go to Vietnam as a 20-year-old and you come back as a 40-year-old," he said.

Result of Improved Niblack's Algorithm



Aftermath of horror

by LEONORA ELLIOTT

MORE than 30 years since serving in the Vietnam War, Phil White still cannot stand musty smells, loud noises or the cinema. And he sleeps with the light on.

He is one of a handful of Vietnam veterans who have formed an education team to take their message of the futility of war to Victorian students.

Last week it was Lilydale High School Years 9 and 10 students' turn.

Mr White was already in the army when he was asked to join [illegible]

[illegible] bomb blasts, and recalled the situation of seeing friends who had lost their feet after standing on land mines.

"You had to laugh and joke about it and become somewhat blase about the whole thing because it was so horrific," he said.

Mr White was one of 60,000 Australians who served in Vietnam during one of the largest wars. It began in 1962, and ended in 1973.

He was stationed at Nui Dat, and while employed as a cook he would often fly on helicopters bringing back the injured or dead.

He was also involved in rebuilding schools and drains.

Reflecting on the experience, Mr White laughed about his return home and the missing hero's welcome that had greeted those who returned from World Wars I and II.

Remembering: *Vietnam veteran Phil White with student Dean Church at the display.* M0211

SERVICE IN STORE

DEAN Church is just 13 years old, but he already knows what he will be doing when he is 17.

The Lilydale High School student was among many who mingled with Vietnam veterans at a special education day and to remind students of Anzac Day or April 25.

Dean intends to join the Army Reserve when he turns 17

"It's sort of a family tradition," he said.

With one uncle still serving in the army and another who retired because of injury, Dean sees the army as a way to serve his country and contribute to the safe-keeping of our nation.

"War is very gruesome and it is not something I would really like to have to participate in, but if the time came I would gladly participate for my country. I would not think twice," he said.

"We were rejected by the RSLs. I went to one club and showed them my discharge papers, and I was told to go and fight a real war. You see, the Vietnam War was never called a war — it was a conflict," he said.

But tell that to the 500 Australians who lost their lives while trying to bring peace to a country, culture and people they knew nothing about.

"The original Australians who were sent over were sent to train the South Vietnamese Army and the village militia who were fighting against the Vietcong and the Vietnamese Army," he said.

He remembers vividly how hot and humid the climate was.

"And it stank of dying fish, and then when the monsoons came the dust would turn to thick mud," he said.

Mr White, like many Vietnam veterans, has been in counselling. For him it's been 14 years of dredging up the past and trying to put to rest the demons that remain from that war.

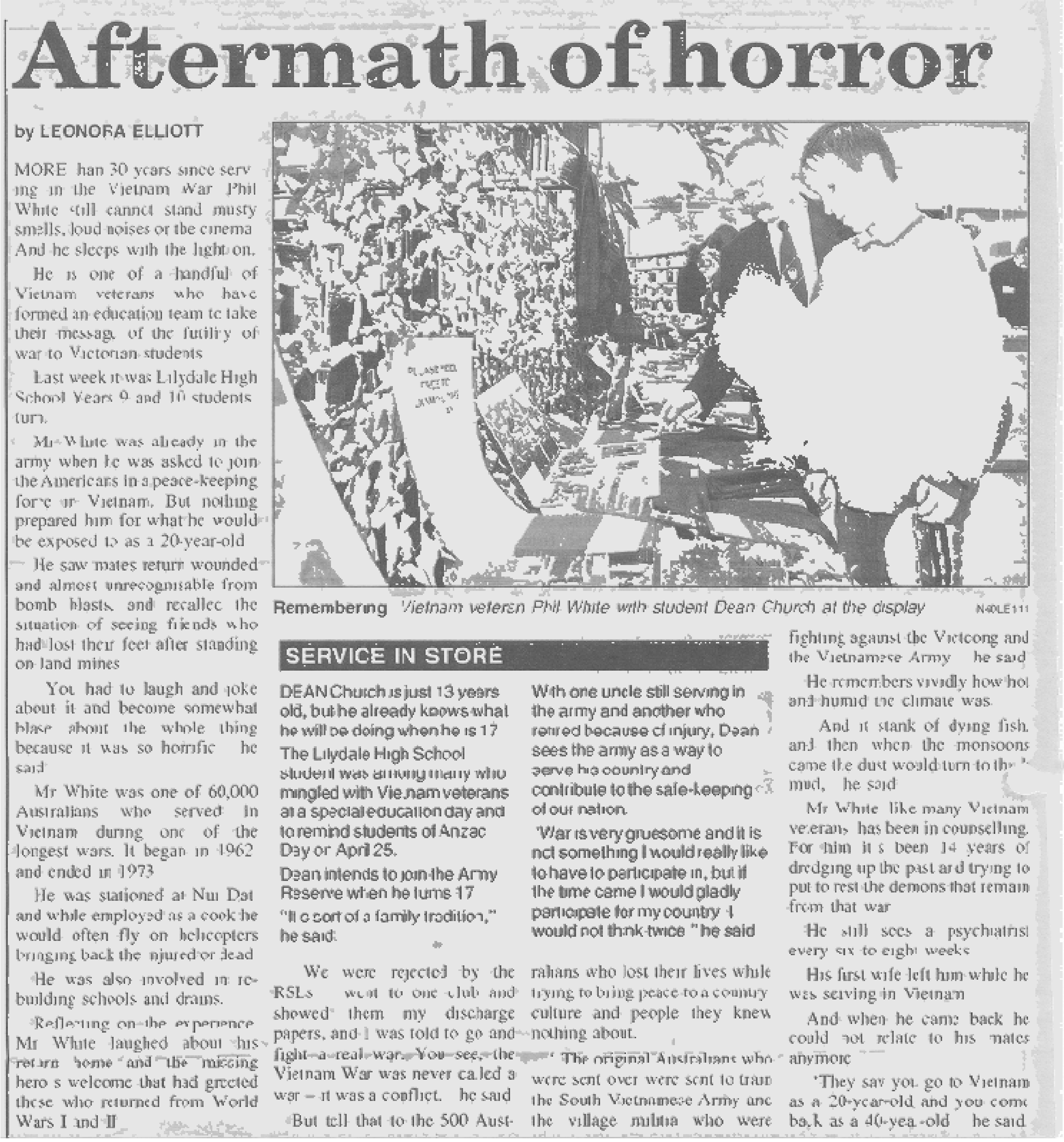
He still sees a psychiatrist every six to eight weeks.

His first wife left him while he was serving in Vietnam.

And when he came back he could not relate to his mates anymore.

"They say you go to Vietnam as a 20-year-old and you come back as a 40 year old," he said

Result of Original QIR's Algorithm



Aftermath of horror

by LEONORA ELLIOTT

MORE than 30 years since serving in the Vietnam War Phil White still cannot stand musty smells, loud noises or the cinema And he sleeps with the light on.

He is one of a handful of Vietnam veterans who have formed an education team to take their message of the futility of war to Victorian students

Last week it was Lilydale High School Years 9 and 10 students turn.

Mr White was already in the army when he was asked to join the Americans in a peace-keeping force in Vietnam. But nothing prepared him for what he would be exposed to as a 20-year-old

He saw mates return wounded and almost unrecognisable from bomb blasts and recalled the situation of seeing friends who had lost their feet after standing on land mines

You had to laugh and joke about it and become somewhat blase about the whole thing because it was so horrific he said

Mr White was one of 60,000 Australians who served in Vietnam during one of the longest wars. It began in 1962 and ended in 1973

He was stationed at Nui Dat and while employed as a cook he would often fly on helicopters bringing back the injured or dead

He was also involved in rebuilding schools and drains.

Reflecting on the experience Mr White laughed about his return home and the missing hero s welcome that had greeted those who returned from World Wars I and II

Remembering *Vietnam veteran Phil White with student Dean Church at the display* N40LE111

SERVICE IN STORE

DEAN Church is just 13 years old, but he already knows what he will be doing when he is 17

The Lilydale High School student was among many who mingled with Vietnam veterans at a special education day and to remind students of Anzac Day or April 25.

Dean intends to join the Army Reserve when he turns 17

"It s sort of a family tradition," he said.

With one uncle still serving in the army and another who retired because of injury, Dean sees the army as a way to serve his country and contribute to the safe-keeping of our nation.

'War is very gruesome and it is not something I would really like to have to participate in, but if the time came I would gladly participate for my country I would not think twice " he said

We were rejected by the RSLs went to one club and showed them my discharge papers, and I was told to go and fight a real war. You see, the Vietnam War was never called a war – it was a conflict. he said

But tell that to the 500 Australians who lost their lives while trying to bring peace to a country culture and people they knew nothing about.

' The original Australians who were sent over were sent to train the South Vietnamese Army and the village militia who were fighting against the Vietcong and the Vietnamese Army he said

He remembers vividly how hot and humid the climate was.

And it stank of dying fish, and then when the monsoons came the dust would turn to the mud, he said

Mr White like many Vietnam veterans has been in counselling. For him it s been 14 years of dredging up the past and trying to put to rest the demons that remain from that war

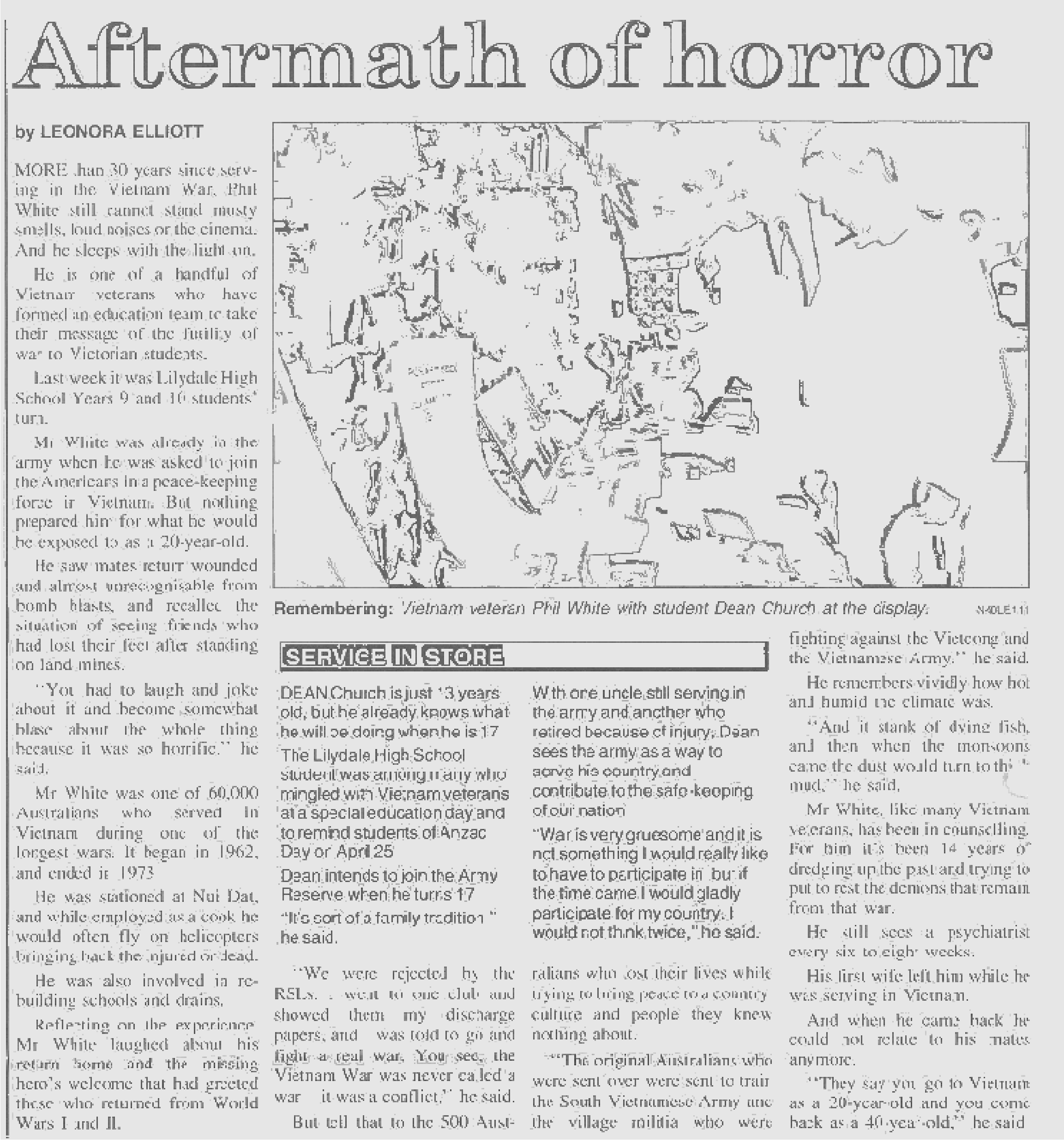
He still sees a psychiatrist every six to eight weeks

His first wife left him while he was serving in Vietnam

And when he came back he could not relate to his mates anymore

'They say you go to Vietnam as a 20-year-old and you come back as a 40-year-old he said

Result of Yanowitz's Algorithm



Aftermath of horror

by LEONORA ELLIOTT

MORE than 30 years since serving in the Vietnam War, Phil White still cannot stand musty smells, loud noises or the cinema. And he sleeps with the light on.

