

# **Document Image Binarization Technique for Degraded Document Images**

**A Project Report**

**Submitted in partial fulfillment of requirements**

**For the Award of the degree of**

**B.Tech in Computer Science and Engineering**

**By**

**Sudheer Nimmagadda (Y13CS958)**

**Shaheera Shaik Sameen (Y13CS950)**

**Anuragh Samineni (Y13CS940)**

**Under the Guidance of**

**Smt. K. Venkata Ramana**

**Associate Professor, Dept. of CSE**



**APRIL 2017**

**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING**

**R.V.R. & J.C. COLLEGE OF ENGINEERING (Autonomous)**

**(Affiliated to Acharya Nagarjuna University)**

**(Accredited by NBA and NAAC-‘A’ GRADE)**

**Chandramoulipuram::Chowdavaram,**

**GUNTUR – 522 019**

**DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING**  
**R.V.R. & J.C. COLLEGE OF ENGINEERING** (Autonomous)  
Chandramoulipuram::Chowdavaram, GUNTUR – 522 019



**CERTIFICATE**

This is to certify that the project report entitled **“Document Image Binarization Technique for Degraded Document Images”** that is being submitted by **Mr. Sudheer Nimmagadda (Y13CS958)**, **Ms. Shaheera Shaik Sameen (Y13CS950)**, and **Mr. Anuragh Samineni (Y13CS940)** who have carried out the work under my guidance and supervision, and submitted in partial fulfillment for the award of the degree of Bachelor of Technology in Computer Science & Engineering.

Signature of Guide

**Smt. K. Venakata Ramana**

Associate Professor, Dept. of CSE

Signature of HOD

**Dr. M. Sreelatha,**

Professor & Head, CSE.

# ACKNOWLEDGEMENT

The successful completion of any task would be incomplete without a proper suggestion, guidance and environment. Combination of these three factors acts like backbone to our project “**Document Image Binarization Technique for Degraded Document Images**”.

We express our sincere thanks to **Smt. K. Venkata Ramana**, Associate Professor for her timely help, guidance and providing us with the most essential materials for the completion of this report.

We are greatly indebted to our Professor and HOD, Department of Computer Science and Engineering **Dr. M. Sreelatha** and **Sri. P. Venkateswara rao**, Associate Professor for their valuable suggestions during course period and inspiring us to select this paper and for their valuable advices to work on this paper.

We regard our sincere thanks to our Principal, **Dr. K. Srinivasu** for providing support and stimulating environment.

We would like to express our gratitude to the management of R.V.R&J.C College of engineering for providing us with a pleasant environment.

We would be thankful to all the teaching and non-teaching staff of the department of Computer Science and Engineering for the cooperation given for the successful completion.

Sudheer Nimmagadda (Y13CS958)

Shaheera Shaik Sameen (Y13CS950)

Samineni Anuragh (Y13CS940)

## **ABSTRACT**

Image binarization is the separation of each pixel values into two collections, black as a foreground and white as a background. Thresholding technique is used for document image binarization. Image binarization plays vital role in Segmentation of text from badly degraded document images is a very challenging task due to the high inter/intra-variation between the document background and the foreground text of different document images. A novel document image binarization technique that addresses these issues by using adaptive image contrast.

The adaptive image contrast is a combination of the local image contrast and the local image gradient that is tolerant to text and background variation caused by different types of document degradations. In the proposed technique, an adaptive contrast map is first constructed for an input degraded document image.

The contrast map is then binarized and combined with Canny's edge map to identify the text stroke edge pixels. The document text is further segmented by a local threshold that is estimated based on the intensities of detected text stroke edge pixels within a local window. The proposed method is simple, robust, and involves minimum parameter tuning.

# CONTENTS

<b>Abstract</b>	iii
<b>List of Figures</b>	vi
<b>List of Tables</b>	ix
<b>List of Abbreviations</b>	x
<b>1: Introduction</b>	1
1.1 Background	3
1.1.1 Document Image Binarization	3
1.1.2 Degraded Document Images	7
1.1.3 Applications of Document Image Binarization	8
1.2 Problem Statement	10
1.3 Need for present study	11
1.4 Significance of the Work	13
<b>2: Literature Review</b>	14
2.1 Previous work	14
2.2 Challenges on Degraded Document Image Binarization	18
<b>3: System Analysis</b>	20
3.1 System Requirements Specification	20
3.1.1 Functional Requirements	20
3.1.2 Non-Functional Requirements	21
3.2 UML diagrams under Analysis Phase	21
3.2.1 Use Case Diagram	22
3.2.2 Activity Diagram	23
3.2.3 State Chart Diagram	25
<b>4: System Design</b>	27
4.1 Architecture of the proposed system	27
4.2 UML diagrams under Design Phase	27

4.2.1 Sequence Diagram	27
4.2.2 Collaboration Diagram	29
4.3 Workflow of the proposed system	30
4.4 Module Description	31
<b>5: Implementation</b>	38
5.1 Algorithms Used	38
5.2 Description of Datasets	41
5.3 Metrics Calculated	43
5.4 Methods Compared	45
<b>6: Testing</b>	46
6.1 Datasets Tested	46
6.2 Test cases	46
<b>7: Results</b>	52
7.1 Actual Results of the work	52
7.2 Analysis of the results obtained	59
<b>8: Conclusion and Future Work</b>	65
<b>9: References</b>	66

## LIST OF FIGURES

Figure no	Figure description	Page no
1.1	The process of thresholding along with inputs and outputs	4
1.2	Taxonomy of thresholding schemes	5
	Examples of Document analysis problem types in binarization	
1.3	(a) Illumination-object separation	6
	(b) Stain problem or mixed graphics/text	
1.4	Example of good binarization on degraded sample image	7
1.5	A degraded document image	8
2.1	Two degraded document images examples, which are obtained from Document Image Binarization contest (DIBCO) dataset	14
2.2	Binarization Results using Otsu's method of images in Figure. 2.1	14
2.3	Binarization results using Niblack's method of images in Figure. 2.1	15
2.4	Binarization results using sauvola's method of images in Figure. 2.1	16
3.1	Use Case Diagram for whole system	23
3.2	Activity Diagram for whole system	24
3.3	State Chart Diagram for whole system	26
4.1	Architecture of proposed System	27
4.2	Sequence Diagram for whole system	29
4.3	Collaboration Diagram for whole system	30
4.4	The work flow for proposed document image binarization technique	31
4.5	Degraded Document Image taken from DIBCO 2009 dataset	33
4.6	Combination of Local image contrast and Local image gradient	34
4.7	Binary map obtained by using otsu's algorithm	35
4.8	Canny edge map of the sample document in Figure. 4.5 respectively	36

<b>Figure no</b>	<b>Figure Description</b>	<b>Page no</b>
4.9	Combined edge maps of the sample document in Figure. 4.5 respectively	36
5.1	Histogram of the distance between adjacent edge pixels of the image in Figure. 4.5	40
7.1	Degraded Document Image taken from DIBCO 2009 handwritten dataset	52
7.2	Binarization result of the document image for Figure. 7.1 taken from DIBCO handwritten 2009 dataset	53
7.3	Degraded Document image taken from DIBCO 2009 machine printed dataset	53
7.4	Binarized Document Image for Figure. 7.3 taken from DIBCO 2009 machine printed dataset	53
7.5	Degraded Document Image taken from H-DIBCO 2010 dataset	54
7.6	Binarized Document Image for Figure. 7.5 taken from H-DIBCO 2010 dataset	54
7.7	Degraded Document Image taken from H-DIBCO 2010 dataset	54
7.8	Binarized Document Image for Figure. 7.7 taken from H-DIBCO 2010 dataset	54
7.9	Degraded Document image taken from DIBCO handwritten 2011 dataset	55
7.10	Binarized Document Image for Figure. 7.9 taken from DIBCO handwritten 2011 dataset	55
7.11	Degraded Document Image taken from DIBCO 2011 machine printed dataset	55



<b>Figure no</b>	<b>Figure Description</b>	<b>Page no</b>
7.12	Binarized Document Image for Figure. 7.11 taken from DIBCO 2011 machine printed dataset	56
7.13	Badly Degraded Document Image taken from Bickley Diary dataset	57
7.14	Binarized Document Image for badly degraded document image in Figure. 7.13 taken from Bickley Diary dataset	58
7.15	Testing on competition dataset DIBCO 2009	60
7.16	Testing on competition dataset H-DIBCO 2010	61
7.17	Testing on competition dataset DIBCO 2011	62
7.18	Testing on bickley diary dataset	63

## LIST OF TABLES

<b>Table no</b>	<b>Table description</b>	<b>Page no</b>
2.1	Document Image Binarization Methods	19
6.1	Test Case 1	49
6.2	Test Case 2	49
6.3	Test Case 3	49
6.4	Test Case 4	50
6.5	Test Case 5	51

## **LIST OF ABBREVIATIONS**

1. OCR: Optical Character Recognition
2. EW: Estimated stroke edge Width
3. F-measure: Harmonic Mean of recall and precision
4. PSNR: Peak Signal to Noise Ratio
5. NRM: Negative Rate Metric
6. MPM: Misclassification Penalty Metric
7. DRD: Distance Reciprocal Distortion
8. LMM: Local Maximum and Minimum filter
9. BE: Background Estimation
10. OTSU: Otsu's method
11. SAUV: Sauvola's method
12. NIBL: Niblack's method
13. BERN: Bernsen's method
14. GATO: Gatos method
15. LELO: Lore method
16. HOWE: N.Howe's method
17. DIBCO: Document Image Binarization Contest
18. H-DIBCO: Handwritten Document Binarization Contest

# CHAPTER 1

## INTRODUCTION

There is huge amount of textual information that is embedded within images. For example, more and more documents are digitalized everyday via camera, scanner and other equipment, many digital images contain texts, and a large amount of textual information is embedded in web images. It would be very useful to turn the characters from image format to textual format by using optical character recognition (OCR). This converted text information is very important for document mining, document image retrieval and so on. However, in many cases, the document images cannot be directly fed to an OCR system due to the following reasons:

- The original document papers suffer from different kinds of degradation including smear, ink-bleeding through and intensity variation, especially for historical documents.
- The process of obtaining digital images from the real world is not perfect. There are many factors that may cause image distortion, such as incorrect focal length, over/under exposure, camera shaking/object movement, low resolution, etc.

Document Image Enhancement is a technique that improves the quality of a document image to enhance human perception and facilitate subsequent automated image processing. It is widely used in the pre-processing stage of different document analysis tasks. Document image enhancement problem is essentially an ill-posed problem, because a number of enhanced images can be generated from the same input image. Moreover, the quality of enhancement techniques is mainly judged by human perception, which makes the quantitative measures hard to be applied. The main aim of the study is to propose some document image enhancement techniques for better accessibility to the textual information embedded in the images. The specific objectives are:

- Propose some document binarization techniques for degraded document images that achieved good performance for degraded documents and can be used in different document analysis applications.

- Develop better frameworks for improving and combining existing binarization methods by employing domain knowledge and image statistics.

The proposed techniques can be used in different applications, such as optical character recognition, document image retrieval, optical musical recognition, image segmentation, and depth recovery and image retrieval.

Document image binarization (threshold selection) refers to the conversion of a gray-scale image into a binary image. It is the initial step of most document image analysis and understanding systems. Usually, it distinguishes text areas from background areas, so it is used as a text locating technique. Binarization plays a key role in document processing since its performance affects quite critically the degree of success in a subsequent character segmentation and recognition. When processing degraded document images, binarization is not an easy task. Degradations appear frequently and may occur due to several reasons which range from the acquisition source type to environmental conditions. Examples of degradation influence may include the appearance of variable background intensity caused by non-uniform intensity, shadows, smear, smudge and low contrast. Thresholding has created to be a well-known technique used for binarization of document images. Thresholding is further divide into the global and local thresholding technique. In document with uniform contrast delivery of background and foreground, global thresholding is has found to be best technique. In degraded documents, where extensive background noise or difference in contrast and brightness exists i.e. there exists many pixels that cannot be effortlessly categorized as foreground or background. In such cases, local thresholding has significant over available techniques.

Document image binarization is the process that segments the grayscale or color document image into text and background by removing any existing degradations (such as bleed-through, large ink stains, non-uniform illumination and faint characters). It is an important pre-processing step of the document image processing and analysis pipeline that affects further stages as well as the final Optical Character Recognition (OCR) stage. To analyze the document, its image is binarized before processing it. It is nothing but segmenting the document background & the foreground text. For the confirmation of document image processing task an accurate document image binarization technique is a must.

## **1.1 Background**

Segmentation subdivides an image into its constituent regions or objects. The level to which the subdivision is carried depends on the problem being solved. That is, segmentation should stop when the objects of interest have been isolated. For example, in the automated inspection of electronic assemblies, interest lies in analyzing images of the products with the objective of determining the presence or absence of specific anomalies, such as missing components or broken connection paths. There is no reason to carry segmentation past the level of detail required to identify those elements.

Segmentation of nontrivial images is one of the most difficult tasks in image processing. Segmentation accuracy determines the eventual success or failure of computerized analysis procedures. For this reason, considerable care should be taken to improve the probability of rugged segmentation. In some situations, such as industrial inspection applications, at least some measure of control over the environment is possible at times. In others, as in remote sensing, user control over image acquisition is limited principally to the choice of imaging sensors.

Segmentation algorithms for monochrome images generally are based on one of two basic properties of image intensity values: discontinuity and similarity. In the first category, the approach is to partition an image based on abrupt changes in intensity, such as edges. The principal approaches in the second category are based on partitioning an image into regions that are similar according to a set of predefined criteria.

### **1.1.1 Document Image Binarization**

Optical scanning of the rock inscription yields an image (file of pixels) that forms the raw input to the Optical Character Recognition System. The output is the set of recognized characters. Pre-processing is the first phase of document analysis. The purpose of pre-processing is to improve the quality of the image being processed. It makes the subsequent phases of image processing like recognition of characters easier. Thresholding is one of the pre-processing methods. In thresholding, the colour image or grey scale image is reduced to a binary image.



Figure. 1.1 The process of thresholding along with its inputs and outputs.

Most document analysis algorithms are built on taking advantage of the underlying binarized image data. The use of a bi-level information decreases the computational load and enables the utilization of the simplified analysis methods compared to 256 levels of grey-scale or colour image information. Document image understanding methods require logical and semantic content preservation during thresholding. For example, a letter connectivity must be maintained for optical character recognition and textual compression. This requirement narrows down the use of a global threshold in many cases. Binarization has been a subject of intense research interest during the last ten years. Most of the developed algorithms rely on statistical methods, not considering the special nature of document images. However, recent developments on document types, for example documents with mixed text and graphics, call for more specialized binarization techniques.

In current techniques [5], the binarization (threshold selection) is usually performed either globally or locally. Some hybrid methods have also been proposed. The global methods use one calculated threshold value to divide image pixels into object or background classes, whereas the local schemes can use many different adapted values selected according to the local area information. Hybrid methods use both global and local information to decide the pixel label.

The main situations in which single global thresholds are not sufficient are caused by changes in lamination (illumination), scanning errors and resolution, poor quality of the source document and complexity in the document structure (e.g. graphics is mixed with text). When character recognition is performed, the melted sets of pixel clusters (characters) are easily misinterpreted if binarization labelling has not successfully separated the clusters. Other misinterpretations occur easily if meant to be clusters are wrongly divided. Figure. 1.2 depicts the taxonomy (called MSLG) and general division into thresholding techniques according to

level of semantics and locality of processing used. The MSLG can be applied in pairs, for example (ML), (SL), (MG) and (SG).

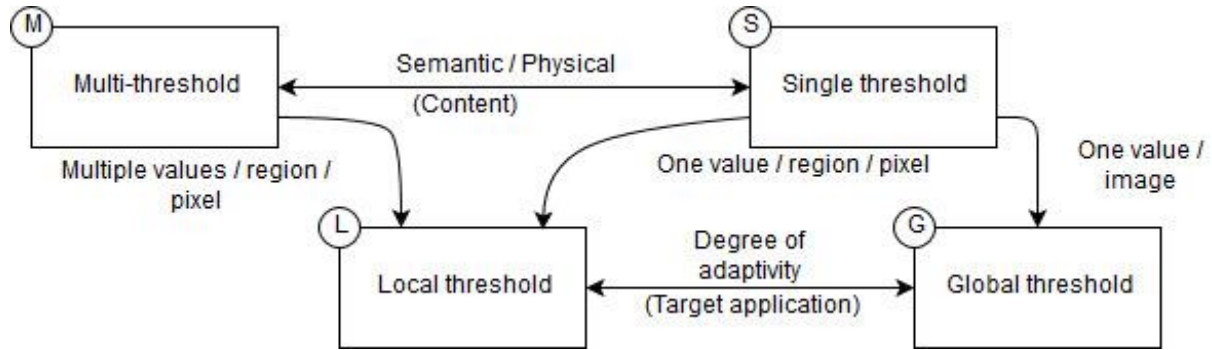


Fig. 1.2 Taxonomy of thresholding schemes

The most conventional approach is a global threshold, where one threshold value (single threshold) is selected for the entire image according to global/local information. In local thresholding the threshold values are determined locally, e.g. pixel by pixel, or region by region. Then, a specified region can have single threshold that is changed from region to region according to threshold candidate selection for a given area. Multi-thresholding is a scheme, where image semantics are evaluated. Then, each pixel can have more than one threshold value depending on the connectivity or other semantic dependency related to physical, logical or pictorial contents.

Many binarization techniques that are used in processing tasks are aimed at simplifying and unifying the image data at hand. The simplification is performed to benefit the oncoming processing characteristics, such as computational load, algorithm complexity and real-time requirements in industrial like environments. One of the key reasons when the binarization step fails to provide the subsequent processing a high-quality data is caused by the different types and degrees of degradation introduced to the source image. The reasons for the degradation may vary from poor source type, the image acquisition process to the environment that causes problems for the image quality directly. Since the degradation is unquestionably one of the main reasons for processing to fail, it is very important to design the binarization technique to detect and alter possible imperfections from becoming the subject for processing and potential cause of errors for post-processing steps. Most degradation types in document images affect both physical and



semantic understand ability in the document analysis tasks, such as page segmentation, classification and optical character recognition. Therefore, the result after all the desired processing steps can be entirely unacceptable, just because of the poorly performed binarization.

Figure. 1.3 depicts two types of typical degradation, when dealing with scanned grey-scale document images. In Figure. 1.3(a) the threshold 'base line' is changing due to illumination effect or implanted (designed) entity. Then, each object has a different base level that affects the object/non-object separation decision in selecting thresholds. In Figure. 1.3 (b) a general type stain problem is presented. In this case, the background and object levels are fluctuating from clear separation to non-clear separation and small level difference between object/non-object. The optimal threshold lines are drawn to both images to depict the base line that a successful binarization algorithm should mimic.

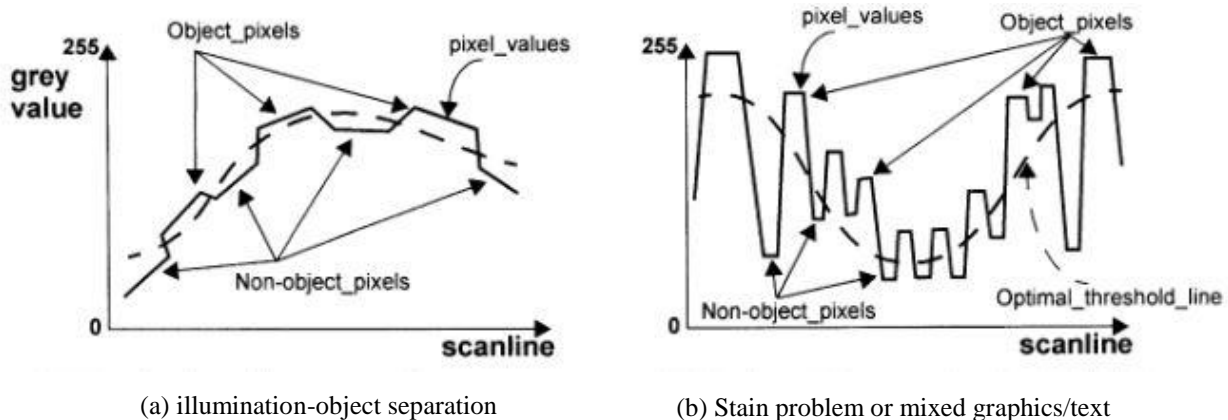


Figure. 1.3 Examples of Document analysis problem types in binarization.

Figure. 1.4 presents another type of problem, frequently occurring in scanned document images: more than two different levels are visible in textual areas due to transparency of the next page. Then, a binarization algorithm should cope with at least two different threshold candidates: background-transparent text and background text. The binarized example presents a correct binarization result.



Fig. 1.4 Example of good binarization on degraded sample image

The research on binarization techniques originates from the traditional ‘scene image’ processing needs to optimize the image processing tasks in terms of image data at hand. While the image types have become more complex the algorithms developed have gained wider theoretical grounds. Current trend seems to move forward image domain understanding based binarization and the control of different source image types and qualities. The state-of-the-art techniques are able to adapt to some degree of errors in a defined category, and focus on few image types. In images needing multi-thresholding, the problem seems to be ever harder to solve, since the complexity of image contents, including textual documents has increased rapidly.

### 1.1.2 Degraded Document Images

Recent years have witnessed the rapid growth of degraded images due to the increasing power of computing and the fast development of Internet. Because of this tremendous increase of quality of degraded images, there is an urgent need of image content description to facilitate automatic retrieval. Image is described by several low level image features, such as colour, texture, shape or the combination of these features. Shape is an important low level image feature. Image processing is very vast field & one of the most important part of image processing is thresholding. Thresholding, which is an important pre-processing steps for the degraded image to enhance their quality & has been studies in relation to various images. There are different algorithms that have been used & studied for various factors of image analysis [33]. The value of thresholding is based on which segmentation has been performed.

Poor quality documents are obtained in various situations such as historical document collections, legal archives, security investigations, and documents found in clandestine locations. Such documents are often scanned for automated analysis, further processing, and archiving.

Due to the nature of such documents, degraded document images are often hard to read, have low contrast, and are corrupted by various artifacts.

Degraded documents are archived and preserved in large quantities worldwide. Electronic scanning is a common approach in handling such documents in a manner which facilitates public access to them. Such document images are often hard to read, have low contrast, and are corrupted by various artifacts. Thus, given an image of a faded, washed out, damaged, crumpled or otherwise difficult to read document, one with mixed handwriting, typed or printed material, with possible pictures, tables or diagrams, it is necessary to enhance its readability and comprehensibility. Documents might have multiple languages in a single page and contain both handwritten and machine printed text. Machine printed text might have been produced using various technologies with variable quality. It is common for libraries to provide public access to historical and ancient document image collections. For such type of document images require specialized processing in order to remove background noise and become more legible.



Fig. 1.5 A degraded Document Image

### 1.1.3 Applications of Document Image Binarization

#### Optical Character Recognition

Optical character recognition (also optical character reader, OCR) is the mechanical or electronic conversion of images of typed, handwritten or printed text into machine-encoded text, whether from a scanned document, a photo of a document, a scene-photo (for example the text

on signs and billboards in a landscape photo) or from subtitle text superimposed on an image (for example from a television broadcast). It is widely used as a form of information entry from printed paper data records, whether passport documents, invoices, bank statements, computerized receipts, business cards, mail, printouts of static-data, or any suitable documentation. It is a common method of digitizing printed texts so that they can be electronically edited, searched, stored more compactly, displayed on-line, and used in machine processes such as cognitive computing, machine translation, (extracted) text-to-speech, key data and text mining. OCR is a field of research in pattern recognition, artificial intelligence and computer vision. So, Document image binarization is usually performed in the preprocessing stage of different document image processing related applications such as optical character recognition (OCR).

### **Document Image Retrieval**

Document image binarization is usually performed in the preprocessing stage of different document image processing related applications such as document image retrieval. Document image retrieval is a very attractive field of research with the continuous growth of interest and increasing security requirements for the development of the modern society. Good documents essentially play an important role in our day to day life. Complex documents present a great challenge to the field of document recognition and retrieval. The primary task of processing these complex documents is to isolate the different contents present in the documents. Once the contents are separated out, then they can now be called as indexed documents which are ready to use for a content-based image retrieval system. The document image understanding, covering a variety of documents such as bank checks, business letters, forms, and technical articles, has been an interesting research area for a long time. In the context of document image retrieval, logo provides an important form of indexing that enable effective explanation of data. Given a large collection of documents, searching for a specific logo is a highly effective way of retrieving documents from the associated organization. Building an effective access to these document images requires designing a mechanism for effective search and retrieval of image data from document image collection.

## 1.2 Problem Statement

The Technology is connecting the whole world together by the medium of internet. Each segment of the data is present in the form of digital document. People are able to store, duplicate, and backup the data in digital form. But what about the old data which is available in the form of traditional document. Sometimes the old documents plays important role in a major challenge. Many of the paper data is being degraded due to lack of attention. Many of these degraded documents have their front data mix up with rear data. To make this front data separate from backend data a good binarization technique is processed on the degraded documents.

Though document image binarization has been studied for many years, the thresholding of degraded document images is still an unsolved problem due to the high inter/intra-variation between the text stroke and the document background across different document images. The handwritten text within the degraded documents often shows a certain amount of variation in terms of the stroke width, stroke brightness, stroke connection, and document background. In addition, historical documents are often degraded by the bleed through where the ink of the other side seeps through to the front. In addition, historical documents are often degraded by different types of imaging artifacts. These different types of document degradations tend to induce the document thresholding error and make degraded document image binarization a big challenge to most state-of-the-art techniques. Text Segmentation from a degraded document images is a very difficult task as the document image might contain lot of variations between the foreground and the background part.

Examples of degradations include shadows and non-uniform illumination, ink seeping, smear and strains. To deal with degradations, the current trend is to use local information that guides the threshold value pixel wise in an adaptive manner. Most of the adaptive local binarization methods ignore the edge property and lead to erroneous results due to the creation of fake shadows. For this, there exist approaches that also incorporate edge information as in wherein they find seeds near the image edges and present an edge connection method to close the image edges. Then, they use closed image edges to partition the binarized image that is generated using a high threshold, and obtain a primary binarization result by filling the partitioned high-

threshold binary image with the seeds. An effective solution is developed in the case of low contrast, noise and non-uniform illumination.

The degraded images are in the form of mixed foreground and background format. Hence, it is important to separate this background from the foreground text. Many thresholding [22], [23], [13], [12] techniques have been proposed for document image binarization. The concept of adaptive image contrast is introduced in the proposed method which is tolerant to text and background variation that is caused by different types of document degradations. As many degraded documents do not have a clear bimodal pattern, global thresholding [7], [2], [1], [11] is usually not a suitable approach for the degraded document binarization. Adaptive thresholding [3], [6], [21], [9], [18], [5], [10], which estimates a local threshold for each document image pixel, is often a better approach to deal with different variations within degraded document images.

### **1.3 Need for Present Study**

The proposed work is to segment text from badly degraded document images. Some of the drawbacks in existing methods are listed here.

They are-

1. Existing techniques are not able to binarize the degraded/handwritten document images.
2. The high inter/intra variation between the text stroke and the document background is still an unsolved problem.
3. The over-normalization problem of the local maximum minimum algorithm is not solved under existing technique.
4. Bernsen's [3] method is simple but cannot work properly on degraded document images with a complex document background.
5. The existing techniques does not deal with minimum parameter tuning and cannot combine different types of image information and domain knowledge and are often complex.

The proposed method is to eliminate the above challenges and to improve performance on the datasets. A new method is designed. The proposed method is

1. Able to overcome normalization problem
2. Performance is going to be improved
3. Deals with minimum parameter tuning
4. Deals document images with complex background

With the above stated existing method problems, the performance of the system will degrade. To improve the performance and accuracy the present study is helpful. The proposed method will produce stable, good results. In Otsu's [1] method cluster-based image thresholding, has been used for the reduction of a gray level image to a binary image. The algorithm tries to reduce combined spread (intra class variance) by assuming that the image contains two classes of pixels.

It assumes that an image follows a bimodal histogram i.e. it contains foreground and background pixels. It then calculates the optimum threshold separating the two classes to ensure that its combined spread is minimal. This method gives acceptable results when the pixels in each class are close to each other. The limitations of this method are that many degraded document images do not have a clear bimodal pattern. Also another limitation is that minimization of intra class variances maximizes between class scatter.

Another approach for document images binarization has been adopted by Sauvola [18]. In the sauvola's [18] method the page is considered as a collection of subcomponents such as text, background and picture. To define a threshold for each pixel of the background and pictures a soft decision method is used. To define a threshold for each pixel of textual and line drawing areas a text binarization method is used. Finally the results of these algorithms are combined. Although sauvola's [18] method solves the problem posed by Niblack's [5] approach but in many cases the characters become extremely thinned and broken.

To eliminate the above challenges the proposed method is helpful and the proposed method is tolerant to different types of document degradations such as uneven illumination and document smear. The proposed method is simple, robust, only few parameters are involved. It also works for complex background document images. The proposed method which uses the adaptive image contrast map is first constructed and the text stroke edges are then detected through the combination of the binarized adaptive contrast map and the canny edge map.

The text is then segmented based on local threshold that is estimated from the detected text stroke edge pixels. And the post processing is further applied to improve the document binarization quality. The proposed method is simple, robust and capable of handling different types of degraded document images with minimum parameter tuning.