He is one of a handful of Vietnam veterans who have formed an education team to take their message of the futility of war to Victorian students.

Last week it was Lilydale High School Years 9 and 10 students' turn.

Mr White was already in the army when he was asked to join the Americans in a peace-keeping force in Vietnam. But nothing prepared him for what he would be exposed to as a 20-year-old.

He saw mates return wounded and almost unrecognisable from bomb blasts, and recalled the situation of seeing friends who had lost their feet after standing on land mines.

"You had to laugh and joke about it and become somewhat blase about the whole thing because it was so horrific," he said.

Mr White was one of 60,000 Australians who served in Vietnam during one of the longest wars. It began in 1962, and ended in 1973.

He was stationed at Nui Dat, and while employed as a cook he would often fly on helicopters bringing back the injured or dead.

He was also involved in rebuilding schools and drains.

Reflecting on the experience Mr White laughed about his return home and the missing hero's welcome that had greeted those who returned from World Wars I and II.

"We were rejected by the RSLs. I went to one club and showed them my discharge papers, and I was told to go and fight a real war. You see, the Vietnam War was never called a war — it was a conflict," he said.

But tell that to the 500 Australians who lost their lives while trying to bring peace to a country culture and people they knew nothing about.

"The original Australians who were sent over were sent to train the South Vietnamese Army and the village militia who were fighting against the Vietcong and the Vietnamese Army," he said.

He remembers vividly how hot and humid the climate was.

"And it stank of dying fish, and then when the monsoons came the dust would turn to thick mud," he said.

Mr White, like many Vietnam veterans, has been in counselling. For him it's been 14 years of dredging up the past and trying to put to rest the demons that remain from that war.

He still sees a psychiatrist every six to eight weeks.

His first wife left him while he was serving in Vietnam.

And when he came back he could not relate to his mates anymore.

"They say you go to Vietnam as a 20-year-old and you come back as a 40-year-old," he said.

Remembering: *Vietnam veteran Phil White with student Dean Church at the display.* N40LE111

SERVICE IN STORE

DEAN Church is just 13 years old, but he already knows what he will be doing when he is 17.

The Lilydale High School student was among many who mingled with Vietnam veterans at a special education day and to remind students of Anzac Day on April 25.

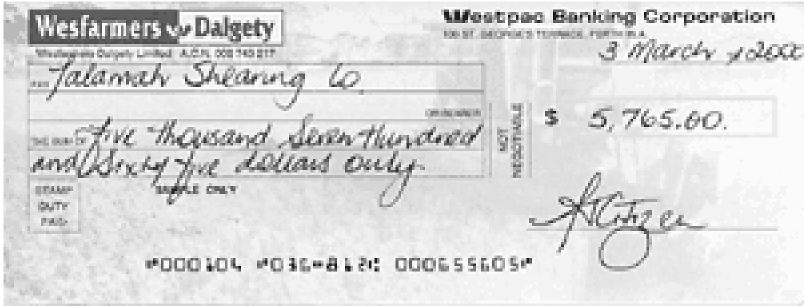
Dean intends to join the Army Reserve when he turns 17.

"It's sort of a family tradition," he said.

With one uncle still serving in the army and another who retired because of injury, Dean sees the army as a way to serve his country and contribute to the safe-keeping of our nation.

"War is very gruesome and it is not something I would really like to have to participate in, but if the time came I would gladly participate for my country. I would not think twice," he said.

Result of Mean-Gradient Algorithm



Wesfarmers Dalgety
Wesfarmers Dalgety Limited A.C.N. 008 743 217

Westpac Banking Corporation
109 ST. GEORGE'S TERRACE, PERTH W.A.

3 March 2000

Jalamah Shearing Co

OR BEARER

THE SUM OF Five thousand Seven hundred and Sixty five dollars only

NOT NEGOTIABLE

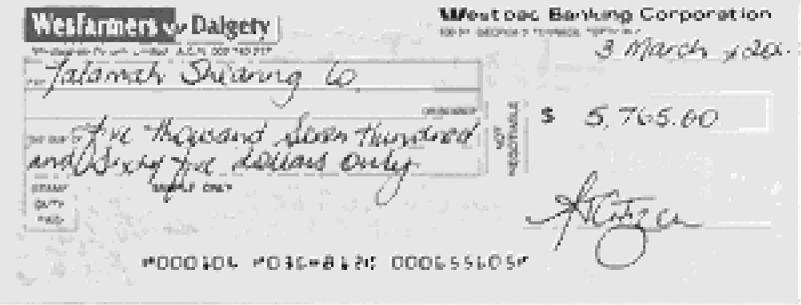
$ 5,765.60.

STAMP DUTY PAID

SAMPLE ONLY

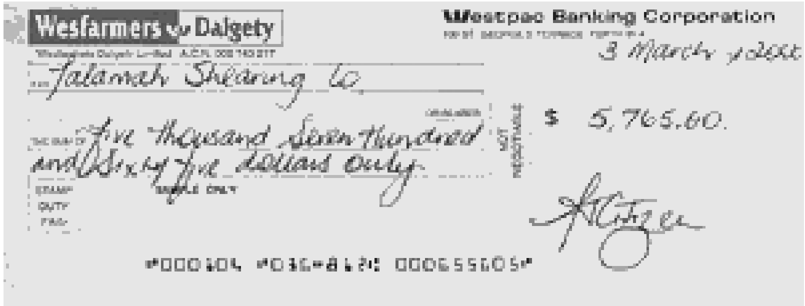
⑈000104 ⑈036⑉812⑆ 000655605⑈

Original Cheque Image 1



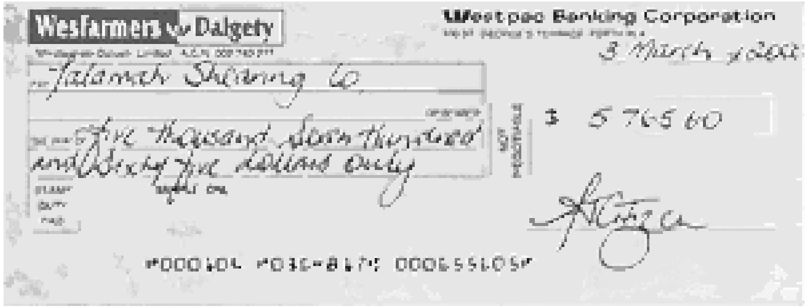
Wesfarmers Dalgety
Westpac Banking Corporation
3 March 2000
Jalamah Shearing Co
Five thousand Seven hundred and Sixty five dollars only
$ 5,765.60
⑈000104 ⑈036⑉812⑆ 000655605⑈

Result of Improved Niblack's Algorithm



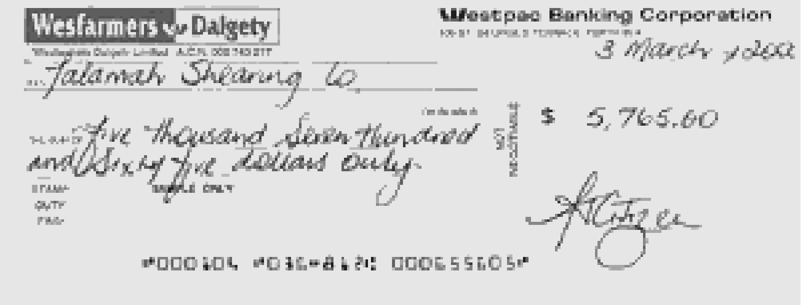
Wesfarmers Dalgety
Westpac Banking Corporation
3 March 2000
Jalamah Shearing Co
Five thousand Seven hundred and Sixty five dollars only
$ 5,765.60.
⑈000104 ⑈036⑉812⑆ 000655605⑈

Result of Original QIR's Algorithm



Wesfarmers Dalgety
Westpac Banking Corporation
3 March 2000
Jalamah Shearing Co
Five thousand Seven hundred and Sixty five dollars only
$ 5 765 60
⑈000104 ⑈036⑉817⑆ 000655605⑈

Result of Yanowitz's Algorithm



Wesfarmers Dalgety

Westpac Banking Corporation

3 March 2002

Falamah Shearing Co.

five thousand Seven hundred and Sixty five dollars only

NOT NEGOTIABLE

$ 5,765.60

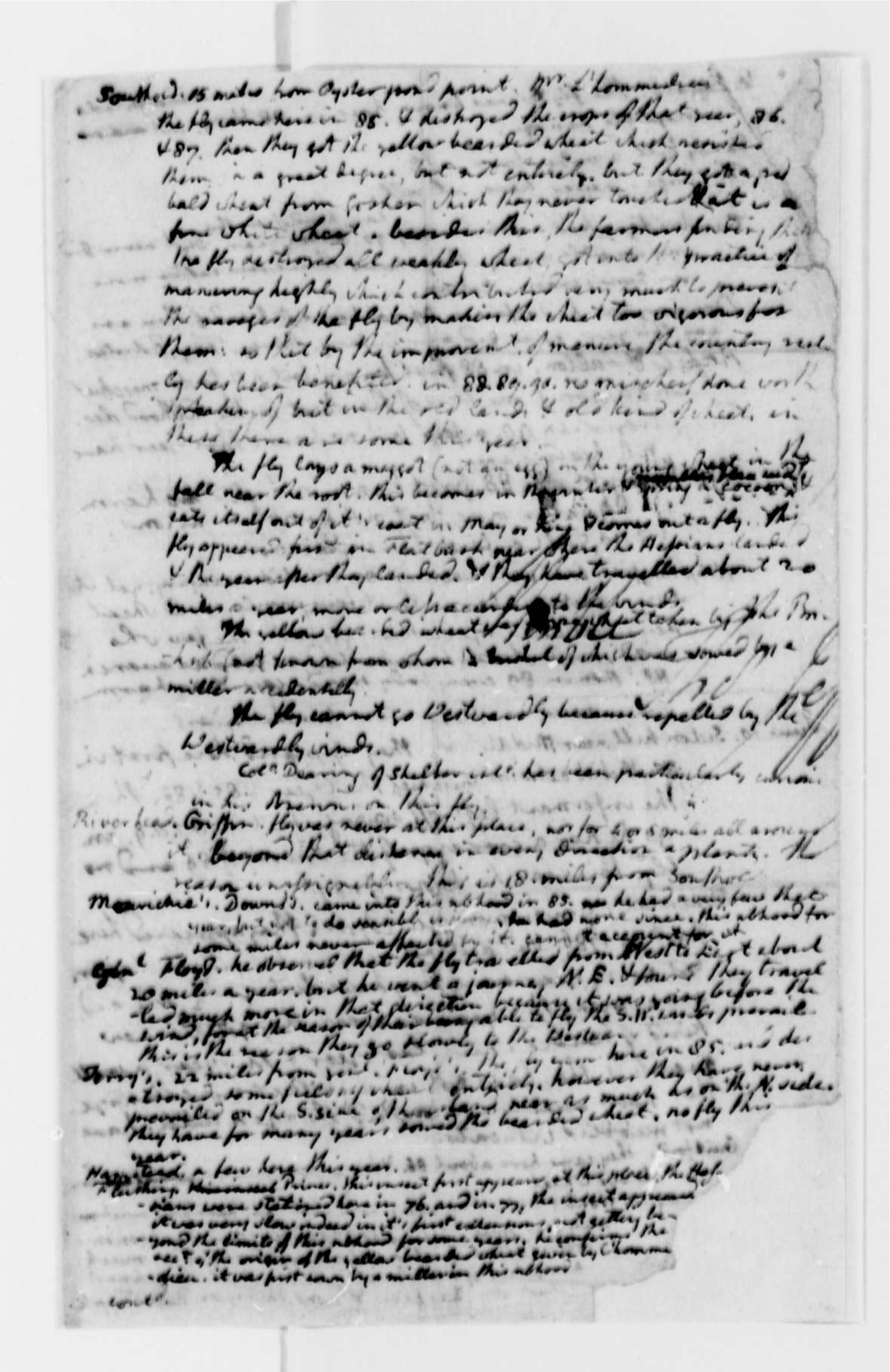
STAMP DUTY PAID

SAMPLE ONLY

⑈000104 ⑈036⑈812⑆ 000655605⑈

Result of Mean-Gradient Algorithm

APPENDIX 3: Experimental Results of Decompose Algorithm and other 6 Methods



Southold. 15 miles from Oyster pond point. Mr. L'Hommedieu.
the fly came here in 85. & destroyed the crops of that year, 86. & 87. then they got the yellow bearded wheat which resisted them in a great degree, but not entirely. but they got a red bald wheat from Goshen which they never touched. that is a fine white wheat. besides this, the farmers finding that the fly destroyed all weakly wheat, got into the practice of manuring highly which contributed very much to prevent the ravages of the fly by making the wheat too vigorous for them: so that by the improvemt. of manure the country really has been benefited. in 88. 89. 90. no mischief done worth speaking of but in the red lands & old kind of wheat. in these there are some this year.

The fly lays a maggot (not an egg) in the young wheat in the fall near the root. this becomes in November a [illegible] cocoon & eats itself out of it's coat in May or June & comes out a fly. this fly appeared first on Flatbush near where the Hessians landed & the year after they landed. & they have travelled about 20 miles a year more or less according to the winds.

The yellow bearded wheat says [illegible] taken by the Br. which first known from whom a bushel of which was sowed by a miller accidentally.

the fly cannot go Westwardly because repelled by the Westwardly winds.

Col. Dearing of Shelter isld. has been particularly curious in his researches on this fly.

Riverhead. Griffen. fly was never at this place, nor for 4 or 5 miles all around it. beyond that distance in every direction a plenty. the reason is unaccountable. this is 18 miles from Southold.

Moriche's. Downs's. came into this neighbd in 85. he had a very few that year but not to do sensible injury. has had none since. this neighbd for some miles never affected by it. cannot account for it.

Genl. Floyd. he observed that the fly travelled from West to East about 20 miles a year. but he went a journey N.E. & found they travelled much more in that direction because it was going before the wind, for the reason of their being able to fly the S.W. winds prevail. this is the reason they go slowly to the Westward.

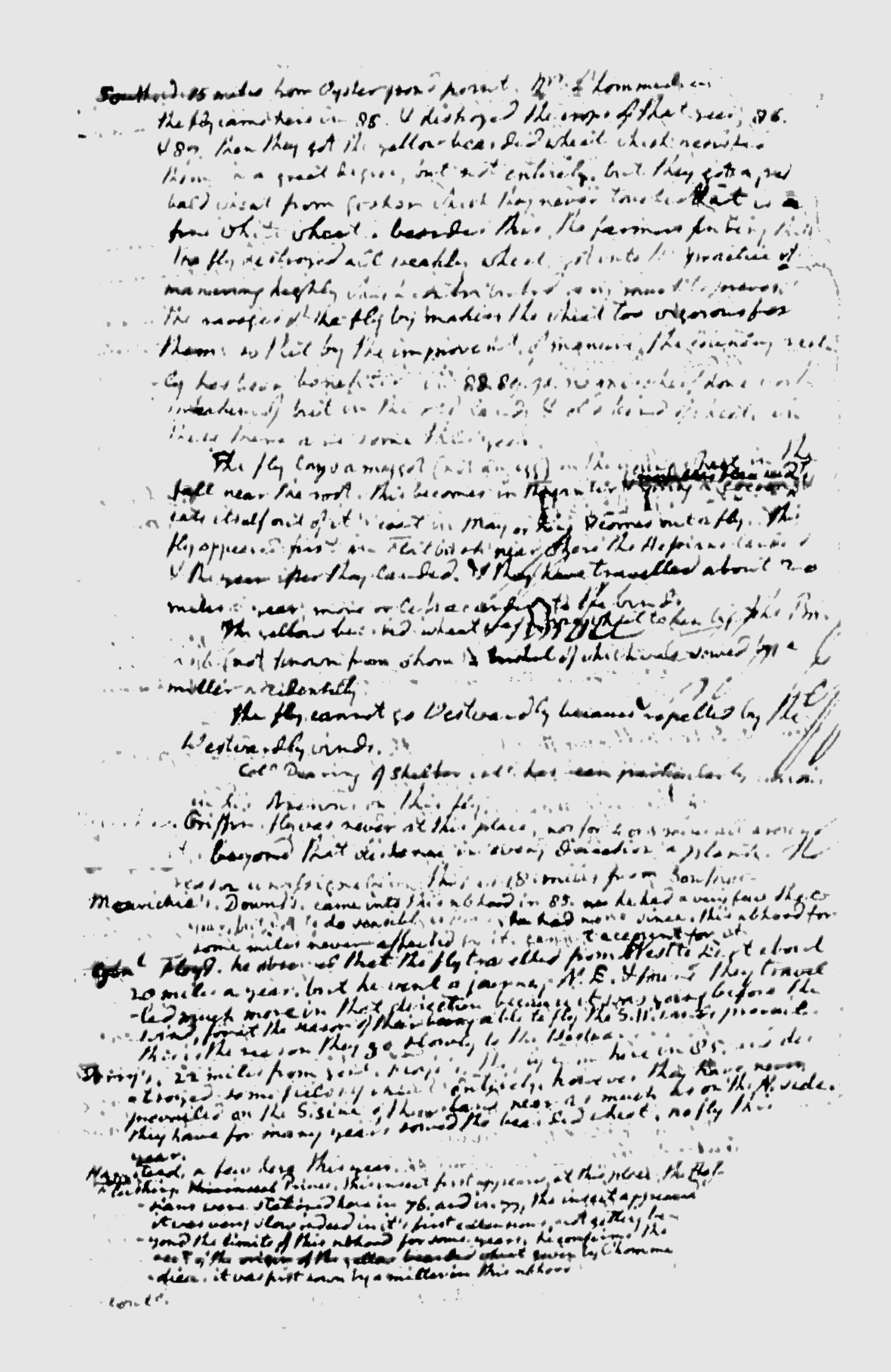
Terry's. 22 miles from Southold. fly came here in 85. and destroyed some fields of wheat entirely. however they have never prevailed on the S. side of the island near as much as on the N. side. they have for many years sowed the bearded wheat. no fly this year.

Hampstead. a few here this year.

Flatbush. Hummond Prince. this insect first appeared at this place. the Hessians were stationed here in 76. and in 77. the insect appeared. it was very slow indeed in it's first extension, not getting beyond the limits of this neighbd for some years. he confirms the acct of the origin of the yellow bearded wheat given by L'Hommedieu. it was first sown by a miller in this neighbd

cont'd.

Original Image 1



Southold. 95 miles from Oysterpond point. Mr. L'Hommedieu

the fly came here in 85. & destroyed the crops of that year, 86. & 87. then they got the yellow bearded wheat which resisted them in a great degree, but not entirely. but they got a yellow bald wheat from Groton which they never touched. that is a fine white wheat. besides this, the farmers finding that the fly destroyed all weakly wheat, got into the practice of manuring highly [illegible] the ravages of the fly by making the wheat too vigorous for them. so that by the improvement of manure, the country really has been benefited. in 88. 89. 90. [illegible] in the [illegible] & [illegible] of wheat, in these farms as in some [illegible] years.

The fly lays a maggot (not an egg) on the young wheat in the fall near the root. this becomes in the winter a chrysalis & generates itself out of it in May or June & becomes a fly. the fly appeared first on Flatbush near where the Hessians landed & the year after they landed. & they have travelled about 20 miles a year more or less according to the winds.

The yellow bearded wheat was first taken by [illegible] not known from whom. a bushel of which was sowed by a miller accidentally.

the fly cannot go Westwardly because repelled by the Westwardly winds.

Colo. Deering of Shelter isld. has been particularly [illegible] in his [illegible] on this fly.

Griffin. the fly was never at this place, nor for 4 or 5 miles around it. beyond that distance in every direction [illegible]. the reason is unassignable. this is 18 miles from Southold.

McCurdie's. Dounds. came into this neighborhood in 85. [illegible] he had a very fine [illegible] [illegible] to do sensibly [illegible] he had none since. this neighborhood for some miles never affected by it. cannot account for it.

Genl. Floyd. he observed that the fly travelled from West to East about 20 miles a year. but he went a journey N. E. & found they travel much more in that direction because it was years before the [illegible] able to fly the S. W. winds prevail [illegible] the reason they go slowly to the Westward.

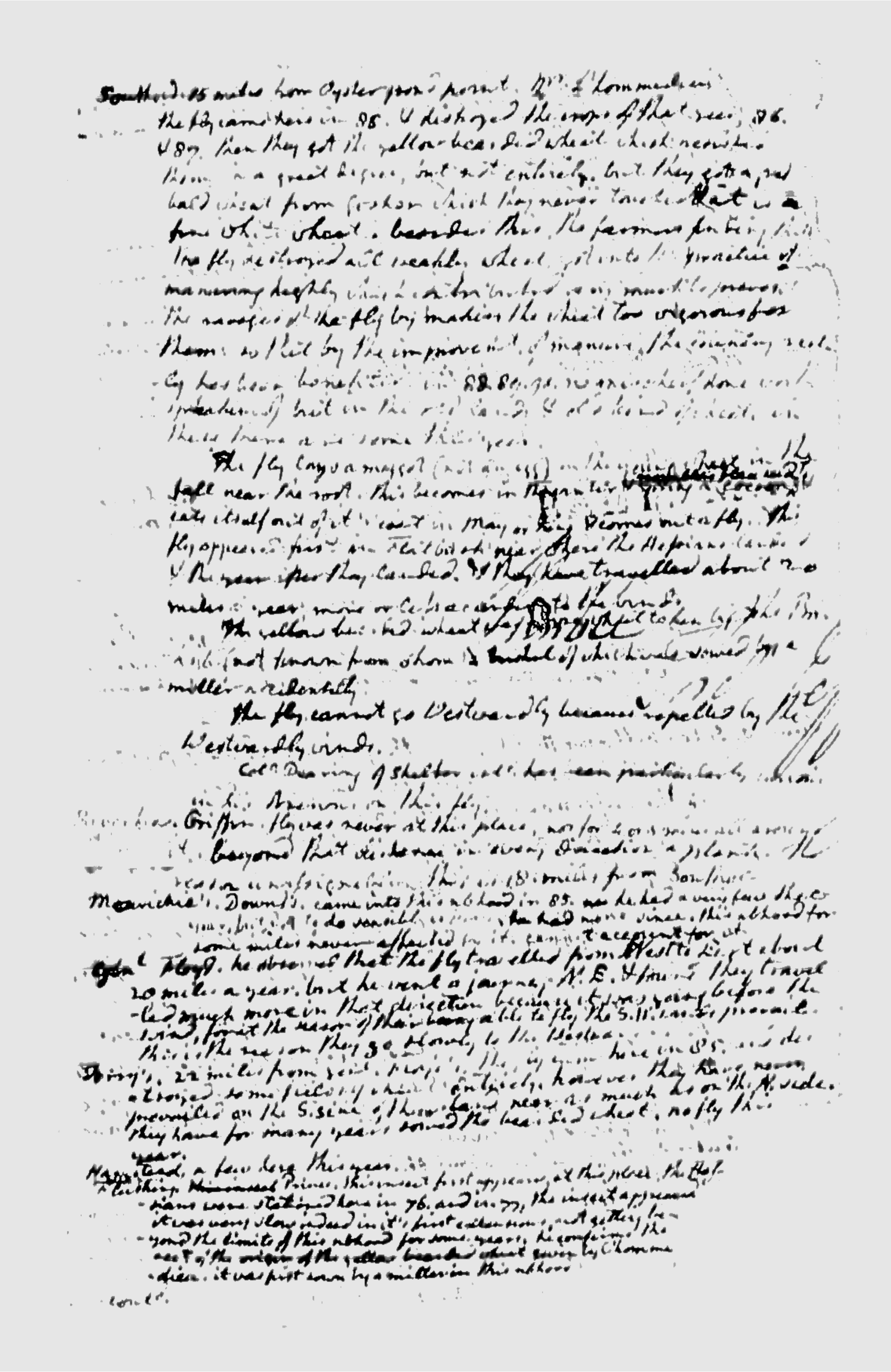
[illegible]'s. 22 miles from [illegible] here in 85. [illegible] destroyed some fields of wheat entirely. however they have never [illegible] as much as on the N. side. prevailed on the S. side of [illegible] near [illegible] wheat. no fly this year. they have for many years sowed the bearded wheat.

Hempstead. a few [illegible] this year.