The proposed method makes use of adaptive image contrast that combines the local image contrast and local image gradient adaptively and therefore is tolerant to the text and background variation caused by different types of document degradations. In particular, the proposed method addresses the over-normalization problem of the local maximum minimum algorithm [39]. At the same time, the parameters used in the algorithm can be adaptively estimated.

#### **1.4 Significance of the work**

To achieve the above mentioned needs the project work is organized into nine chapters.

The Chapter 1 deals with introduction to importance of document image binarization in the pre-processing stage and applications, problem statement, need for present study and scope of the present study are detailed. Chapter 2 presents various challenges involved in thresholding of degraded document images by various binarization methods and detailed literature survey on existing methods. Chapter 3 presents system requirements specification adopted for binarization of degraded document images and UML diagrams related to the binarization system. Chapter 4 presents in detail description of the proposed system which includes adaptive image contrast, text stroke edge pixel detection, local threshold estimation and post processing procedure. Chapter 5 presents algorithms involved in local threshold estimation, post processing procedure and connected components algorithm in post processing procedure, results and discussions on various degraded document images. Chapter 6 presents various types of testing and test cases involved in document image binarization. Chapter 7 presents results of algorithms and result analysis involved in degraded document image binarization. Chapter 8 describes about conclusions of the proposed work and future work to be addressed by researchers. Chapter 9 describes the references that are in connection with the proposed method.



## CHAPTER 2

### LITERATURE REVIEW

Document image binarization is usually performed in the preprocessing stage of different document image processing related applications such as optical character recognition (OCR) and document image retrieval. It converts a gray-scale document image into a binary document image and accordingly facilitates the ensuing tasks such as document skew estimation and document layout analysis. As more and more text documents are scanned, fast and accurate document image binarization is becoming increasingly important.

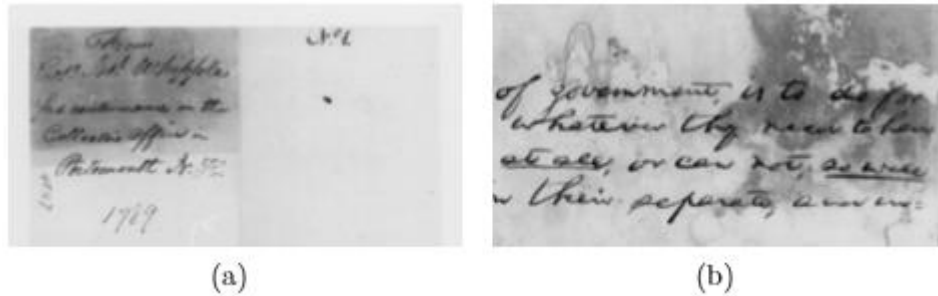
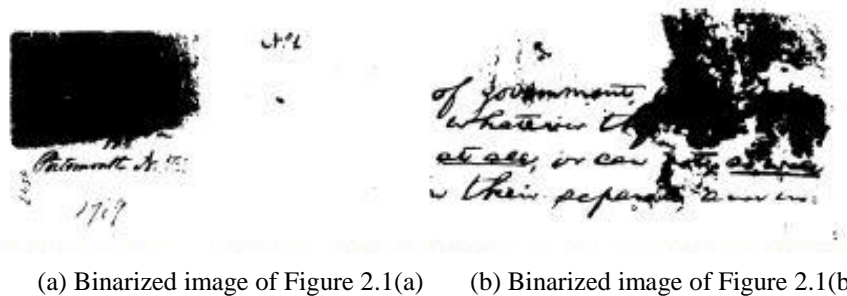


Figure. 2.1 Two degraded document images examples, which are obtained from Document Image Binarization contest (DIBCO) dataset



(a) Binarized image of Figure 2.1(a) (b) Binarized image of Figure 2.1(b)

Figure. 2.2 Binarization Results using Otsu's method of images in Figure.2.1

### 2.1 Previous Work

Generally speaking, the binarization techniques are either global or local. The global binarization techniques assign a single threshold for the whole document image, and the local binarization techniques find a threshold for each pixel in the document image. One of the famous

global thresholding methods is Otsu's method [1], which is a histogram shape-based image thresholding technique. Otsu's method tries to estimate a global threshold that minimizes the intra-class variance, which is defined as a weighted sum of variances of the two classes:

$$\sigma_{\omega}^2 = \omega_1(t)\sigma_1^2(t) + \omega_2(t)\sigma_2^2(t) \quad (2.1.1)$$



(a) Binarized image of Figure 2.1(a)

(b) Binarized image of Figure 2.1(b)

Figure. 2.3 Binarization results using Niblack's method of images in Figure 2.1

where the term  $\omega_i$  is the probabilities of the two classes separated by a threshold  $t$  and the variances of these classes  $\delta_i$ . The term  $\omega_i$  is defined as follows:

$$\omega_1 = \sum_{i=1}^{t-1} p(i) ; \omega_2 = \sum_{i=t}^n p(i) \quad (2.1.2)$$

where the variable  $p(i)$  denotes the number of pixels with gray value level  $i$ . And one of the famous local thresholding methods is Niblack's method [5], which estimates the local threshold by using the local mean  $m$  and the standard variation  $s$ . The local threshold is computed as follows:

$$T = m + K \cdot s \quad (2.1.3)$$

where the parameter  $k$  is a user defined parameter and it normally lies between -1 and 0. The main drawback of this window based thresholding approach is that the thresholding performance depends heavily on the window size and hence the character stroke width. Sauvola et. al. [18] later modify the formula in Equation 2.1.3 and propose a new thresholding formula as follows:

$$T = m \cdot \left( 1 + K \cdot \left( \frac{s}{R} - 1 \right) \right) \quad (2.1.4)$$

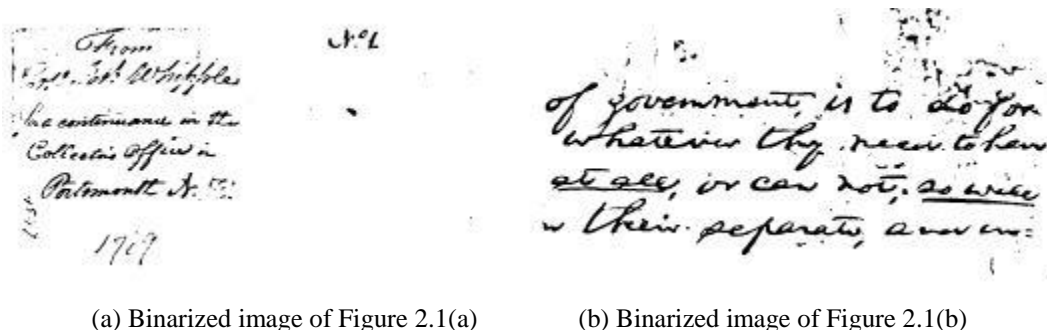


Figure. 2.4 Binarization results using sauvola's method of images in figure. 2.1

where the parameter  $R$  refers to the dynamic range of the standard deviation and the parameter  $k$  instead takes a positive value between 0 and 1. The new thresholding formulas reduce the background noise greatly, but it requires the knowledge of document contrast to set the parameter  $R$  properly.

Figures 2.2, 2.3, 2.4 show the binarization results of the sample document images in Figure 2.1. As shown in the results, Otsu's method [1] requires a bimodal histogram pattern and so cannot handle these document image with severe background variation. Adaptive thresholding methods such as Niblack's/ Sauvola's [5], [18] method may either introduce a certain amount of noise or fail to detect the document text with a low image contrast. Many works [22], [13] have been reported to deal with the high variation within historical document images. As many historical documents do not have a clear bimodal pattern, global thresholding [7], [2], [1] is usually not a suitable approach for the historical document binarization. Adaptive thresholding [3], [6], [21], [9], [18], [5], which estimates a local threshold for each document image pixel, is usually a better approach to handle the high variation associated with historical document images. For example, the early window-based adaptive thresholding [18], [5] techniques estimate the local threshold by using the mean and the standard variation of image pixels within a local neighbourhood window.

There are other approaches have been developed. Background Subtraction [42], [26] tries to subtract a background from the degraded images and use it to binarized the document images, however it is hard to model the document background and separate it from foreground text. Image contrast and edge information [3] which are good indicators of text strokes are used to remove the non-uniform background, although it is difficult to identify the difference between

text stroke edges and document background noise. Some domain knowledge such as Texture feature [14] and cross section sequence graph analysis [28] can also be used to produce better results. But they requires some prior knowledge to the testing document images. Decomposition method [25] tried to divide the document images into smaller regions which are more uniform and easier to be binarized. Energy-based method [44] employs graph-cut algorithm to segment text information by minimizing Laplacian energy. Other approaches have also been intended, including background subtraction [43], [26], texture analysis [14], recursive method [16], [24], decomposition method [25], contour completion [34], [5], [19], Markov Random Field [31], [10], [37], [33], matched wavelet [30], cross section sequence graph analysis [28], self-learning [38], Laplacian energy [44] user assistance [40], [41] and combination of binarization techniques [29], [32]. These methods combine different types of image information and domain knowledge and are often complex. In conclusion, these approaches combine different types of image information and domain knowledge and are often complex and time consuming.

### **Document Image Binarization using Background Estimation**

A document binarization technique which uses background estimation [43] and stroke edge information. The document binarization technique is based on the observations that the text documents usually have a document background of the uniform colour and texture and the document text within it has a different intensity level compared with the surrounding background. The technique makes use of the document background surface and the text stroke edge information. It first estimates a document background surface through an iterative polynomial smoothing procedure. The text stroke edges are then detected by combining the local image variation and the estimated document background surface. After that, the document text is segmented based on the local threshold that is estimated from the detected stroke edge pixels. At the end, a series of post-processing operations are performed to further improve the document binarization performance.

### **Document Image Binarization using Local Maximum and Minimum**

A simple but efficient historical document image binarization technique that is tolerant to different types of document degradation such as uneven illumination and document smear. The

technique makes use of the image contrast that is evaluated based on the local maximum and minimum [39]. Given a document image, it first constructs a contrast image and then extracts the high contrast image pixels by using Otsu’s global thresholding method. After that, the text pixels are classified based on the local threshold that is estimated from the detected high contrast image pixels. The method has also been tested on the dataset that is used in the recent DIBCO contest series.

## 2.2 Challenges on Degraded Document Image Binarization

Table 2.1 shows most state-of-the-art document image binarization techniques with their strengths and weaknesses.

Table 2.1 Document Image Binarization Methods

Methods	Pros	Cons
Global Thresholding	Fast, Produce good results on clean documents	Fail on degraded images
Local Thresholding	Works on degraded documents	Sensitive to window size
Background subtraction	Produce good results when foreground varies	Performance decreased when background non-uniform
Image Contrast	Produce good results when background varies	Performance decreased when foreground non-uniform
Domain knowledge	Preserve text info using domain knowledge	Hard to extract proper domain knowledge
Energy based	Simple but effective	Need to tune a few parameters

Though document image binarization has been studied for many years, the thresholding of degraded document images is still an unsolved problem. This can be explained by the fact that the modelling of the document foreground/background is very difficult due to various types of document degradation such as uneven illumination, image contrast variation, bleeding-through, and smear as illustrated in Figure 2.1. The recent Document Image Binarization Contests (DIBCO) [35], [45] held under the framework of the International Conference on Document

Analysis and Recognition (ICDAR) 2009 and 2011 and the Handwritten Document Image Binarization Contest (H-DIBCO) [42] held under the framework of the International Conference on Frontiers in Handwritten Recognition (ICFHR) show recent efforts on this issue. These contests partially reflect the current efforts on this task as well as the common understanding that further efforts are required for better document image binarization solutions. Many practical document image binarization techniques have been applied on the commercial document image processing systems. These techniques perform well on the documents which do not suffer from serious document degradation. However, the degraded document image binarization is not fully explored and still needs further research.

## **CHAPTER 3**

### **SYSTEM ANALYSIS**

#### **3.1 System Requirements Specification**

Software requirements are:-

- Image processing tool box

In addition to the simple functionality system will also support three simple image processing tools.

- Crop
- Scale
- Contrast Adjustment

##### **Crop Tool**

The crop tool will allow single images to be cropped at a time. The user will be able to select the image from the File Viewer, and define the parameters of the crop. The crop transformation then applied to the image.

##### **Scale Tool**

The scale tool will allow the user to select and rescale a single image in the library the user will be able to select the image from the File View, and then apply scale transformation based on new height, width, or scale parameters supplied by the user.

##### **Contrast Adjustment**

The contrast adjustment will allow the user to select the single image, and change the brightness, and contrast parameters of the image. The user will be able to select the image from the File View and apply the adjustments.

#### **3.1.1 Functional Requirements**

##### **Input:**

DIBCO(2009/2011) datasets, H-DIBCO(2010) dataset ,Bickley diary dataset

## **Output:**

Gives the:

- Adaptive image contrast image
- Combined edge map image
- Initial binarization result image
- Final Binarization result image
- Evaluation results of the image as well as the whole dataset

### **3.1.2 Non-Functional Requirements**

#### **1. Execution qualities:**

- Robustness
- Efficiency

#### **2. Evolution qualities:**

- Testability
- Maintainability
- Extensibility
- Scalability

### **3.2 UML Diagrams under Analysis Phase**

The **Unified Modeling Language (UML)** is a general-purpose modeling language in the field of software engineering, which is designed to provide a standard way to visualize the design of a system. The Unified Modeling Language (UML) offers a way to visualize a system's architectural blueprints in a diagram (see image), including elements such as:

- Any activities (jobs)
- Individual components of the system and how they can interact with other software components.
- How the system will run
- How entities interact with others (components and interfaces)



- External user interface
- How the system is expected to be used.

Although originally intended solely for object-oriented design documentation, the Unified Modeling Language (UML) has been extended to cover a larger set of design documentation (as listed above), and been found useful in many contexts.