[illegible] Prince. the insect first appeared at this place. the Hessians were stationed here in 76. and in 77. the insect appeared it was very slow indeed in it's first advances, not getting beyond the limits of this neighbd for some years. he confirms the acct of the origin of the yellow bearded wheat given by L'Hommedieu. it was first sown by a miller in this neighbd

contd.

Decompose Algorithm



Southold. 85 miles from Oyster pond point. Mr. L'Hommedieu:
the fly came here in 86. & destroyed the crops of that year, 86.
& 87. then they got the yellow bearded wheat which resisted
them in a great degree, but not entirely. but they got a red
bald wheat from Gosham which they never touched. that is a
fine white wheat. besides this the farmers for the 2 first
years the fly destroyed all the early wheat, got into the practice of
manuring highly which [illegible] forwarded very much to prevent
the ravages of the fly by making the wheat too vigorous for
them. so that by the improvement of manure, the country really
has been benefitted. in 88. 89. 90. [illegible] done [illegible]
[illegible] but in the [illegible] [illegible] kind of wheat. in
[illegible] [illegible] for some [illegible] years.

The fly lays a maggot (not an egg) on the young ~~[illegible]~~ in the
fall near the root. this becomes in the winter a ~~[illegible]~~
eats itself out of it & comes in May or June becomes out a fly. the
fly appeared first on Flatbush near where the Hessians landed
& the year after they landed. & they have travelled about 20
miles a year more or less according to the winds.

The yellow bearded wheat [illegible] taken by the Br[illegible]
[illegible] (not known from whom) a bushel of which was sowed by a
miller accidentally.

the fly cannot go Westwardly because repelled by the
Westwardly winds.

Col. Deering of Shelter isld. has been particularly [illegible]
in his [illegible] on this fly.

Riverhead. Griffin. flies never at this place, nor for 4 or 5 miles around
it. beyond that distance in every direction on the island. the
reason is unaccountable. this is 18 miles from Southold.

Moriches. Downs. came into this nbhood in 85. he had a very few the
year, but did no sensible injury. has had none since. this nbhood for
some miles never affected by it. cannot account for it.

Genl. Floyd. he observed that the fly travelled from West to East about
20 miles a year. but he went a journey N.E. & found they travel-
-led much more in that direction because it was going before the
wind, for it the reason of their being able to fly the S.W. winds prevail.
this is the reason they go slowly to the Westward. [illegible] here in 85. [illegible]

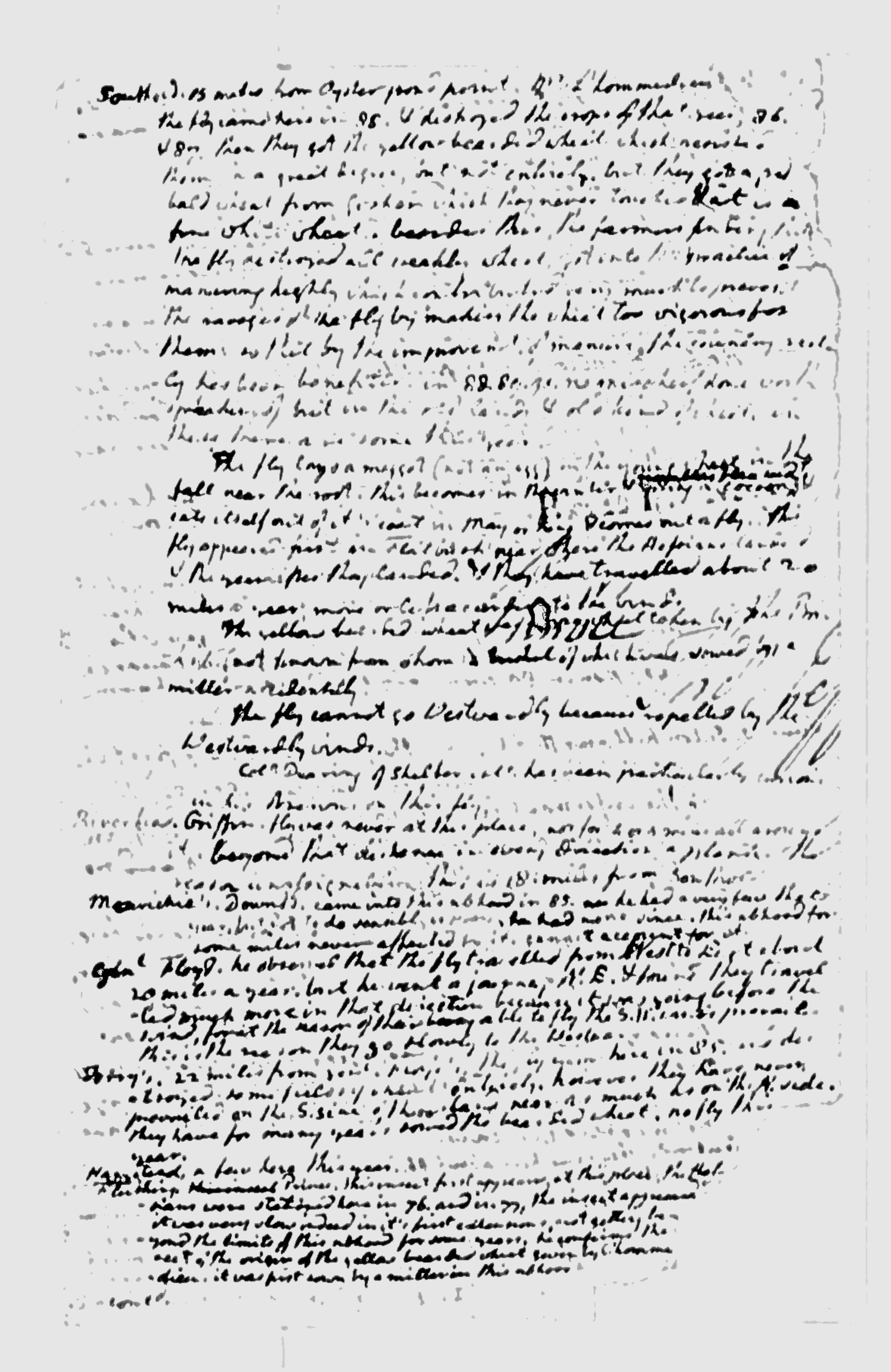
Smith's, 22 miles from [illegible]. [illegible] they have never
destroyed some fields of wheat entirely, however they [illegible] as much as on the N. side.
prevailed on the S. side of the island near as much [illegible]. no fly this
they have for many years sowed the bearded wheat.
year.

Hampstead, a few here this year.

Flushing. ~~[illegible]~~ Prince. this insect first appeared at this place the Hes-
-sians were stationed here in 76. and in 77, the insect appeared
it was very slow indeed in it's first extension, not getting be-
-yond the limits of this nbhood for some years. he confirms the
acct of the origin of the yellow bearded wheat given by L'Homme-
-dieu. it was first seen by a miller in this nbhood.

contd.

Improved Decompose Algorithm



Southold. 95 miles from Oyster pond point. Mr. L'Hommedieu.
the fly came here in 85. & destroyed the crops of that year, 86.
& 87. then they got the yellow bearded wheat which resists
them in a great degree, but not entirely. but they got a red
bald wheat from Goshen which they have found to be that is a
fine white wheat. besides this the farmers finding
the fly destroyed all weakly wheat, got into the practice of
manuring highly which has rendered them much less subject to
the ravages of the fly by making the wheat too vigorous for
them: so that by the improvement of manuring the country really
has been benefited. in 88. 89. 90. 91. no mischief done worth
speaking of but on the poor lands & of a kind of [illegible] on
these lands a [illegible] some [illegible] year.

The fly lays a maggot (not an egg) in the young wheat in the
fall near the root. this becomes in December & spring a cocoon &
eats itself out of it & comes out in May or June a fly. the
fly appeared first on Staten island near where the Hessians landed
& the year after they landed. & they have travelled about 20
miles a year more or less according to the winds.

The yellow bearded wheat [illegible] taken by the [illegible]
it is not known from whom a bushel of which [illegible] sowed in 91 a
miller accidentally

the fly cannot go Westwardly because repelled by the
Westwardly winds.

Colo. Deering of Shelter isl. has been particularly curious
in his observations on this fly.

Riverhead. Griffin. the fly was never at this place, nor for 8 or 9 miles around
it. beyond that distance in every direction a plenty. the
reason is unaccountable. this is 18 miles from Southold.

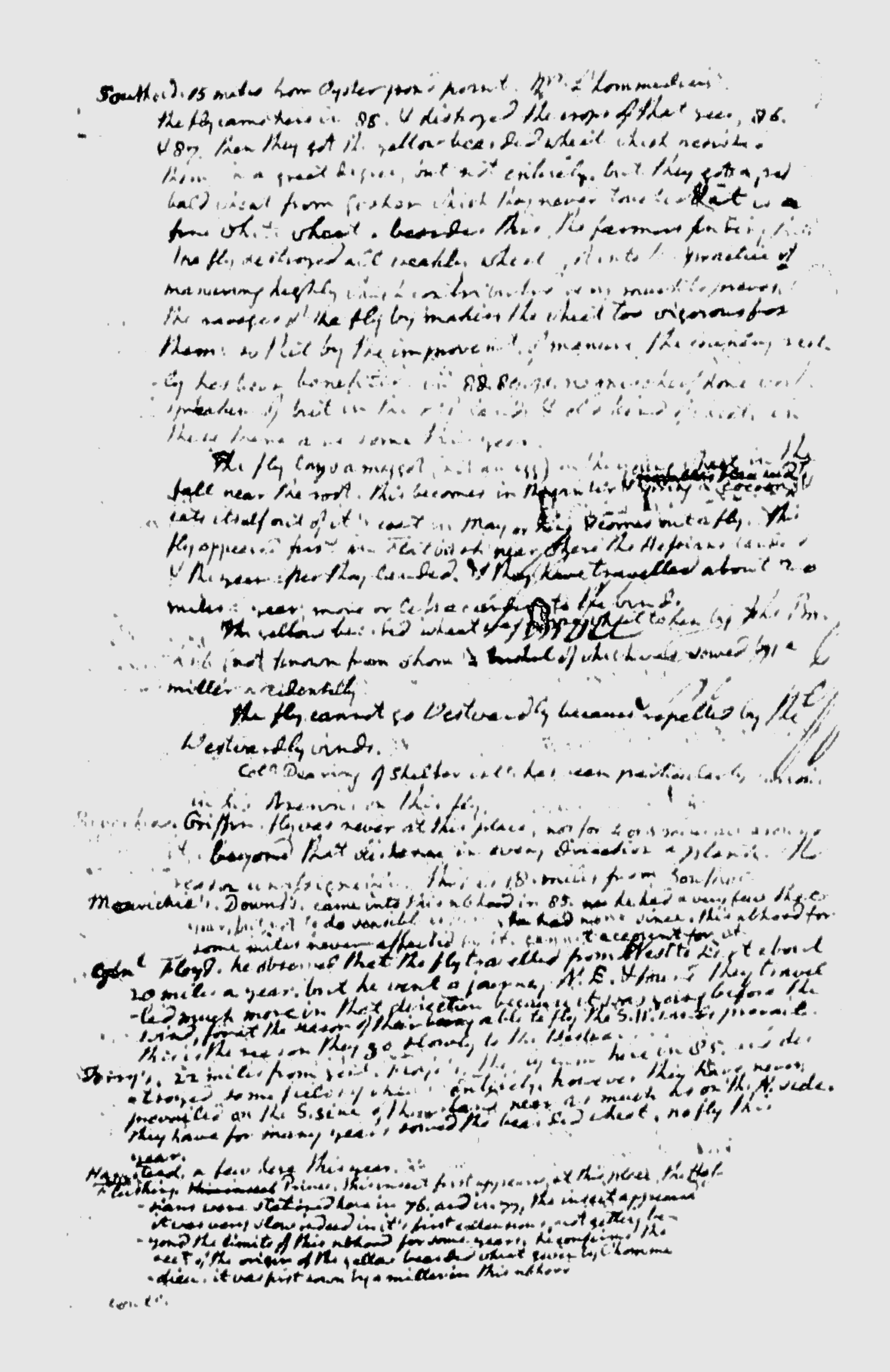
Moriches. Downs. came into this neighborhood in 85. he had a very few the 2
years, but not to do much mischief, he had none since. this neighborhood for
some miles never affected by it. cannot account for it.

Genl. Floyd. he observed that the fly travelled from West to East about
20 miles a year. but he went a journey N.E. & found they travel
[illegible] more in that direction because it was going before the
wind, & that the reason of their being able to fly the S.W. winds prevail.
this is the reason they go slowly to the Westward. here in 85. [illegible]

Brown's. 22 miles from [illegible] they have never
destroyed some fields of wheat entirely. however they have much so on the N. side.
prevailed on the S. side of those lands near as much the bearded wheat. no fly this
they have for many years sowed the bearded wheat.
year.

Hempstead. a few here this year.
Flushing. Hermanus Prince. the insect first appeared at this place. the Hes-
-sians were stationed here in 76. and in 77. the insect appeared.
it was very slow indeed in it's first advances, not getting be-
-yond the limits of this neighborhood for some years. he confirms the
acct. of the origin of the yellow bearded wheat given by Chamm-
-dian. it was first seen by a miller in this neighborhood
cont'd.

ETM's Algorithm



Southold. 15 miles from Oyster pond point. Mr. L'Hommedieu's.
the fly came here in 85. & destroyed the crops of that year, 86. & 87. then they got the yellow bearded wheat which resisted them in a great degree, but not entirely. but they got a red bald wheat from Groton which they never touched. that is a fine white wheat. besides this, the farmers finding that the fly destroyed all weakly wheat, went into the practice of manuring highly [illegible] in order to prevent the ravages of the fly by making the wheat too vigorous for them: so that by the improvement of manure the country really has been benefited. in 88. 89. 90. no [illegible] done [illegible] spoken of but in the [illegible] lands & old kind of wheat, in these [illegible] some few years.

The fly lays a maggot (not an egg) on the young plant in the fall near the root. this becomes in November & spring a cocoon, eats itself out of it's coat in May or June becomes a fly. the fly appeared first on Staten island near where the Hessians landed & the year after they landed. & they have travelled about 20 miles a year more or less according to the wind.

The yellow bearded wheat [illegible] taken by the [illegible] (not known from whom) in a bushel of which was sowed by a miller accidentally.

the fly cannot go Westwardly because repelled by the Westwardly winds.

Col. Deering of Shelter island has been particularly [illegible] in his researches on this fly.

Riverhead. Griffin. fly was never at this place, nor for 4 or 5 miles around it. beyond that distance in every direction a plenty. the reason unknown [illegible]. this is 18 miles from Southold.

Moriches. Downs's. came into this neighborhood in 85. he had a very few the [illegible] year, but not to do sensible [illegible]. he had none since. this neighborhood for some miles never affected by it. cannot account for it.

Genl. Floyd. he observed that the fly travelled from West to East about 20 miles a year. but he went a journey N. E. & found they travel much more in that direction because it was going before the wind, for it the reason of their being able to fly the S. W. [illegible] prevail. this is the reason they go slowly to the Westward. [illegible] here in 85. [illegible]

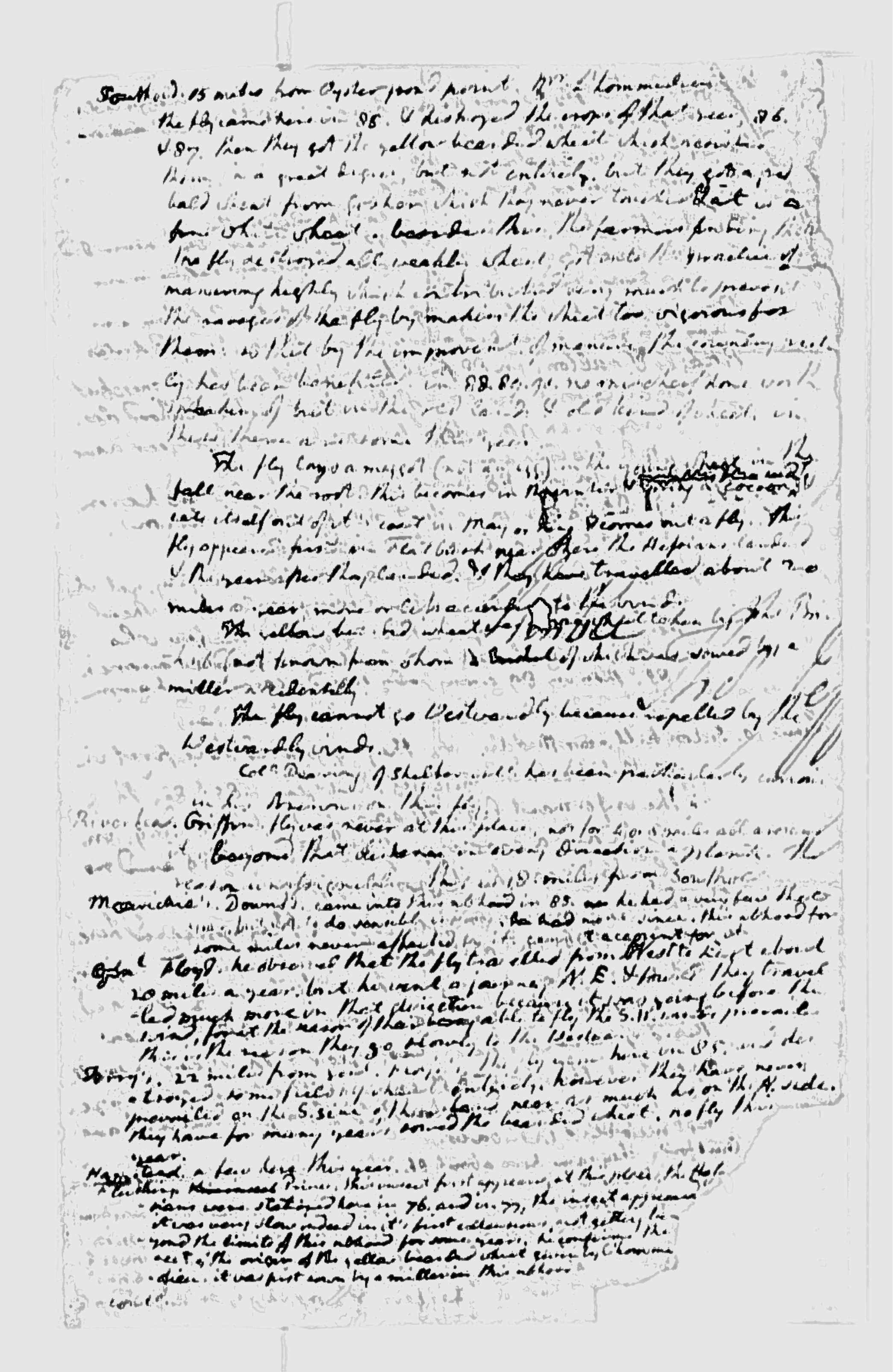
[illegible]'s. 22 miles from [illegible]. they [illegible] entirely. however they have never destroyed some fields of wheat entirely. however they have done as much as on the N. side. prevailed on the S. side of [illegible] near as much. they have for many years sowed the bearded wheat. no fly this year.

Hampstead. a few here this year.

Flushing. ~~[illegible]~~ Prince. the insect first appeared at this place, the Hessians were stationed here in 76. and in 77. the insect appeared. it was very slow indeed in it's first extension, not getting beyond the limits of this neighborhood for some years. he confirms the acct. of the origin of the yellow bearded wheat given by L'Hommedieu. it was first sown by a miller in this neighbo

cont'd.

Improved Niblack's Technique



Southold. 15 miles from Oyster pond point. Mr. L'Hommedieu.
the fly came here in 85. & destroyed the crops of that year, 86.
& 87. then they got the yellow bearded wheat which resisted
them in a great degree, but not entirely. but they got a red
bald wheat from Groshon which they never touched. that is a
fine white wheat. besides this, the farmers finding that
the fly destroyed all weakly wheat, got into the practice of
manuring highly which enables the wheat very much to prevent
the ravages of the fly by making the wheat too vigorous for
them. so that by the improvement of manuring, the country real-
ly has been benefited. in 88. 89. 90. no mischief done worth
speaking of but in the red land. & old kind of wheat. in
these three or four last years.

The fly lays a maggot (not an egg) in the young wheat in the
fall near the root. this becomes in November a [illegible] cocoon &
eats itself out of it's case in May or June & becomes a fly. the
fly appeared first on Flatbush near where the Hessians landed
& the year after they landed. & they have travelled about 20
miles a year more or less according to the wind.

The yellow bearded wheat was [illegible] first taken by [illegible]
[illegible] first known from [illegible] a bushel of which was sowed by a
miller accidentally.

The fly cannot go Westwardly because repelled by the
Westwardly winds.

Colo. Deering of Shelter island has been particularly curious
in his observations on this fly.

Riverhead. Griffin. fly was never at this place, nor for 4 or 5 miles all around
it. beyond that distance in every direction a plenty. the
reason is not yet conceived. this is 18 miles from Southold.

Moriches. Downs. came into this neighborhood in 85. as he had a very fine [illegible]
[illegible] to do sensible injury. he had none since. this neighborhood for
some miles never affected by it. cannot account for it.

Captn. Floyd. he observed that the fly travelled from West to East about
20 miles a year. but he went a journey N.E. & found they travel-
led much more in that direction. because if going before the
wind, [illegible] the reason of their being able to fly, the S.W. winds prevail.
this is the reason they go slowly to the Westward.

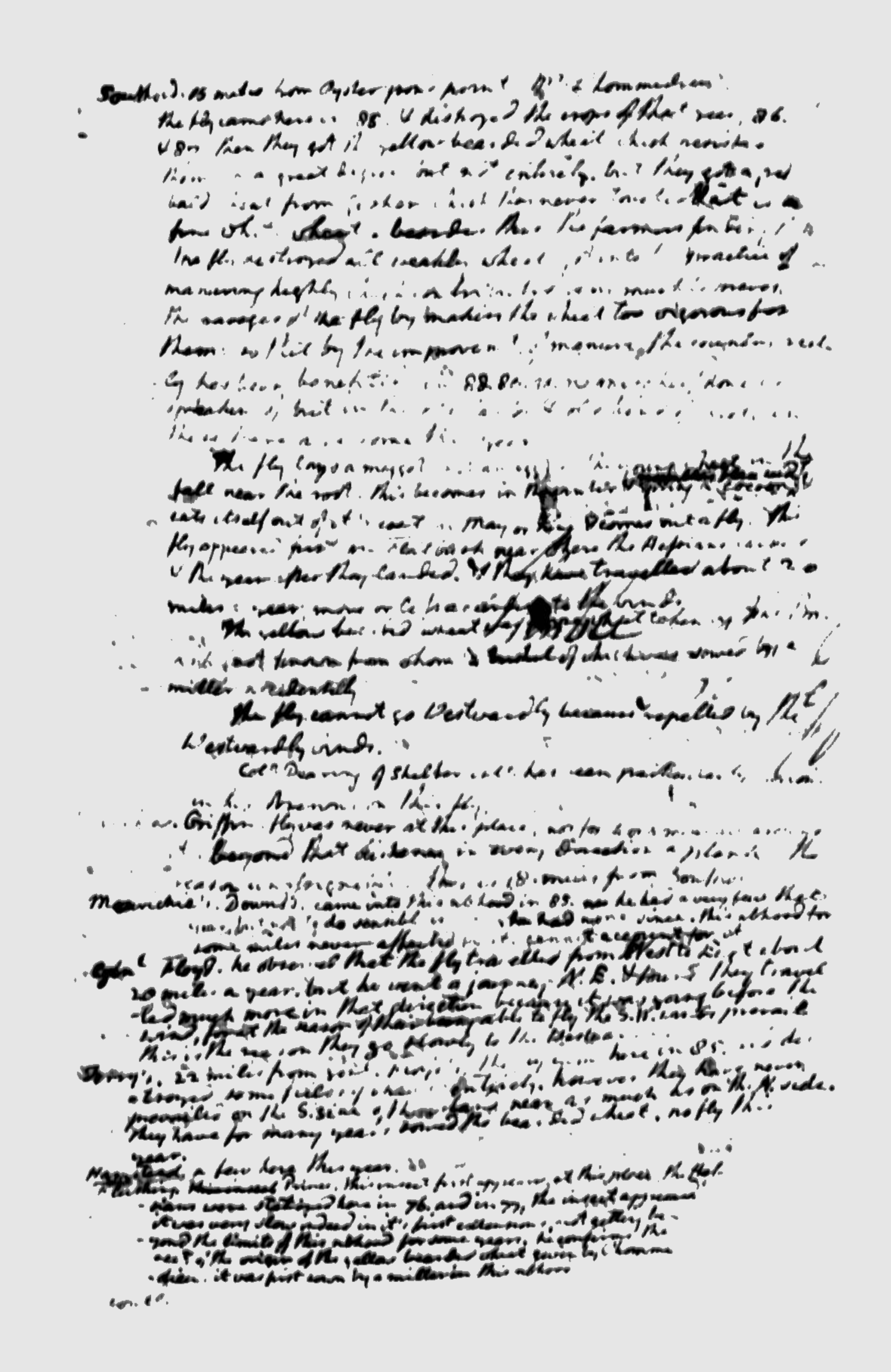
Terry's. 22 miles from Southold. [illegible] the fly came here in 85. and des-
troyed some fields of wheat entirely. however they have never
prevailed on the S. side of the [illegible] near as much as on the N. side.
they have for many years sowed the bearded wheat. no fly this
year.