### **3.2.1 Use Case Diagram**

A use case diagram is a graph of actors, a set of use cases enclosed by a system boundary, communication (participation) associations between the actors and users and generalization among use cases. The use case model defines the outside (actors) and inside (use case) of the system's behavior. Actors are NOT part of the system. Actors represent anyone or anything that interacts with (input to or receive output from) the system. An actor is someone or something that:

1. Interacts with or uses the system
2. Provides input to and receives information from the system
3. is external to the system and has no control over the use cases

Actors are discovered by examining:

1. Who directly uses the system, which is responsible for maintaining the system?
2. External hardware used by the system
3. Other systems that need to interact with the system

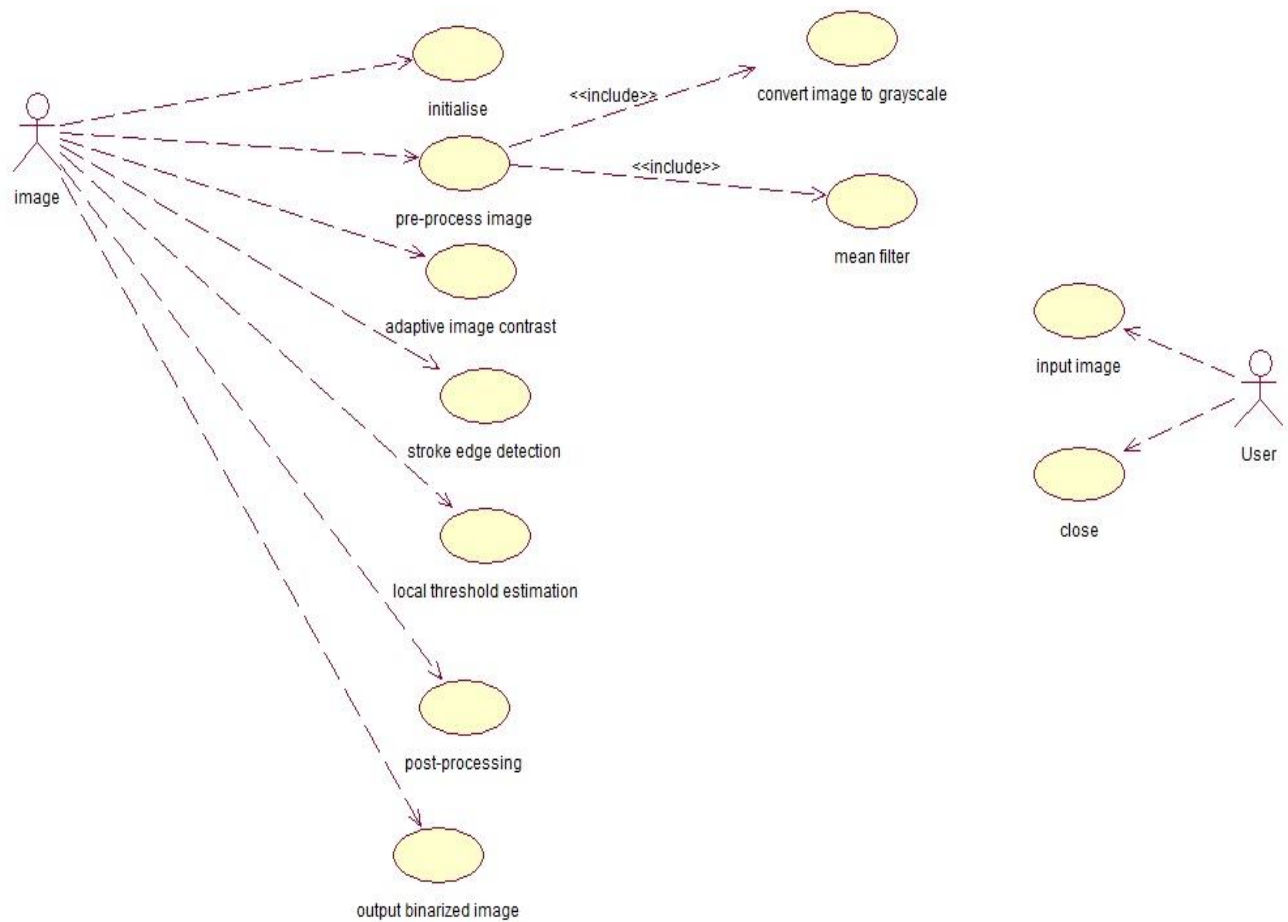


Figure. 3.1 Use Case Diagram for whole system

### 3.2.2 Activity Diagram

Activity diagram is a variation or a special case of a state machine, in which the states are activities representing the performance of operations and the transactions are triggered by the completion of operations. Unlike state diagrams focus on the events occurring to a single object as it responds to messages, an activity diagram can be used to model an entire business process. The purpose of activity diagram is to provide a view of flows and what is going on inside a use case or among several classes. An activity is shown in around box containing the name of the operation. When an operation symbol appears within an activity diagram, it indicates the

execution of object. Executing a particular step within the diagram represents a state within the execution of the overall method.

An activity diagram is used mostly to show the internal state of the object, but external events may appear in them. Actions may be organized into swim lanes, each separated from neighboring swim lanes by vertical solid lines on both sides. Each swim lane represents responsibility for part of the overall activity and may be implemented by one or more objects.

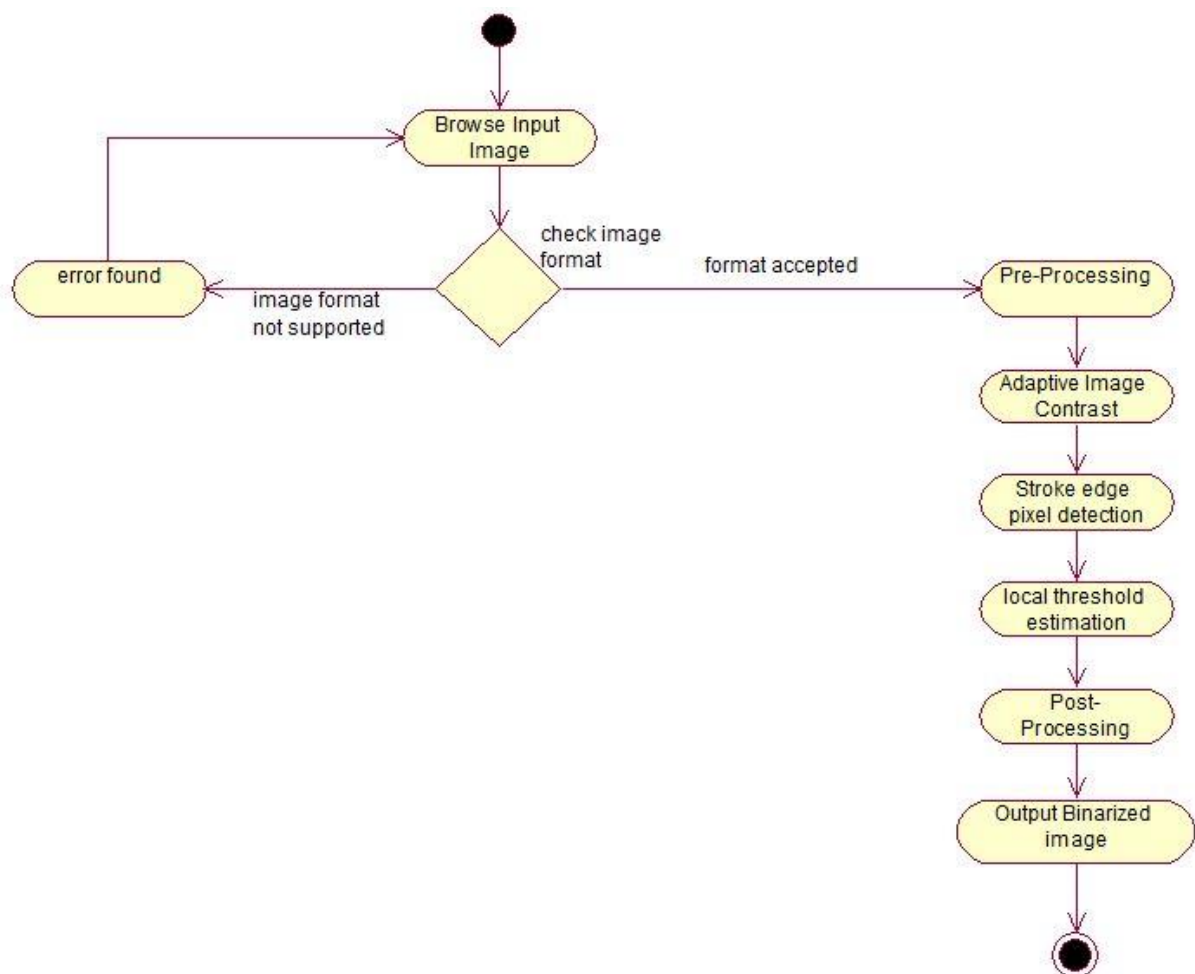


Figure. 3.2 Activity Diagram for whole system

### 3.2.3 State chart diagram

A State chart diagram describes a state machine. State machine can be defined as a machine which defines different states of an object and these states are controlled by external or internal events. State chart diagram defines the states, it is used to model the lifetime of an object. State chart diagram is used to model the dynamic nature of a system. They define different states of an object during its lifetime and these states are changed by events. State chart diagrams are useful to model the reactive systems. Reactive systems can be defined as a system that responds to external or internal events.

State chart diagram describes the flow of control from one state to another state. States are defined as a condition in which an object exists and it changes when some event is triggered. The most important purpose of State chart diagram is to model lifetime of an object from creation to termination. State chart diagrams are also used for forward and reverse engineering of a system. However, the main purpose is to model the reactive system. Following are the main purposes of using State chart diagrams:

- To model the dynamic aspect of a system.
- To model the life time of a reactive system.
- To describe different states of an object during its life time.
- Define a state machine to model the states of an object.

State chart diagram is used to describe the states of different objects in its life cycle. Emphasis is placed on the state changes upon some internal or external events. These states of objects are important to analyse and implement them accurately. State chart diagrams are very important for describing the states. States can be identified as the condition of objects when a particular event occurs.

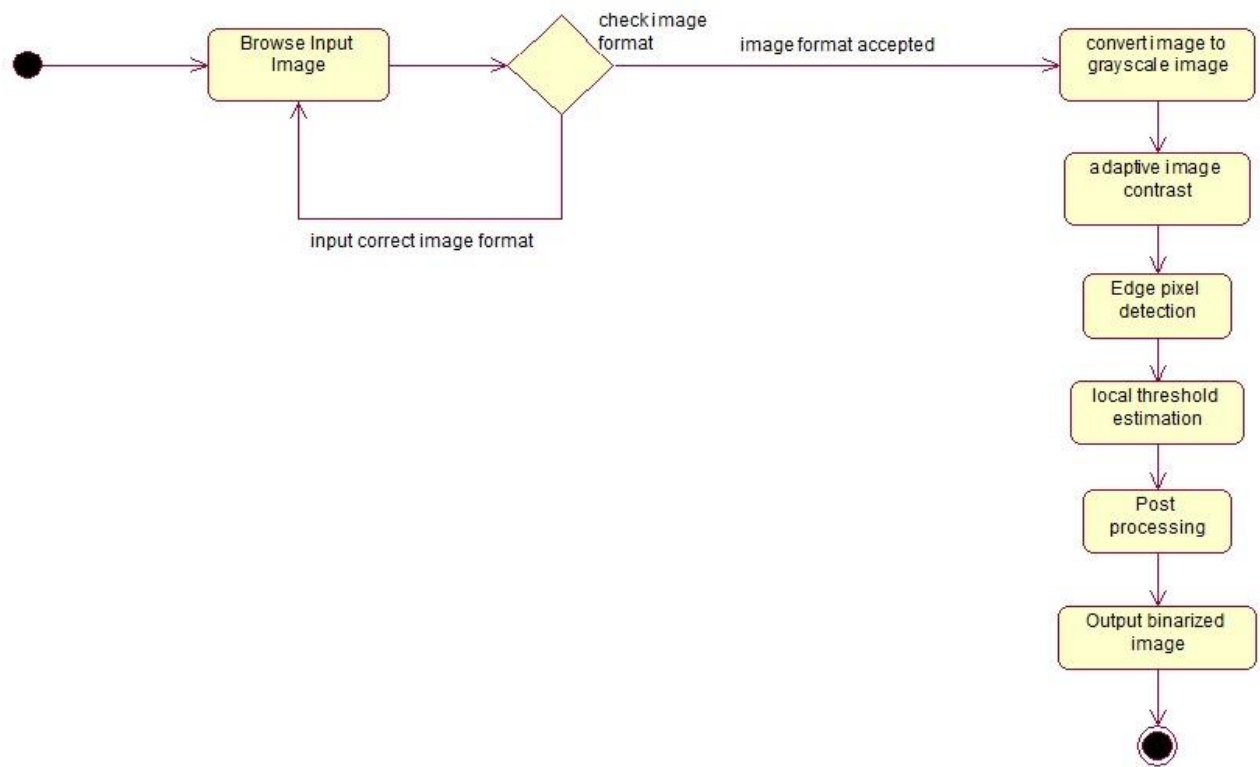


Figure. 3.3 State chart diagram for whole system

## CHAPTER 4

### SYSTEM DESIGN

#### 4.1 Architecture of the Proposed System

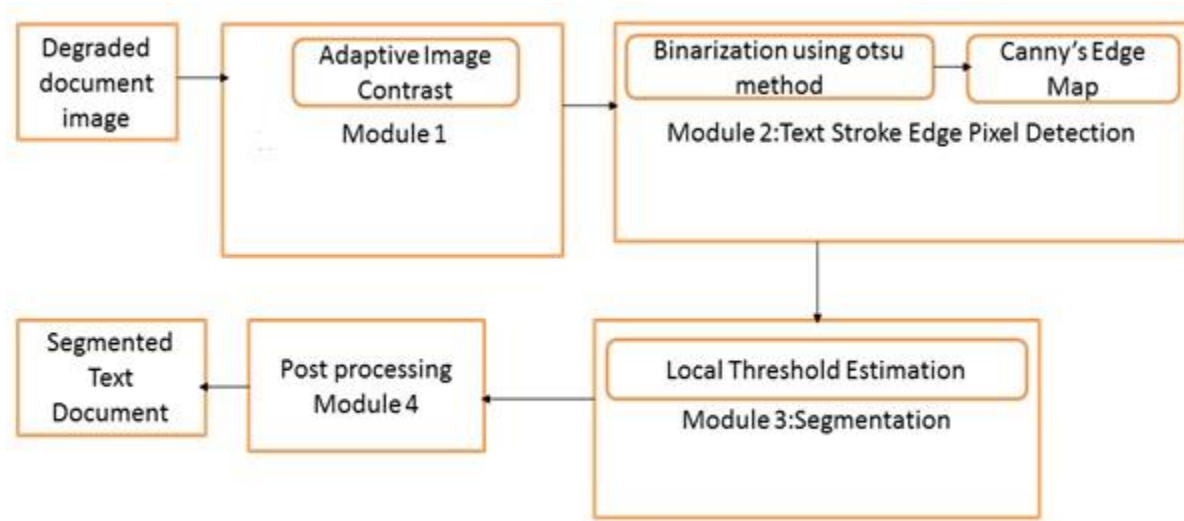


Figure. 4.1 Architecture of proposed System

Given a degraded document image, an adaptive image contrast map is first constructed and the text stroke edges are then detected through the combination of the binarized adaptive contrast map and the canny edge map. The text is then segmented based on the local threshold that is estimated from the detected text stroke edge pixels. Some post-processing is further applied to improve the document binarization quality.

#### 4.2 UML Diagrams under Design Phase

##### 4.2.1 Sequence Diagram

A sequence diagram is an interaction diagram that shows how processes operate with one another and in what order. It is a construct of a Message Sequence Chart. A sequence diagram shows object interactions arranged in time sequence. It depicts the objects and classes involved in the scenario and the sequence of messages exchanged between the objects needed to carry out the functionality of the scenario. Sequence diagrams are typically associated with use case

realizations in the Logical View of the system under development. Sequence diagrams are sometimes called event diagrams. So the purposes of the sequence diagram can be describes as:

- To capture dynamic behavior of a system.
- To describe the message flow in the system.
- To describe structural organization of the objects.
- To describe interaction among objects.

A sequence diagram shows, as parallel vertical lines (*lifelines*), different processes or objects that live simultaneously, and, as horizontal arrows, the messages exchanged between them, in the order in which they occur. This allows the specification of simple runtime scenarios in a graphical manner.

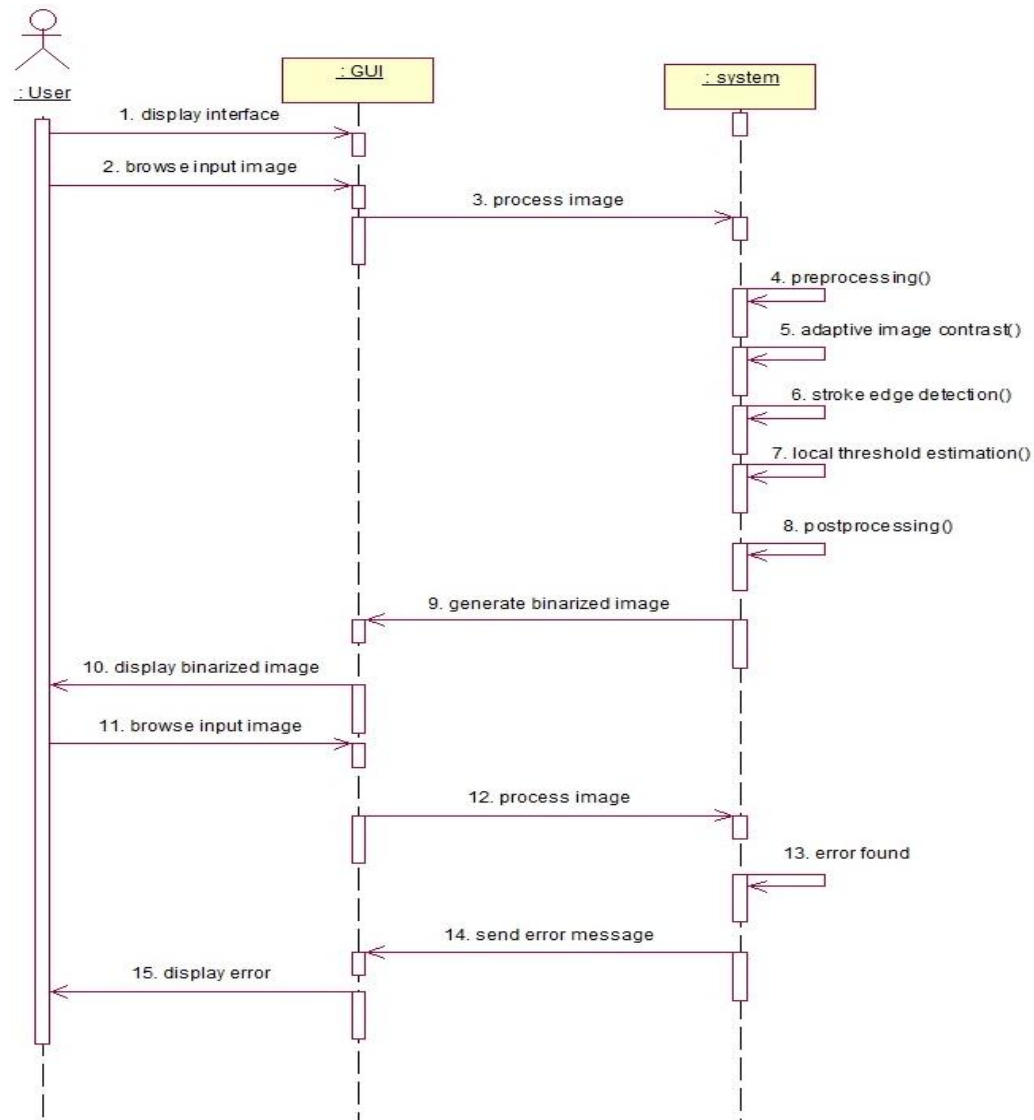


Figure. 4.2 Sequence Diagram for whole system

### 4.2.2 Collaboration diagram

The second interaction diagram is collaboration diagram. It shows the object organization as shown below. Here in collaboration diagram the method call sequence is indicated by some numbering technique as shown below. The number indicates how the methods are called one after another. So, take the same order management system to describe the collaboration diagram. The method calls are similar to that of a sequence diagram. But the



difference is that the sequence diagram does not describe the object organization whereas the collaboration diagram shows the object organization.

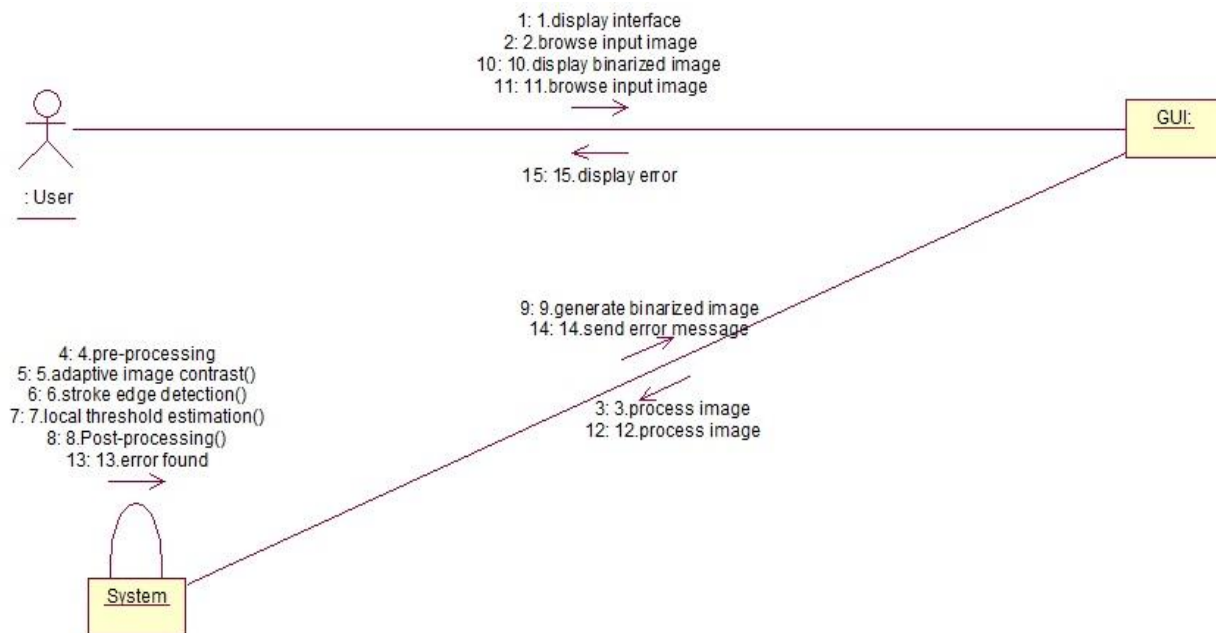


Figure. 4.3 Collaboration Diagram for whole system

### 4.3 Work flow of the Proposed System

The proposed method is an extension to previous method local maximum and minimum. The proposed method is simple, robust and capable of handling different types of degraded document images with minimum parameter tuning. It makes use of the adaptive image contrast that combines the local image contrast and the local image gradient adaptively and therefore is tolerant to the text and background variation caused by different types of document degradation. In particular, the proposed technique addresses the over-normalization problem of the local maximum minimum method. At the same time, the parameters used in the algorithm can be adaptively estimated.

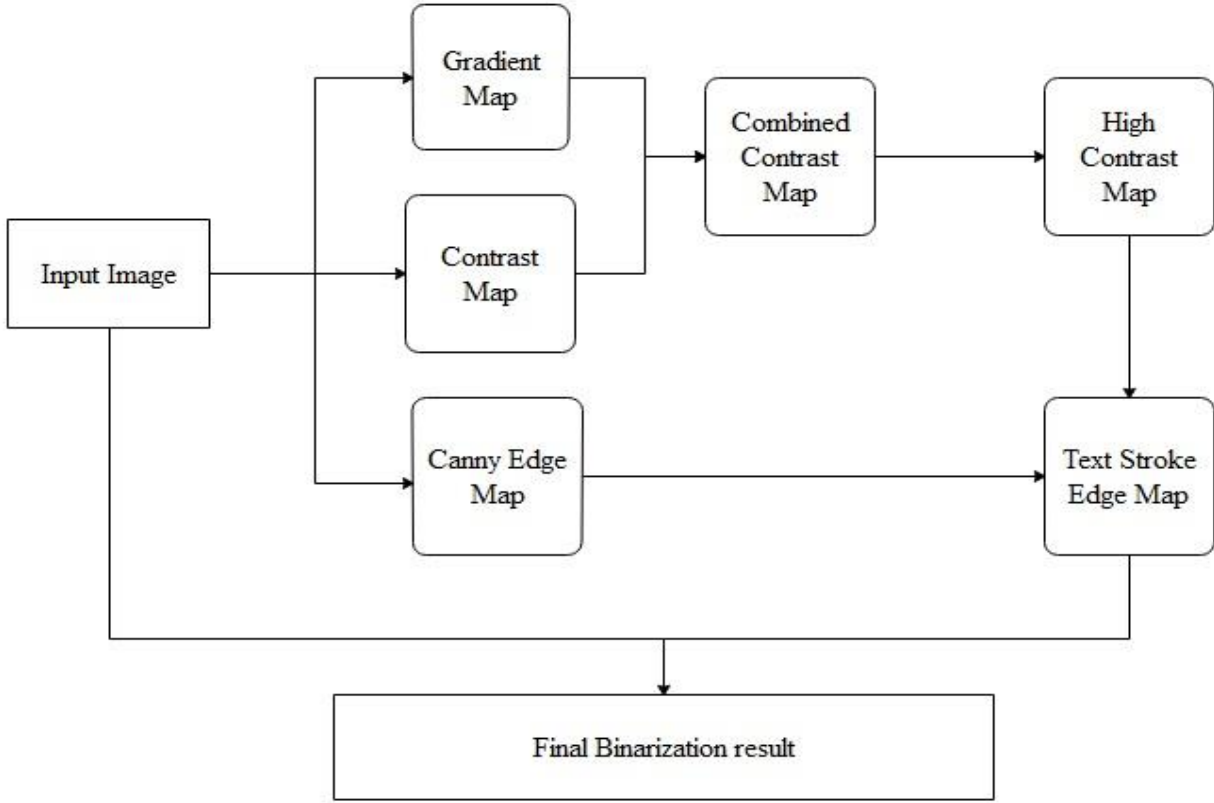


Figure. 4.4 The work flow for proposed document image binarization technique

The overall workflow is shown in Figure 4.4. Given a degraded document image, an adaptive contrast map is first constructed and the text stroke edges are then detected through the combination of the binarized adaptive contrast map and the canny edge map. The text is then segmented based on the local threshold that is estimated from the detected text stroke edge pixels. Some post-processing is further applied to improve the document binarization quality.

#### 4.4 Module Description

##### Contrast Image Construction

The image gradient has been widely used for edge detection [15] and it can be used to detect the text stroke edges of the document images effectively that have a uniform document background. On the other hand, it often detects many non-stroke edges from the background of degraded document that often contains certain image variations due to noise, uneven lighting, bleed-through, etc. To extract only the stroke edges properly, the image gradient needs to be

normalized to compensate the image variation within the document background. In the previous method [39], the local contrast evaluated by the local image maximum and minimum is used to suppress the background variation as described in Equation 4.4.1. In particular, the numerator (i.e. the difference between the local maximum and the local minimum) captures the local image difference that is similar to the traditional image gradient [15].

$$C(i, j) = \frac{I_{\max}(i, j) - I_{\min}(i, j)}{I_{\max}(i, j) + I_{\min}(i, j) + \epsilon} \quad (4.4.1)$$

The denominator is a normalization factor that suppresses the image variation within the document background. For image pixels within bright regions, it will produce a large normalization factor to neutralize the numerator and accordingly result in a relatively low image contrast. For the image pixels within dark regions, it will produce a small denominator and accordingly result in a relatively high image contrast. However, the image contrast in Equation 4.4.1 has one typical limitation that it may not handle document images with the bright text properly. This is because a weak contrast will be calculated for stroke edges of the bright text where the denominator in Equation 4.4.1 [8] will be large but the numerator will be small. To overcome this over-normalization problem, combine the local image contrast with the local image gradient and derive an adaptive local image contrast as follows:

$$C_{\alpha}(i, j) = \alpha C(i, j) + (1 - \alpha)(I_{\max}(i, j) - I_{\min}(i, j)) \quad (4.4.2)$$

where  $C(i, j)$  denotes the local contrast in Equation 4.4.1 and  $(I_{\max}(i, j) - I_{\min}(i, j))$  refers to the local image gradient that is normalized to  $[0, 1]$ . The local windows size is set to 3 empirically.  $\alpha$  is the weight between local contrast and local gradient that is controlled based on the document image statistical information. Ideally, the image contrast will be assigned with a high weight (i.e. large  $\alpha$ ) when the document image has significant intensity variation. So that the proposed binarization technique depends more on the local image contrast that can capture the intensity variation well and hence produce good results. Otherwise, the local image gradient will be assigned with a high weight. The proposed binarization technique relies more on image gradient and avoid the over normalization problem of the previous method [39].

Now, model the mapping from document image intensity variation to  $\alpha$  by a power function as follows:

$$\alpha = \left( \frac{\text{std}}{128} \right)^\gamma \quad (4.4.3)$$

where  $Std$  denotes the document image intensity standard deviation, and  $\gamma$  is a pre-defined parameter. The power function has a nice property in that it monotonically and smoothly increases from 0 to 1 and its shape can be easily controlled by different  $\gamma$ .  $\gamma$  can be selected from  $[0, \infty]$ , where the power function becomes a linear function when  $\gamma = 1$ . Therefore, the local image gradient will play the major role in Equation 4.4.2 when  $\gamma$  is large and the local image contrast will play the major role when  $\gamma$  is small.