Hempstead. a few here this year.

Flushing. ~~[illegible]~~ Prince. this insect first appearing at this place, the Hes-
sians were stationed here in 76. and in 77. the insect appeared.
it was very slow indeed in it's first advances, not getting be-
yond the limits of this neighborhood for some years. he confirms the
acct of the origin of the yellow bearded wheat given by L'Hommedieu.
it was first seen by a miller in this neighborhood.

cont'd.

Bernsen's Method



Southold. 15 miles from Oyster pond point [illegible] Commodore [illegible].
the fly came here in 85. & destroyed the crops of that year, 86.
& 87. then they got the yellow bearded wheat which resists
them in a great degree but not entirely. but they got a red
bald wheat from [illegible] which has never [illegible] that is a
fine wheat. besides this the farmers further [illegible]
the fly destroyed all weakly wheat [illegible] practice of
manuring highly [illegible] much [illegible] more.
the ravages of the fly by making the wheat too vigorous for
them: so that by the improvement of manuring the country [illegible]
[illegible] has been benefitted [illegible] 88. 89. 90. [illegible]
[illegible]
[illegible] year.

The fly lays a maggot [illegible] egg [illegible] in the
fall near the root. this becomes in [illegible] & [illegible]
eats itself out of it [illegible] in May or [illegible] becomes out a fly. the
fly appeared first on [illegible] year [illegible] the Hessians [illegible]
& the year after they landed. & they have travelled about 20
miles a year more or less according to the winds.

The yellow bearded wheat [illegible] taken [illegible]
[illegible] from whom [illegible] bushel of which was sowed by a
miller accidentally.

the fly cannot go Westwardly because repelled by the
Westwardly winds.

Colo. [illegible] of Shelter isld. has [illegible]
in his [illegible] in this fly.

[illegible] Griffin. they was never at this place, nor for [illegible]
[illegible] beyond that distance in [illegible] a place [illegible] the
reason [illegible] this is 18. miles from Southold.

Moriches. Downs. came into this neighbourhood in 89. [illegible] he had a very fine [illegible]
[illegible] he had none since. this neighbourhood for
some miles never affected [illegible] cannot account for it.

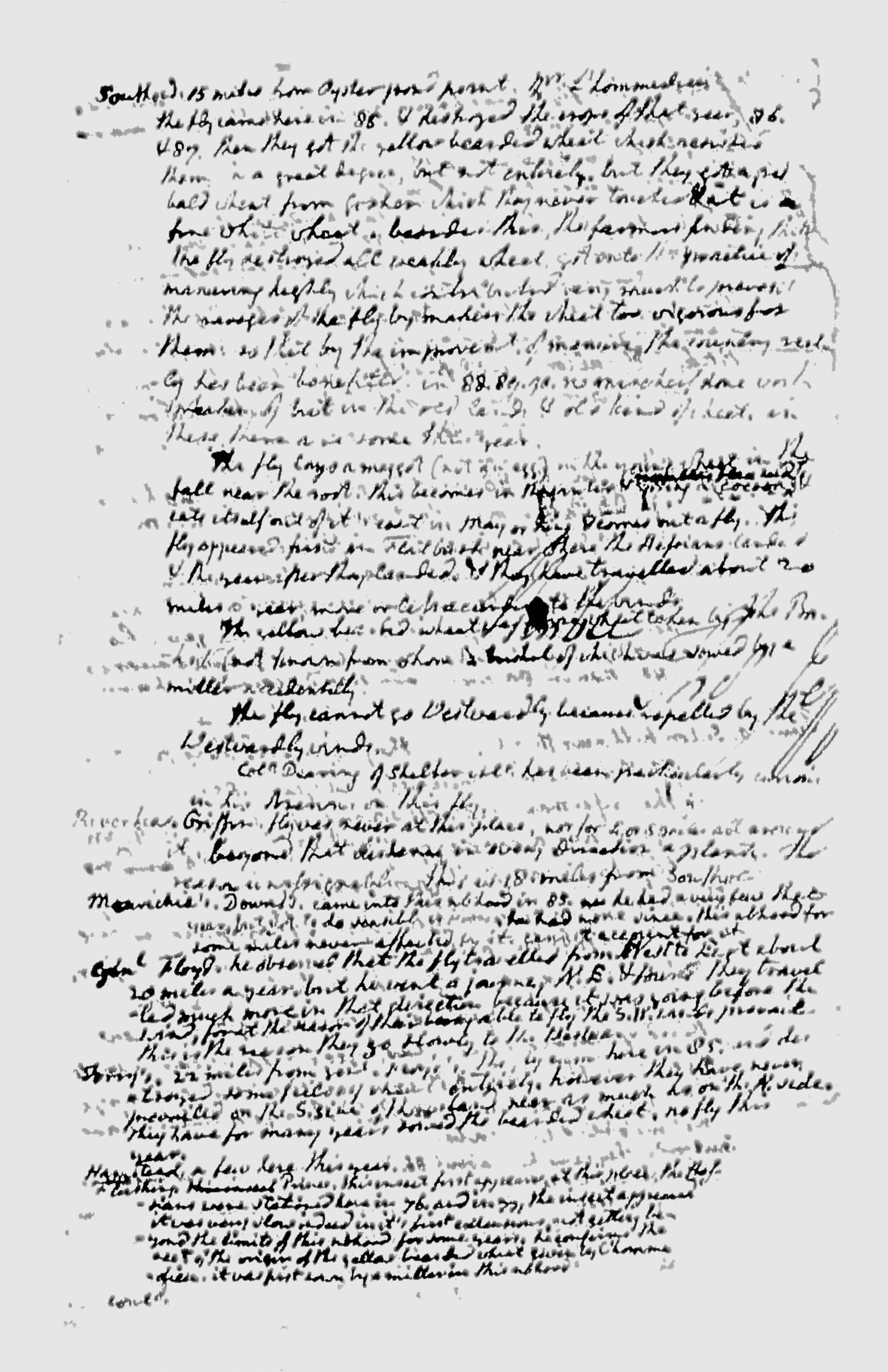
Genl. Floyd. he observed that the fly travelled from West to East about
20 miles a year. but he went a journey N.E. & some S. they travel
[illegible] more in that direction [illegible] before the
wind, [illegible] the reason of their [illegible] able to fly the S.W. [illegible]
this is the reason they go [illegible] to the [illegible] here in 85. [illegible]

Smith's. 22 miles from [illegible] however they [illegible] never
[illegible] some fields [illegible] much as on the N. side.
[illegible] on the S. side [illegible] wheat. no fly [illegible]
they have for many years sowed the bearded wheat [illegible]
year.

Hempstead. a few [illegible] this year. [illegible]
[illegible] this insect first appeared at this place. the [illegible]
[illegible] were stationed here in 76. and in 77. the insect appeared
it was very slow indeed in its first advances, not getting be-
yond the limits of this neighbourhood for some years. he confirms the
[illegible] of the origin of the yellow bearded wheat given by [illegible]
[illegible] it was first sown by a miller in this neighbourhood

[illegible]

Otsu’s Technique



Southold. 15 miles from Oyster pond point. Mr. L'Hommedieu.
the fly came here in 86. & destroyed the crops of that year, 86.
& 87. then they got the yellow bearded wheat which resists
them in a great degree, but not entirely. but they escaped
bald wheat from Goshen which they never touched. that is a
fine white wheat. besides this, the farmers finding that
the fly destroyed all weakly wheat, got into the practice of
manuring highly which he thinks has been very much to prevent
the ravages of the fly by making the wheat too vigorous for
them: so that by the improvement of manuring the country really
has been benefited. in 88. 89. 90. no mischief done worth
speaking of but in the red lands, & old kind of wheat. in
these [illegible] a [illegible] some this year.

The fly lays a maggot (not an egg) in the young wheat in the
fall near the root. this becomes in the winter [illegible] a cocoon &
eats itself out of it & [illegible] in May or June comes out a fly. the
fly appeared first in Flatbush near where the Hessians landed
& the year after they landed. & they have travelled about 20
miles a year more or less according to the winds.

The yellow bearded wheat says [illegible] was taken by the Br.
[illegible] not known from whence. a bushel of which was sowed by a
miller accidentally

the fly cannot go Westwardly because repelled by the
Westwardly winds.

Colo. Deering of Shelter isld. has been particularly curious
in his researches on this fly.

Riverhead. Griffin. flies never at this place, nor for 4 or 5 miles all around
it. beyond that distance in every direction it plenty. the
reason is unaccountable. this is 18 miles from Southold.

Moriches's. Downs. came into this neighborhood in 85. [illegible] he had a very few that
year, but did no damage. [illegible] he had none since. this neighborhood for
some miles never affected by it. cannot account for it.

Genl. Floyd. he observes that the fly travelled from West to East about
20 miles a year. but he went a journey N.E. & found they travel
much more in that direction because it was going before the
winds. & [illegible] the reason of their being able to fly the S.W. winds prevailing.
this is the reason they go slowly to the Westward.

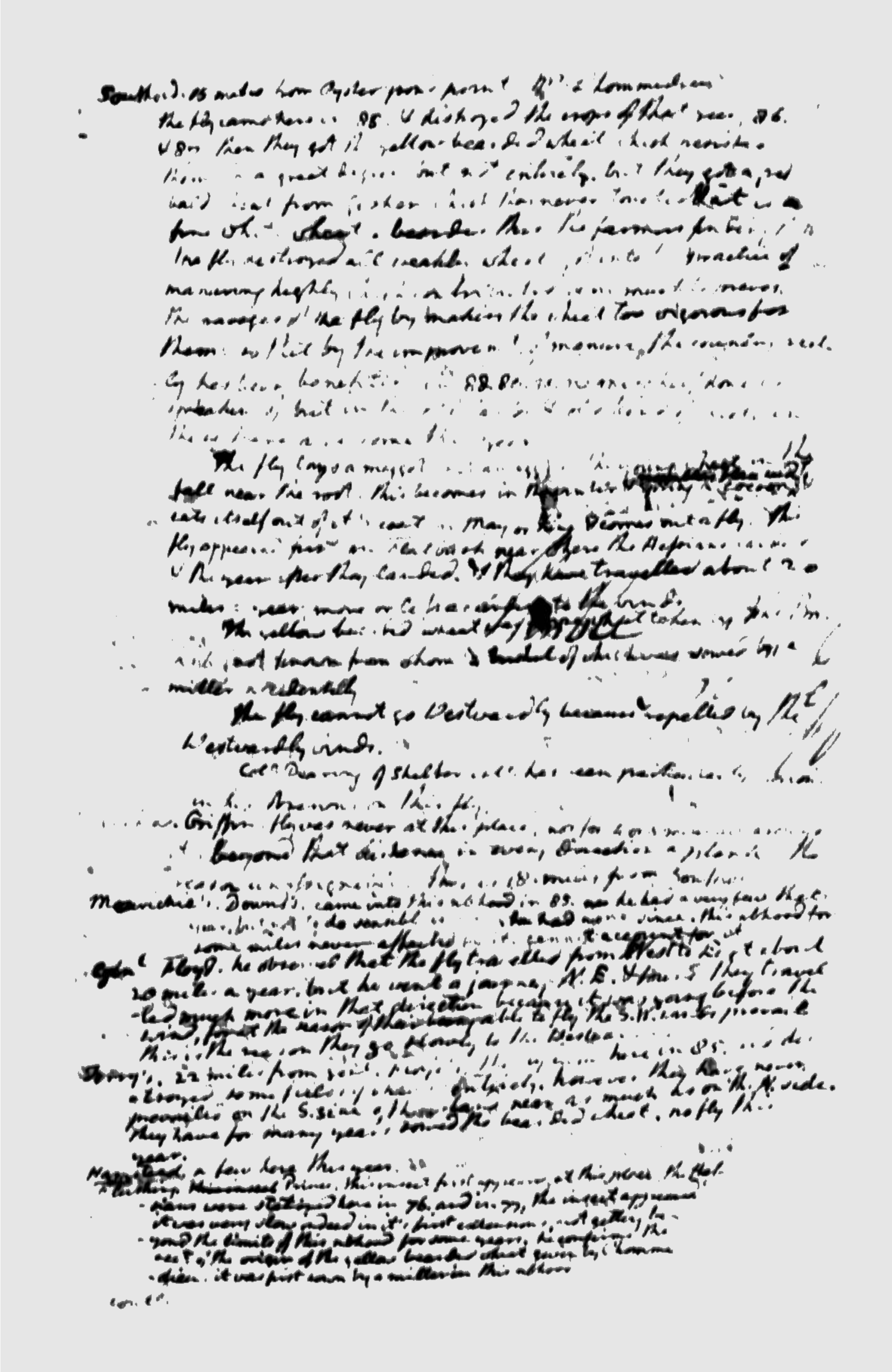
Strong's. 22 miles from [illegible] they [illegible] here in 85. and destroyed some fields of wheat entirely. however they have never
prevailed on the S. side of the island near as much as on the N. side. no fly this
year. they have for many years sowed the bearded wheat.

Hampstead. a few here this year.

Flushing. Prince. the insect first appeared at this place, the Hessians were stationed here in 76. and in 77. the insect appeared.
it was very slow indeed in it's first advances, not getting be-
yond the limits of this neighborhood for some years. he confirms the
acct. of the origin of the yellow bearded wheat given by L'Hommedieu.
it was first seen by a miller in this neighborhood

contd.

Yanowitz's Algorithm



Southold. 25 miles from Oyster pond point. [illegible] communicates
the fly came here in 85. & destroyed the crops of that year, 86.
& 87. then they got the yellow bearded wheat which resists
them in a great degree but not entirely. but they got a red
bald wheat from [illegible] which has never [illegible] that is a
fine white wheat. besides this the farmers fortify [illegible]
the fly destroyed all the weakly wheat, [illegible] practice of
manuring highly [illegible] a [illegible] much [illegible] manure.
the ravages of the fly by making the wheat too vigorous for
them: so that by the improvement of manuring the country will
by have been benefited [illegible] 88. 89. [illegible]
[illegible]
[illegible] some [illegible] years.

The fly lays a maggot [illegible] in the [illegible] in the
fall near the root. this becomes in the winter a [illegible] & [illegible]
eats itself out of it's case in May or June becomes out a fly. the
fly appeared first on Staten island near where the Hessians [illegible]
& the year after they landed. & they have travelled about 20
miles a year more or less according to the winds.

The yellow bearded wheat [illegible] taken up [illegible]
with [illegible] from whom 2 bushels of which was sowed by a
miller accidentally

the fly cannot go Westwardly because repelled by the
Westwardly winds.

Col. Deering of Shelter isld. has [illegible] particular [illegible]
in his [illegible] in this fly.

[illegible] Griffin. the fly was never at this place, nor for [illegible]
[illegible] beyond that distance in every direction a [illegible] the
reason [illegible]. this is 18 miles from Southold.

Moriches. Downs's. came into this neighborhood in 89. as he had a very few the
year before [illegible] he had none since. this neighborhood for
some miles never affected [illegible] cannot account for it

Capt. Floyd. he observed that the fly travelled from West to East about
20 miles a year. but he went a journey N.E. [illegible] they travel
-led much more in that direction because [illegible] before the
wind, that the reason of their being able to fly the S.W. winds prevail
this is the reason they go chiefly to the Eastward.

Smith's. 22 miles from [illegible]. [illegible] here in 85. [illegible]
destroyed some fields of wheat entirely. however they have never
[illegible] as much as on the N. side.
[illegible] on the S. side of [illegible] near as much
they have for many years sowed the bearded wheat. no fly this
year.

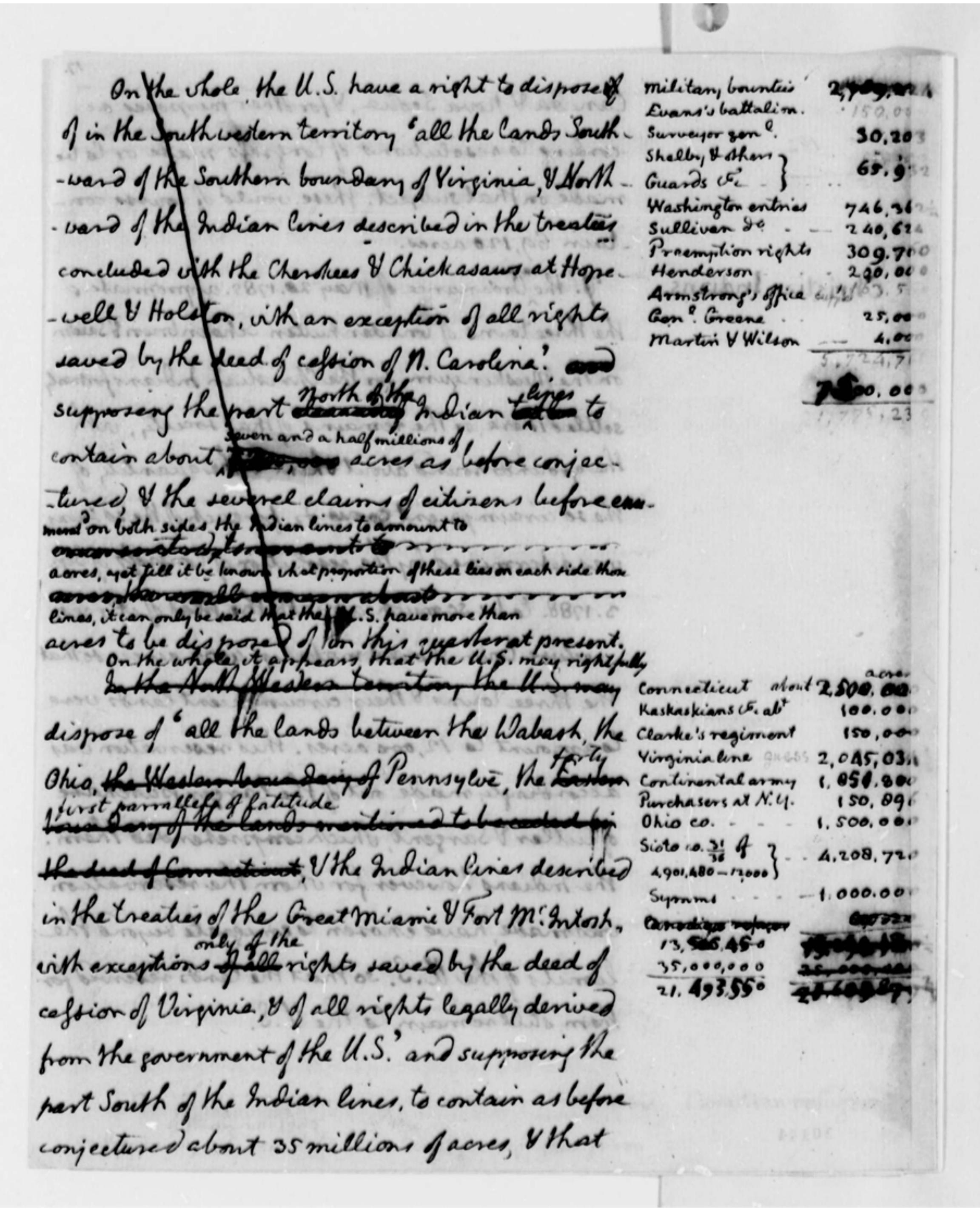
Hampstead. a few here this year.

Flushing. Prince. the insect first appeared at this place. the Hes-
-sians were stationed here in 76. and in 77. the insect appeared
it was very slow indeed in it's first extension, not getting be-
-yond the limits of this neighborhood for some years. he confirms the
acct of the origin of the yellow bearded wheat given by [illegible]
-dien. it was first sown by a miller in this neighborhood

con'd.

QIR Technique

APPENDIX 4: Experimental Results of ICA-Based Segmentation Algorithm



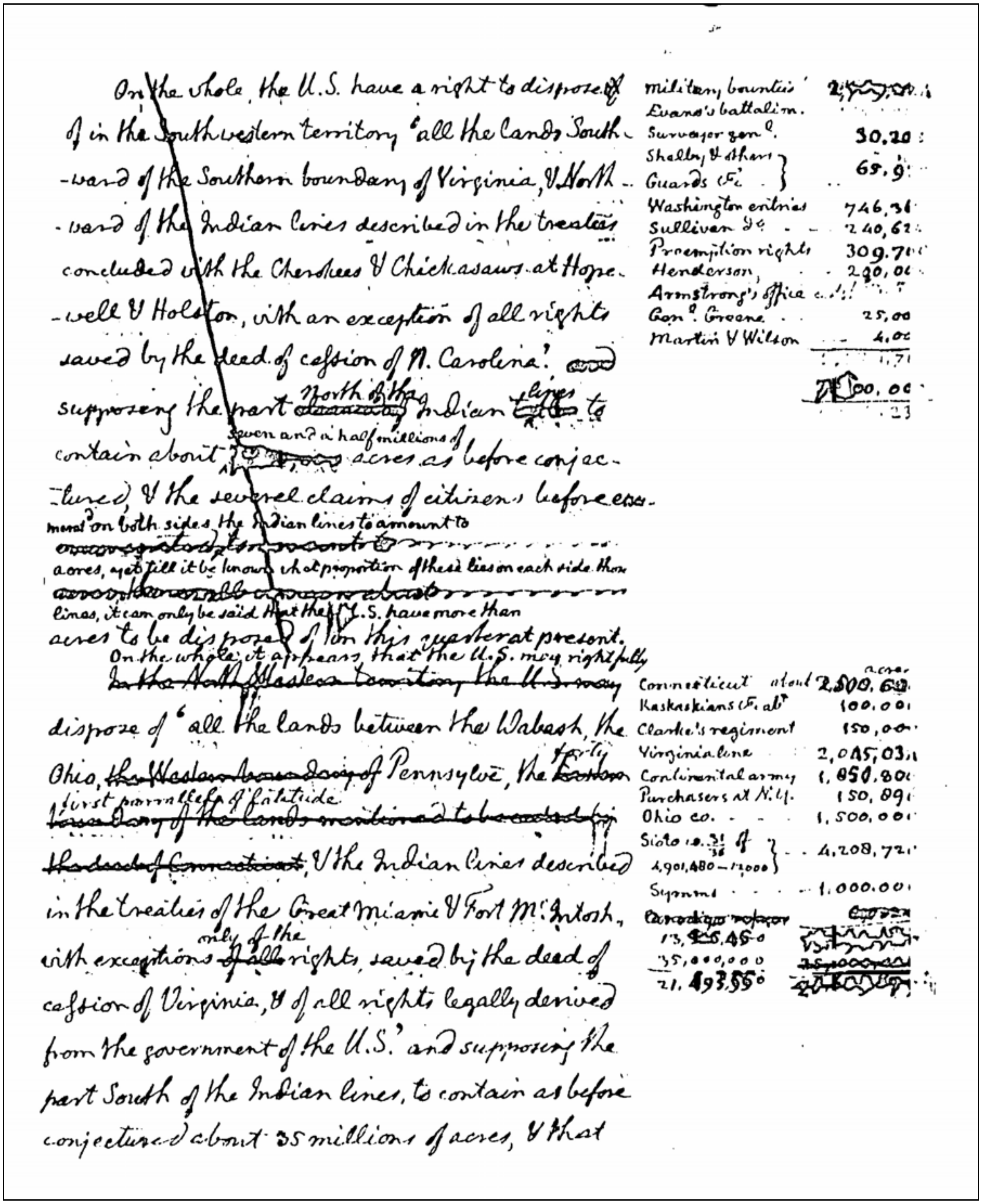
On the whole the U.S. have a right to dispose of in the Southwestern territory 'all the lands Southward of the Southern boundary of Virginia, & Northward of the Indian lines described in the treaties concluded with the Cherokees & Chickasaws at Hopewell & Holston, with an exception of all rights saved by the deed of cession of N. Carolina.' ~~and~~ supposing the part North of the ~~[illegible]~~ Indian lines ~~[illegible]~~ to contain about seven and a half millions of ~~[illegible]~~ acres as before conjectured, & the several claims of citizens before enumerated on both sides the Indian lines to amount to ~~[illegible]~~ acres, yet till it be known what proportion of these lies on each side those ~~[illegible]~~ lines, it can only be said that the U.S. have more than acres to be disposed of on this quarter at present.