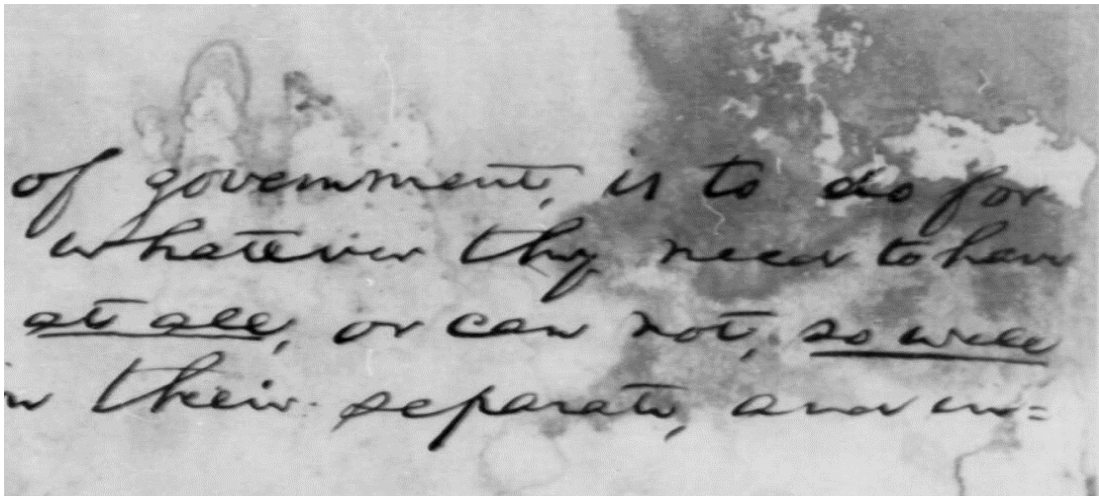


Figure. 4.5 Degraded Document Image taken from DIBCO 2009 dataset

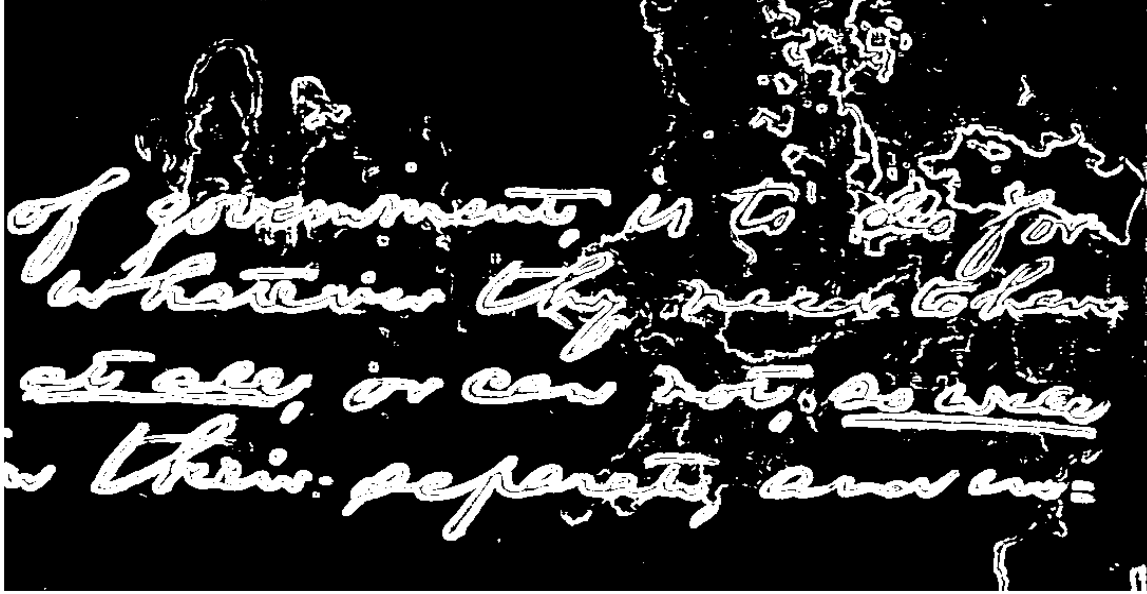


Figure. 4.6 Combination of Local image contrast and Local image gradient

Figure.4.6 shows the contrast map of the sample document images in Figure. 4.5 that are created by using the proposed method in Equation 4.4.2, respectively. For the sample document with a complex document background in Figure. 4.5, the use of the local image contrast [39] produces a better result compared with the result by the local image gradient [15] (because the normalization factors in Equation 4.4.2 helps to suppress the noise). But for the sample degraded document which has small intensity variation within the document background but large intensity variation within the text strokes, the use of the local image contrast removes many light text strokes improperly in the contrast map whereas the use of local image gradient is capable of preserving those light text strokes. As a comparison, the adaptive combination of the local image contrast and the local image gradient in Equation 4.4.2 can produce proper contrast maps for document images with different types of degradation as shown in Figure. 4.6. In particular, the local image contrast in Equation 4.4.2 gets a high weight for the document image with high intensity variation within the document background whereas the local image gradient gets a high weight for the document image in Figure. 4.5.

## Text Stroke Edge Pixel Detection

The purpose of the contrast image construction is to detect the stroke edge pixels of the document text properly. The constructed contrast image has a clear bi-modal pattern [39], where the adaptive image contrast computed at text stroke edges is obviously larger than that computed within the document background. Therefore detect the text stroke edge pixel candidate by using Otsu's global thresholding method. For the contrast images in Figure. 4.6, Figure. 4.7 shows a binary map by Otsu's algorithm that extracts the stroke edge pixels properly. As the local image contrast and the local image gradient are evaluated by the difference between the maximum and minimum intensity in a local window, the pixels at both sides of the text stroke will be selected as the high contrast pixels.

The binary map can be further improved through the combination with the edges by Canny's edge detector [4], because Canny's edge detector has a good localization property that it can mark the edges close to real edge locations in the detecting image. In addition, canny edge detector uses two adaptive thresholds and is more tolerant to different imaging artifacts such as shading [17]. It should be noted that Canny's edge detector by itself often extracts a large amount of non-stroke edges as illustrated in Figure. 4.8 without tuning the parameter manually. In the combined map, keep only pixels that appear within both the high contrast image pixel map and canny edge map. The combination helps to extract the text stroke edge pixels accurately as shown in Figure. 4.9.

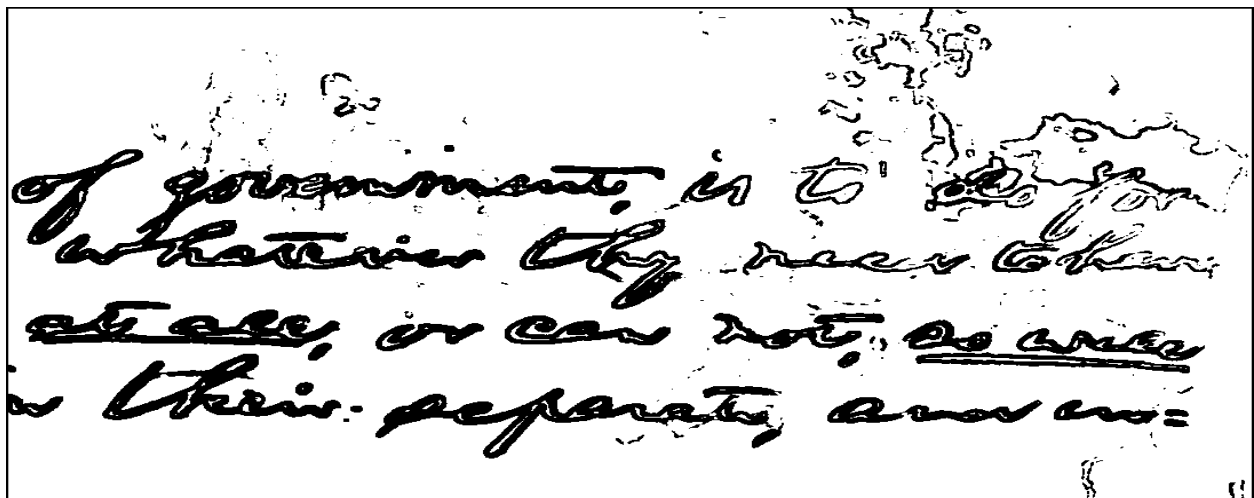


Figure. 4.7 Binary map obtained by using otsu's algorithm

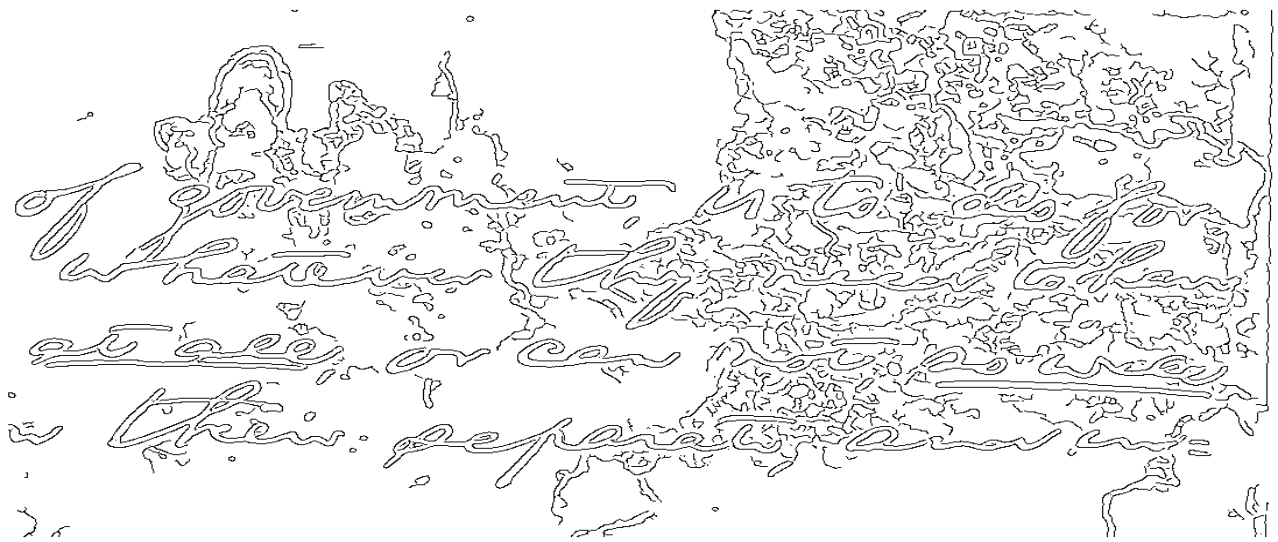


Figure. 4.8 Canny edge map of the sample document in Figure. 4.5 respectively

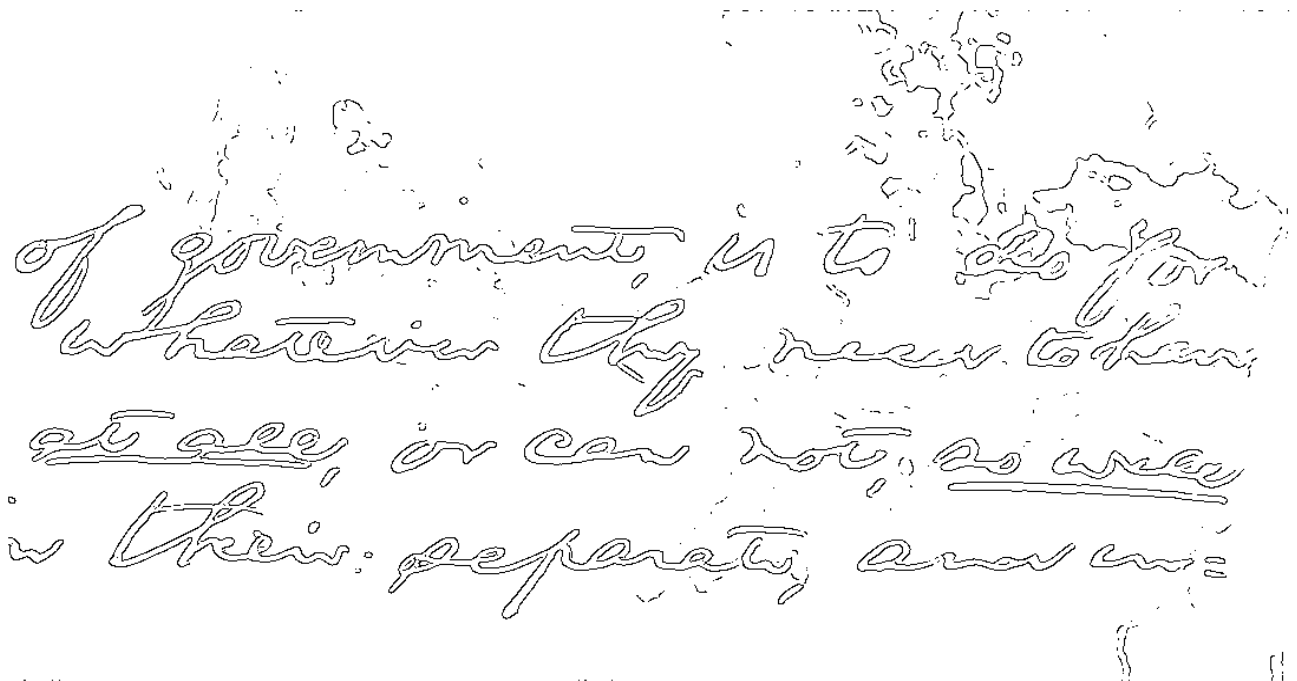


Figure. 4.9 Combined edge maps of the sample document in Figure. 4.5 respectively

### **Local Threshold estimation**

The text can then be extracted from the document background pixels once the high contrast stroke edge pixels are detected properly. Two characteristics can be observed from different kinds of document images [39]: First, the text pixels are close to the detected text stroke edge pixels. Second, there is a distinct intensity difference between the high contrast stroke edge pixels and the surrounding background pixels. After high contrast text stroke edge pixels are detected properly, segment the foreground text from the document background by a local threshold that is estimated based on the intensities of the detected text stroke edge pixels.

### **Post-Processing Procedure**

Once the initial binarization result is obtained from the previous stage then the binarization result can further be improved by incorporating certain domain knowledge. Finally, single pixel artifacts [43] are removed using several logical operators.



## CHAPTER 5

### IMPLEMENTATION

The proposed method for binarizing the degraded uses local threshold estimation to extract the text based on the detected text stroke edge pixels. The proposed method for further improving the binarization result some post processing procedure is applied through incorporating certain domain knowledge.

#### 5.1 Algorithms Used

##### Algorithm 1: Edge Width Estimation

Require: The Input Document Image  $I$  and Corresponding Binary Text Stroke Edge Image  $Edg$

Ensure: The Estimated Text Stroke Edge Width  $EW$

- 1: Get the width and height of  $I$
- 2: for Each Row  $i = 1$  to height in  $Edg$  do
- 3: Scan from left to right to find edge pixels that meet the following criteria:
  - a) its label is 0 (background);
  - b) the next pixel is labeled as 1(edge).
- 4: Examine the intensities in  $I$  of those pixels selected in Step 3. And remove those pixels that have a lower intensity than the following pixel next to it in the same row of  $I$ .
- 5: Match the remaining adjacent pixels in the same row into pairs, and calculate the distance between the two pixels in pair.
- 6: end for
- 7: Construct a histogram of those calculated distances.
- 8: Use the most frequently occurring distance as the estimated stroke edge width  $EW$ .

The document image text can thus be extracted based on the detected text stroke edge pixels as follows:

$$R(x, y) = \begin{cases} 1 & I(x, y) \leq E_{\text{mean}} + \frac{E_{\text{std}}}{2} \\ 0 & \text{otherwise} \end{cases} \quad (5.1.1)$$

where  $E_{\text{mean}}$  and  $E_{\text{std}}$  are the mean and standard deviation of the intensity of the detected text stroke edge pixels within a neighborhood window  $W$ , respectively. The neighborhood window should be at least larger than the stroke width in order to contain stroke edge pixels. So the size of the neighborhood window  $W$  can be set based on the stroke width of the document image under study,  $EW$ , which can be estimated from the detected stroke edges [shown in Figure. 4.8] as stated in Algorithm 1.

Since a precise stroke width is not needed, just calculate the most frequently distance between two adjacent edge pixels (which denotes two sides edge of a stroke) in horizontal direction and use it as the estimated stroke width. First the edge image is scanned horizontally row by row and the edge pixel candidates are selected as described in step 3. If the edge pixels, which are labelled 0 (background) and the pixels next to them are labelled to 1 (edge) in the edge map ( $Edg$ ), are correctly detected, they should have higher intensities than the following few pixels (which should be the text stroke pixels). So those improperly detected edge pixels are removed in step 4. In the remaining edge pixels in the same row, the two adjacent edge pixels are likely the two sides of a stroke, so these two adjacent edge pixels are matched to pairs and the distance between them are calculated in step 5. After that a histogram is constructed that records the frequency of the distance between two adjacent candidate pixels. The stroke edge width  $EW$  can then be approximately estimated by using the most frequently occurring distances of the adjacent edge pixels as illustrated in Figure. 5.1. The window size  $W$  is closely related to the stroke width  $EW$ . Generally, a larger local window size will help to reduce the classification error that is often induced by the lack of edge pixels within the local neighbourhood window. In addition, the performance of the proposed method becomes stable when the local window size is larger than  $2EW$  consistently on the three datasets.  $W$  can therefore be set around  $2EW$  because a larger local neighbourhood window will increase the computational load significantly.

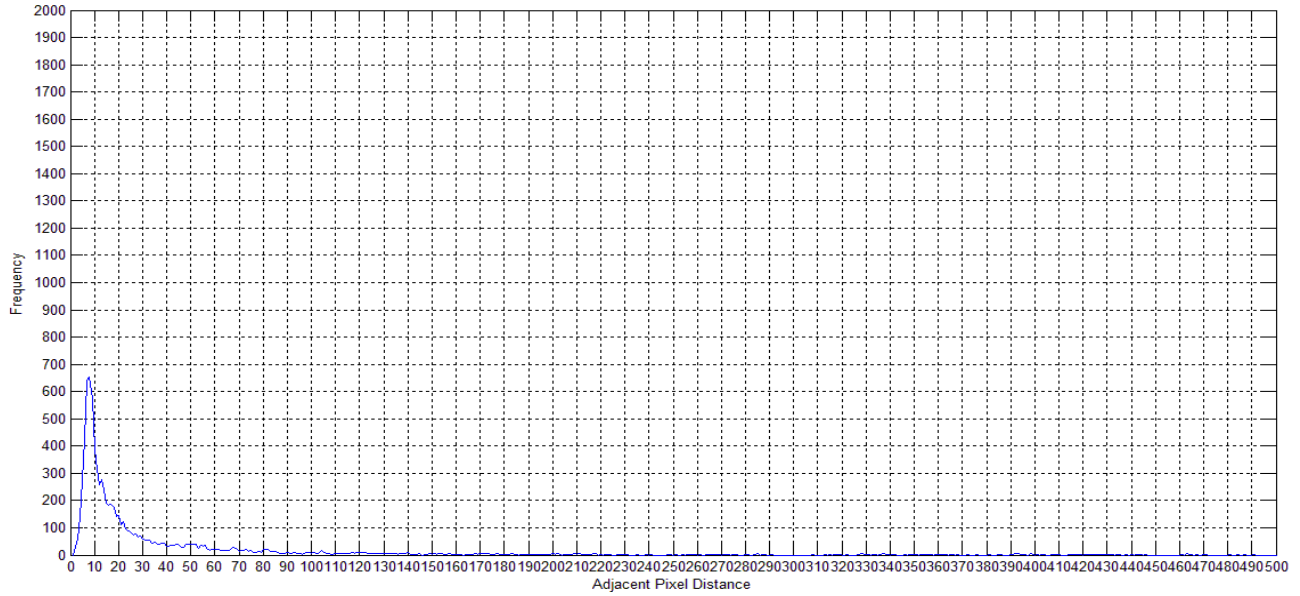


Figure. 5.1 Histogram of the distance between adjacent edge pixels of the image for Figure. 4.5

### Algorithm 2: Post-Processing Procedure

Require: The Input Document Image  $I$ , Initial Binary Result  $B$  and corresponding Binary Text Stroke Edge Image  $Edg$

Ensure: The Final Binary Result  $B_f$

- 1: Find out all the connect components of the stroke edge pixels  $Edg$ .
- 2: Remove those pixels that do not connect with other pixels.
- 3: for Each remaining edge pixels  $(i, j)$ :do
- 4: Get its neighborhood pairs:  $(i-1, j)$  and  $(i+1, j)$ ;  $(i, j-1)$  and  $(i, j+1)$
- 5: if The pixels in the same pairs belong to the same class (both text or background) then
- 6: Assign the pixel with lower intensity to foreground class (text), and the other to background class.
- 7: end if
- 8: end for
- 9: Remove single-pixel artifacts [43] along the text stroke boundaries after the document thresholding.
- 10: Store the new binary result to  $B_f$ .

Once the initial binarization result is derived as described in equation 5.1.1, the binarization result can be further improved by incorporating certain domain knowledge as described in Algorithm 2. First, the isolated foreground pixels that do not connect with other foreground pixels are filtered out to make the edge pixel set precisely. Second, the neighborhood pixel pair that lies on symmetric sides of a text stroke edge pixel should belong to different classes (i.e., either the document background or the foreground text). One pixel of the pixel pair is therefore labeled to the other category if both of the two pixels belong to the same class. Finally, some single-pixel artifacts [43] along the text stroke boundaries are filtered out by using several logical operators.

## **5.2 Description of Data Sets**

### **Document Image Binarization Contest (2009)**

It is the first international document image binarization contest, the general objective is to record recent advances in document image binarization using established evaluation performance measures. Here the plan is to create a benchmarking dataset that is representative of the potential problems which are challenging in the binarization process and use a common evaluation platform in order to test and compare the submitted algorithms for document image binarization. The DIBCO 2009 dataset will contain images that range from gray scale to color, from machine printed to handwritten, and finally, from real to synthetic. Sample image documents along with the corresponding ground truth will be given. The DIBCO 2009 [35] is held under the framework of the International Conference on Document Analysis and Recognition (ICDAR) 2009 [35].

### **Handwritten Document Image Binarization Contest (2010)**

Document image binarization is an important step in the document image analysis and recognition pipeline. Therefore, it is imperative to have a benchmarking dataset along with an objective evaluation methodology in order to capture the efficiency of current document image binarization practices. Following the success of DIBCO 2009 organized in conjunction with ICDAR'09, the follow-up of this contest is organized in conjunction with ICFHR 2010. In H-DIBCO 2010 (**H**andwritten **D**ocument **I**mage **B**inarization **C**ontest) [42], the general objective is to record recent advances in handwritten document image binarization using established

evaluation performance measures. It has a benchmarking dataset that is representative of the potential problems which are challenging in the binarization process and use a common evaluation platform in order to test and compare the submitted algorithms for handwritten document image binarization.

### **Document Image Binarization Contest (2011)**

Document image binarization is an important step in the document image analysis and recognition pipeline. Therefore, it is imperative to have a benchmarking dataset along with an objective evaluation methodology in order to capture the efficiency of current document image binarization practices. Following the success of DIBCO 2009 organized in conjunction with ICDAR'09 as well as of H-DIBCO 2010 organized in conjunction with ICFHR 2010, the follow-up of these contests in the framework of ICDAR 2011 [45] is organized. The general objective of DIBCO 2011 is to record recent advances in document image binarization using established evaluation performance measures and a benchmarking dataset that is representative of the potential challenges met in the binarization process. The DIBCO 2011 dataset will contain images that range from gray scale to color and from machine printed to handwritten. For the preparation of the users (participants) using the new dataset, the existing publicly available datasets of previous contests (DIBCO 2009 and H-DIBCO 2010) along with the corresponding ground truth could be used.

### **Bickley Diary Dataset**

This dataset is the dairy of Ms. Anna Felton Bickley, the wife of Bishop George H. Bickley, one of the earliest Methodist missionaries to come to Malaysia. The dairy is designed such that each page is the same calendar day but for up to five different years. This dataset consists of 92 grayscale images of a photocopy of the original dairy. The images are from a single volume (Jan-Mar, 1922-1926) of a four volume set. This dataset is particularly challenging as the original diary suffers from discolorization and water stains, differences in ink contrast for the different years, and overall noise from the photocopying. Also, Ms. Bickley's handwriting is challenging to read. It has manually binarized 7 images to serve as ground truth for the example-based binarization method. Mr. Erin Bickley, Jr (grandson of Bishop Bickley)

has made the dataset is public for other researchers to use. The images from Bickley diary dataset [40] are taken from a photocopy of a diary that is written about 100 years ago. These images suffer from different kinds of degradation, such as water stains, ink bleed-through, and significant foreground text intensity and are more challenging than the previous two DIBCO and H-DIBCO datasets. Seven ground truth images that are annotated manually using Pix Labeler [36] to evaluate the proposed method with the other methods.

### 5.3 Metrics Calculated

The binarization performance is evaluated by using F-Measure, pseudo F-Measure, Peak Signal to Noise Ratio (PSNR), Negative Rate Metric (NRM), Misclassification Penalty Metric (MPM) and Distance Reciprocal Distortion (DRD) that are adopted from DIBCO 2009, H-DIBCO 2010 and DIBCO 2011 [35, 42, 45]. Due to the lack of ground truth data in some datasets (such as the skeleton ground truth), not all of the metrics are applied on every image. In particular, the F-Measure is defined as follows:

$$FM = \frac{2*RC*PR}{RC+PR} \quad (5.3.1)$$

$$RC = \frac{CTP}{CTP+CFN} \quad (5.3.2)$$

$$PR = \frac{CTP}{CTP+CFP} \quad (5.3.3)$$

where the terms  $RC$  and  $PR$  refer the recall and the precision of the method in Equation 5.3.1. The terms  $CTP$ ,  $CFP$ , and  $CFN$  denote the numbers of true positive pixels, false positive pixels, and false negative pixels, respectively. This measure evaluates how well an algorithm can retrieve the desire pixels. The pseudo F-Measure is defined as follows:

$$pFM = \frac{2*pRC*PR}{pRC+PR} \quad (5.3.4)$$

$$pRC = \frac{\sum_{i,j} SG(i,j)B(i,j)}{\sum_{i,j} SG(i,j)} \quad (5.3.5)$$

where the term  $pRC$  refers the pseudo recall of the method in Equation 5.3.4, the term  $PR$  is the same as in Equation 5.3.1. The term  $SG$  denotes the skeletonized ground truth image that has 0 at background and 1 in text, respectively. The term  $B$  denotes the resultant binary image. This measure evaluates how well an algorithm can preserve the character skeleton. The measure PSNR is defined as follows:

$$PSNR = 10\log\left(\frac{C^2}{MSE}\right) \quad (5.3.5)$$

$$MSE = \frac{\sum_{x=1}^M \sum_{y=1}^N (I(x,y) - I'(x,y))^2}{MN} \quad (5.3.6)$$

where the term  $C$  is a constant that denotes the difference between foreground and background. This constant can be set to 1. The term  $PSNR$  measures how close the resultant image to the ground truth image. The measure NRM is defined as follows:

$$NRM = \frac{NR_{FN} + NR_{FP}}{2} \quad (5.3.7)$$

$$NR_{FN} = \frac{N_{FN}}{N_{FN} + N_{TP}}$$

$$NR_{FP} = \frac{N_{FP}}{N_{FP} + N_{TN}}$$

where the terms  $N_{TP}$ ,  $N_{FP}$ ,  $N_{TN}$ , and  $N_{FN}$  denote the number of true positives, false positives, true negatives, and false negatives respectively. This metric measures pixel mismatch rate between the ground truth image and resultant image. The measure MPM is defined as follows:

$$MPM = \frac{MP_{FN} + MP_{FP}}{2} \quad (5.3.8)$$

$$MP_{FN} = \frac{\sum_{i=1}^{N_{FN}} d_{FN}^i}{D}$$

$$MP_{FP} = \frac{\sum_{j=1}^{N_{FP}} d_{FP}^j}{D}$$

where the terms  $d_{FN}^i$  and  $d_{FP}^j$  denote the distance of the  $i^{th}$  false negative and the  $j^{th}$  false positive pixel from the contour of the ground truth segmentation. The normalization factor  $D$  is the sum over all the pixel-to-contour distances of the ground truth object. This metric measures how well the resultant image represents the contour of ground truth image. The measure DRD as follows:

$$DRD = \frac{\sum_{K=1}^S DRD_K}{N} \quad (5.3.9)$$

$$DRD_K = \sum_{i=-2}^2 \sum_{j=-2}^2 |GT_K(x+i, y+j) - B_K(x, y)| \times W_{Nm}(i, j)$$

where the term  $DRD_k$  denotes the distortion of the  $k$  - th flipped pixel and calculated using a weight matrix  $W_{Nm}$  that defined in [34]. The pair  $(x, y)$  denotes the index of the  $k$  - th flipped pixel, The terms  $B$  and  $GT$  denote the resultant binary image and ground truth image, respectively. The term  $N$  refers to the number of the non-uniform (which contains background and text pixels)  $8 \times 8$  blocks in the GT image. This metric properly correlates with the human visual perception and measures the distortion for all the flipped pixels.