militany bounties	2,7[illegible]
Evans's battalion.	150,00
Surveyor genl.	30,203
Shelby & others Guards &c.	65,932
Washington entries	746,36
Sullivan do.	240,624
Praemption rights	309,760
Henderson	200,000
Armstrong's office	[illegible]3.5
Genl. Greene	25,000
Martin & Wilson	4,000
	5,724,71
	7,500,000
	1,775,23

On the whole, it appears that the U.S. may rightfully ~~In the North Western territory, the U.S. may~~ dispose of 'all the lands between the Wabash, the Ohio, ~~the Western boundary~~ of Pennsylvā, the forty first parallel of latitude ~~Eastern boundary of the lands mentioned to be ceded by the deed of Connecticut~~, & the Indian lines described in the treaties of the Great Miami & Fort McIntosh, with exceptions only of the ~~of all~~ rights saved by the deed of cession of Virginia, & of all rights legally derived from the government of the U.S.' and supposing the part South of the Indian lines, to contain as before conjectured about 35 millions of acres, & that

	acres
Connecticut about	2,500,000
Kaskaskians &c. abt	100,000
Clarke's regiment	150,000
Virginia line	2,045,03
Continental army	1,05[illegible],800
Purchasers at N.Y.	150,896
Ohio co.	1,500,000
Sioto co. 31/36 of 4,901,480 − 12,000	4,208,720
Symmes	1,000,000
~~[illegible]~~	~~[illegible]~~
13,506,450	~~[illegible]~~
35,000,000	~~35,000,000~~
21,493,550	~~[illegible]~~

Image 1



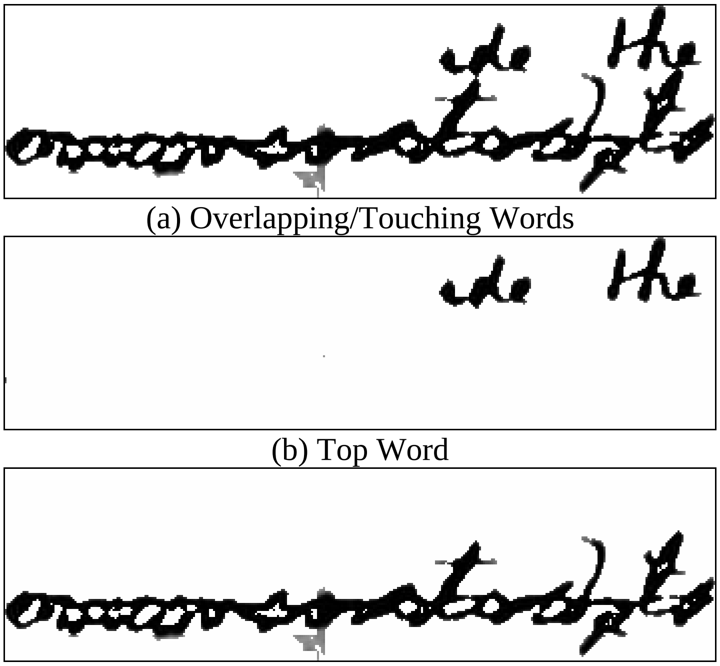
On the whole, the U.S. have a right to dispose of in the Southwestern territory 'all the lands Southward of the Southern boundary of Virginia, & Northward of the Indian lines described in the treaties concluded with the Cherokees & Chickasaws at Hopewell & Holston, with an exception of all rights saved by the deed of ceſsion of N. Carolina.' ~~and~~ supposing the part ~~[illegible]~~ North of the Indian ~~[illegible]~~ lines to contain about seven and a half millions of ~~[illegible]~~ acres as before conjectured, & the several claims of citizens before enumerated on both sides the Indian lines to amount to ~~[illegible]~~ acres, yet till it be known what proportion of these lies on each side those ~~[illegible]~~ lines, it can only be said that the U.S. have more than acres to be disposed of on this quarter at present.

On the whole, it appears that the U.S. may rightfully ~~In the North Western territory, the U.S. may~~ dispose of 'all the lands between the Wabash, the Ohio, ~~the Western boundary~~ of Pennsylva, the ~~Eastern~~ forty first parallels of latitude ~~boundary of the lands mentioned to be ceded by the deed of Connecticut~~, & the Indian lines described in the treaties of the Great Miami & Fort McIntosh, with exceptions ~~of all~~ only of the rights, saved by the dead of ceſsion of Virginia, & of all rights legally derived from the government of the U.S.' and supposing the part South of the Indian lines, to contain as before conjectured about 35 millions of acres, & that

military bounties	2,[illegible]
Evans's battalion.	
Surveyor genl.	30,20
Shelby & others Guards &c.	65,9
Washington entries	746,3
Sullivan do.	240,62
Preemption rights	309,70
Henderson,	200,00
Armstrong's office &c.	[illegible]
Genl. Greene	25,00
Martin & Wilson	4,00
	[illegible]1,71
	7,[illegible]00,00
	23

	acres
Connecticut about	2,500,00
Kaskaskians &c. abt	100,00
Clarke's regiment	150,00
Virginia line	2,045,03
Continental army	1,850,80
Purchasers at N.Y.	150,89
Ohio co.	1,500,00
Sioto co. $\frac{31}{36}$ of 4,901,480 − 12,000	4,208,72
Symmes	1,000,00
~~Connecticut reserve~~	~~[illegible]~~
13,925,450	[illegible]
35,000,000	~~35,000,000~~
21,493,550	~~[illegible]~~

Binary Result of Image 1

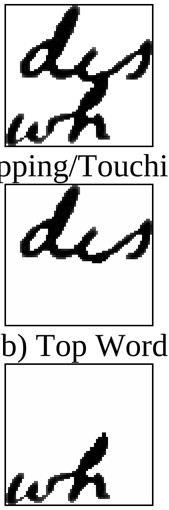


(a) Overlapping/Touching Words

(b) Top Word

(c) Bottom Word

Overlapping/ Touching Words Segmentation Result 1 of Image 1



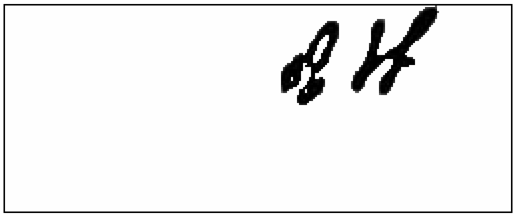
(a) Overlapping/Touching Words

(b) Top Word

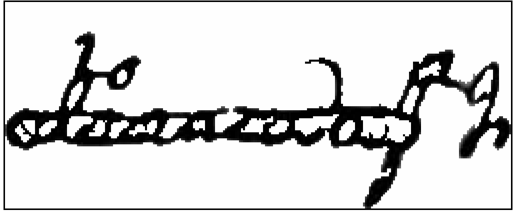
(c) Bottom Word

Overlapping/ Touching Words Segmentation Result 2 of Image 1

(a) Overlapping/Touching Words

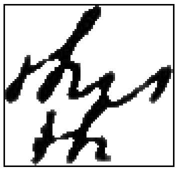


(b) Top Word

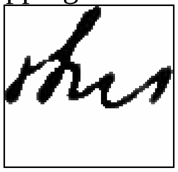


(c) Bottom Word

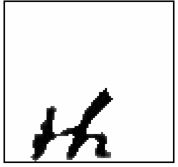
Overlapping/ Touching Words Segmentation Result 3 of Image 1



(a) Overlapping/Touching Words

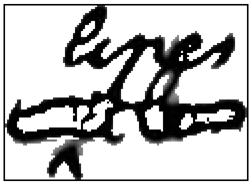


(b) Top Word

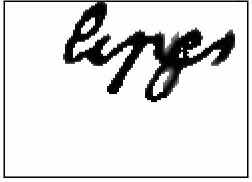


(c) Bottom Word

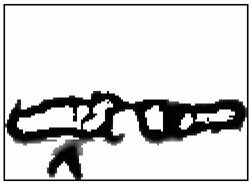
Overlapping/ Touching Words Segmentation Result 4 of Image 1



(a) Overlapping/Touching Words

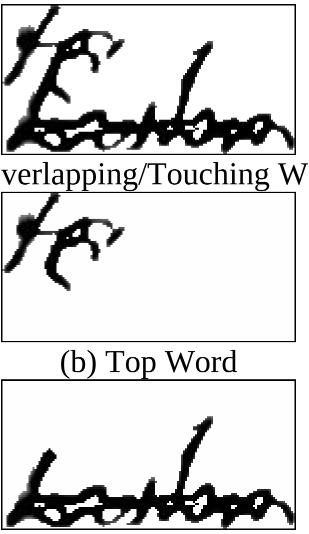


(b) Top Word



(c) Bottom Word

Overlapping/ Touching Words Segmentation Result 5 of Image 1

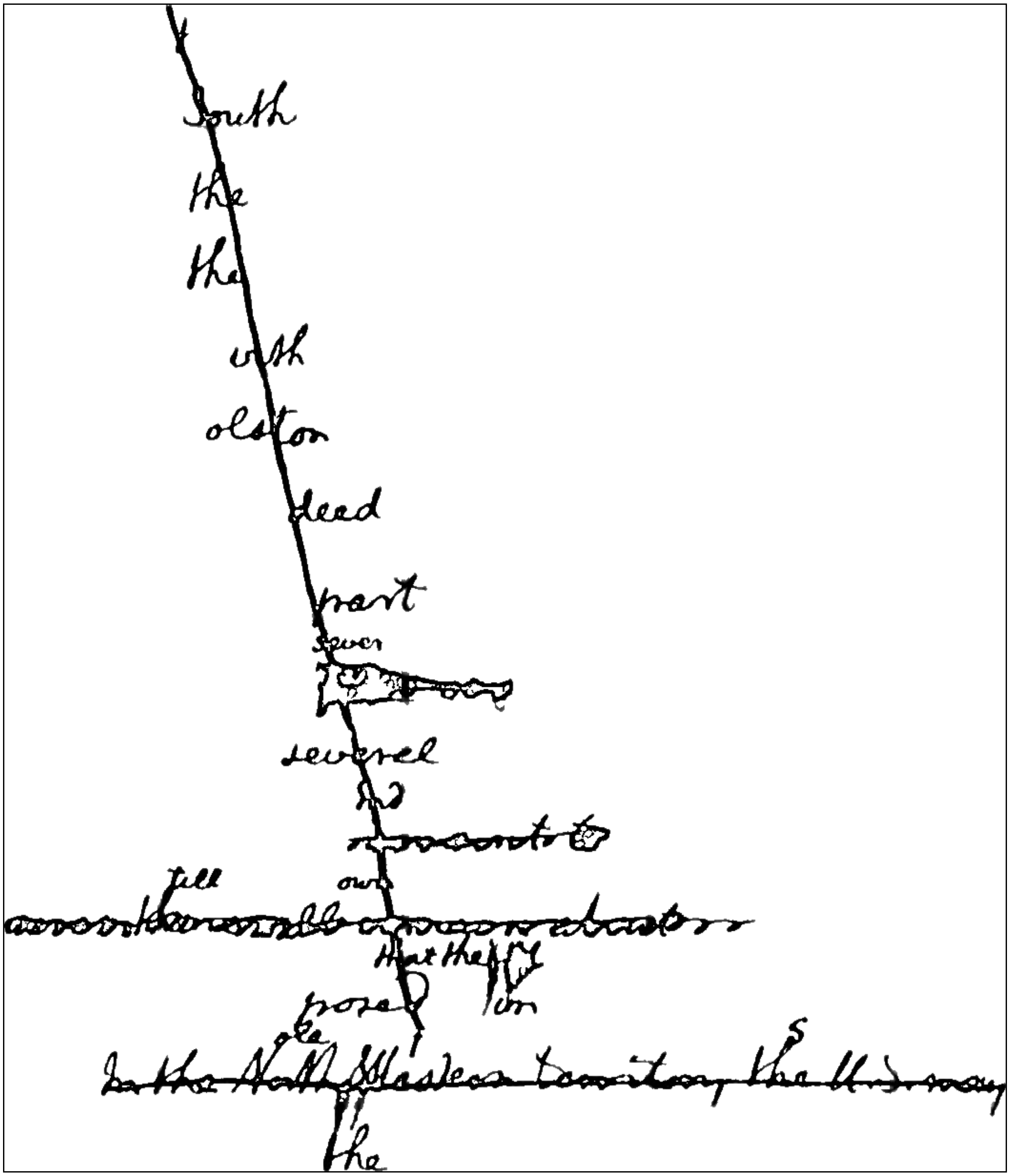


(a) Overlapping/Touching Words

(b) Top Word

(c) Bottom Word

Overlapping/ Touching Words Segmentation Result 6 of Image 1



Overlapping/ Touching Words Segmentation Fail Result 1 of Image 1