## 5.4 Methods Compared

In the experiment, the proposed method is quantitatively compared with other state-of-the-art techniques on DIBCO 2009, H-DIBCO 2010 and DIBCO 2011 datasets. These methods include Otsu's method (OTSU) [1], Sauvola's method (SAUV) [18], Niblack's method (NIBL) [5], Bernsen's method (BERN) [3], Gatos et al.'s method (GATO) [26] and the previous methods LMM [39], BE [43]. Besides the comparison methods mentioned above, the proposed method is also compared with the top algorithms namely Lelore et al.'s method (LELO) [46], the method submitted by a team (SNUS) for the DIBCO 2011 dataset. The three datasets are composed of the same series of document images that suffer from several common document degradations such as smear, smudge, bleed-through and low contrast. The DIBCO 2009 dataset contains ten testing images that consist of five degraded handwritten documents and five degraded printed documents. The H-DIBCO 2010 dataset consists of ten degraded handwritten documents. The DIBCO 2011 dataset contains eight degraded handwritten documents and eight degraded printed documents. In total, 36 degraded document images with ground truth are available.



## CHAPTER 6

### TESTING

A healthy attitude towards testing is an aggregate of the following perspectives listed in increasing order of maturity. Each attitude is translated into a hard objective you can include in your test planning documentation. Note that the most advanced and effective attitude towards testing is summed up at maturity level 4 “Testing is a state of mind”.

#### 6.1 Datasets Tested

The algorithm is tested publicly available DIBCO(2009),DIBCO(2011),H-DIBCO(2010) datasets and also the algorithm is tested on more challenging data set called Bickley Diary Dataset. The proposed algorithm is compared with the state of art techniques namely Otsu’s method [1], Sauvola’s method[18], Niblack’s method[5], Bernsen’s method[3], Gato’s method[26].

#### 6.2 Test Cases

Characteristics of a good test case are -

Accurate: Exacts the purpose

Economical: No unnecessary steps or word

Traceable: Capable of being traced to requirements

Repeatable: Can be used to perform the test over and over

Reusable: Can be reused if necessary.

Test Case Summary: - To convert a (color/grayscale) degraded document image into a binarized document image by using adaptive image contrast.

Related Requirement: - Mat Lab

Test Case Procedure: - Given a degraded document image, an adaptive contrast map is first constructed and the text stroke edges are then detected through the combination of the binarized adaptive contrast map and the canny edge map. The text is then segmented based on the local threshold that is estimated from the detected text stroke edge pixels. Some post-processing is

further applied to improve the document binarization quality. So, different types of testing on the project is performed. They are-

## **System Testing**

System testing is the stage of implementation, which aimed at ensuring that the system works accurately and efficiently before the live operation commences. Testing is the process of executing a program with the intent of finding an error. A good test case is one that has a high probability of finding a yet undiscovered error. A successful test is one that answers a yet undiscovered error.

Testing is vital to the success of the system. System testing makes a logical assumption that if all parts of the system are correct, the goal will be successfully achieved. The candidate system is subject to variety of tests-on-line response, Volume Street, recovery and security and usability test. A series of tests are performed before the system is ready for the user acceptance testing. Any engineered product can be tested in one of the following ways. Knowing the specified function that a product has been designed to from, test can be conducted to demonstrate each function is fully operational. Knowing the internal working of a product, tests can be conducted to ensure that “al gears mesh”, that is the internal operation of the product performs according to the specification and all internal components have been adequately exercised.

## **Unit Testing**

Unit testing is the testing of each module and the integration of the overall system is done. Unit testing becomes verification efforts on the smallest unit of software design in the module. This is also known as ‘module testing’. The modules of the system are tested separately. This testing is carried out during the programming itself. In this testing step, each model is found to be working satisfactorily as regard to the expected output from the module. There are some validation checks for the fields. For example, the validation check is done for verifying the data given by the user where both format and validity of the data entered is included. It is very easy to find error and debug the system.

Here, unit testing is performed on each and every module namely adaptive image contrast module, text stroke edge pixel detection module, local threshold estimation module, post-processing procedure module. So, each and every module is tested separately and scrutinized each and every operation.

### **Integration Testing**

Data can be lost across an interface, one module can have an adverse effect on the other sub function, when combined, may not produce the desired major function. Integrated testing is systematic testing that can be done with sample data. The need for the integrated test is to find the overall system performance. Is the phase in software testing in which individual software modules are combined and tested as a group.

Here after unit testing, integration testing is performed where the modules are integrated into one module such that each module's output is given as input to the next module. So, all the 4 (adaptive image contrast, stroke edge pixel detection, local threshold estimation, post-processing) modules are grouped into one module and integration testing is performed.

### **Performance testing**

To determine the responsiveness, throughput, reliability, and scalability of a system under a given workload. When the proposed algorithm is tested on the DIBCO datasets the results are fast than the results of Bickley diary dataset.

### **Validation Testing**

Process of checking that system meets specifications and that it fulfills its intended purpose.

Expected Result: - Binarized Document Image

Actual Result: - Binarized Document Image

Table 6.1 Test Case 1

Test case Name	Test case description	Test steps			Test status (P/F)
		Steps	expected	actual	
Upload image	Uploading an invalid image format	Upload an image	Has to show error message	error message	P

Table 6.2 Test Case 2

Test case Name	Test case description	Test steps			Test status (P/F)
		Steps	expected	actual	
Upload valid document image	Compute adaptive local image contrast	Combine local image contrast and local image gradient	No error and output of contrast image	Contrast image	P

Table 6.3 Test Case 3

Test case Name	Test case description	Test steps			Test status (P/F)
		Steps	expected	actual	
Input contrast image and input valid document image	Detect the text stroke edge pixel candidates	Combine binary contrast map and canny edge map	Combined edge map with the detected text stroke edge pixels	Combined edge map	P

Table 6.4 Test Case 4

Test case Name	Test case description	Test steps			Test status (P/F)
		Steps	expected	actual	
Input Document Image and Binary text stroke edge image	To calculate the most frequently occurring distance as the estimated stroke edge width	Find edge pixels and examine the intensities in document image and remove pixels that have lower intensity than the following pixel. Finally calculate the adjacent edge pixel distance and find estimated stroke edge width	Histogram of calculated distances and estimated text stroke edge width without errors	The estimated Text stroke edge width	P

Table 6.5 Test Case 5

Test case Name	Test case description	Test steps			Test status (P/F)
		Steps	expected	actual	
Input Document image and initial binary result and binary text stroke edge image	Post processing procedure to improve the binarization result by incorporating domain knowledge, to obtain final binary result	Remove those pixels that don't connect with other pixels. The neighborhood pixel pair of a text stroke edge pixel should belong to different classes. Finally remove single pixel artifacts.	Obtain the new binary result from post processing procedure	Obtain the new binary result from post processing procedure	P

## CHAPTER 7

### RESULTS

In the section of results the resultant binarized document images of different degraded document input images are shown and the proposed algorithm is compared with the state of art techniques over on well-known competition datasets: DIBCO 2009, H-DIBCO 2010, and DIBCO 2011. The proposed algorithm is further tested on a very challenging Bickley Diary dataset.

#### 7.1 Actual Results of the work

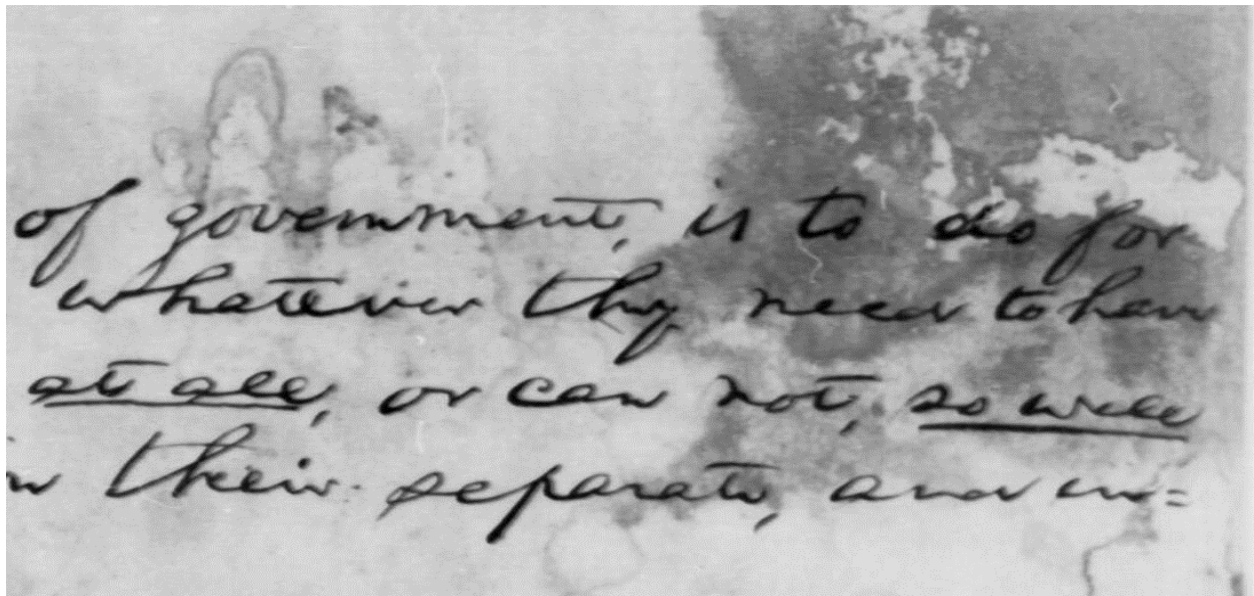


Figure. 7.1 Degraded Document Image taken from DIBCO 2009 handwritten dataset

of government, is to do for  
 whatever they need to have  
at all, or can not, so well  
 in their separate, manner:

Figure. 7.2 Binarization result of the document image for Figure. 7.1 taken from DIBCO  
 handwritten 2009 dataset

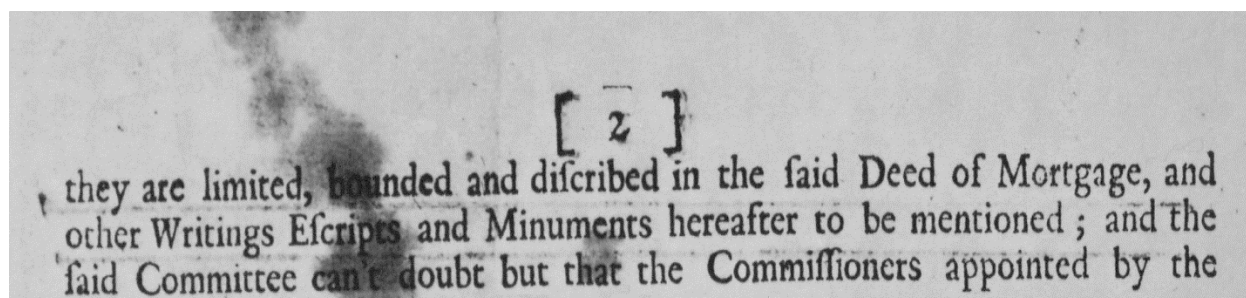


Figure. 7.3 Degraded Document image taken from DIBCO 2009 machine printed dataset

[ 2 ]  
 they are limited, bounded and discribed in the said Deed of Mortgage, and  
 other Writings Escripts and Minuments hereafter to be mentioned ; and the  
 said Committee can't doubt but that the Commissioners appointed by the

Figure. 7.4 Binarized Document Image for Figure. 7.3 taken from DIBCO 2009 machine printed dataset



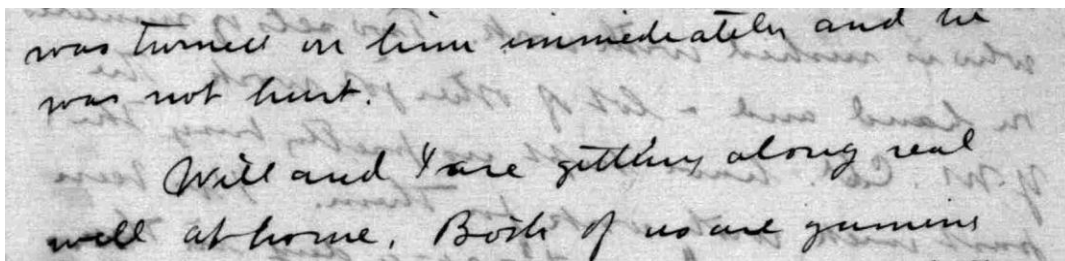


Figure. 7.5 Degraded Document Image taken from H-DIBCO 2010 dataset

was turned in time immediately and he was not hurt.

Will and I are getting along real well at home. Both of us are getting

Figure. 7.6 Binarized Document Image for Figure. 7.5 taken from H-DIBCO 2010 dataset

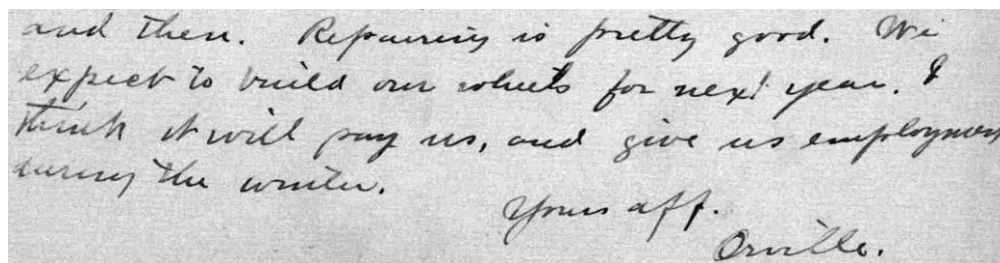


Figure. 7.7 Degraded Document Image taken from H-DIBCO 2010 dataset

and then. Repairing is pretty good. We expect to build our wharfs for next year. I think it will pay us, and give us employment during the winter.

Yours aff. Orville.

Figure. 7.8 Binarized Document Image for Figure. 7.7 taken from H-DIBCO 2010 dataset

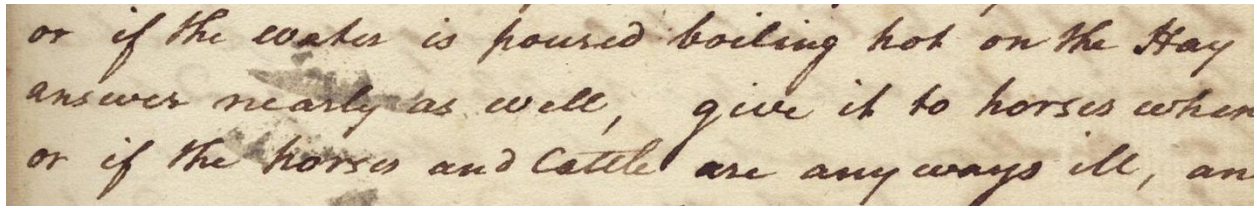


Figure. 7.9 Degraded Document image taken from DIBCO handwritten 2011 dataset

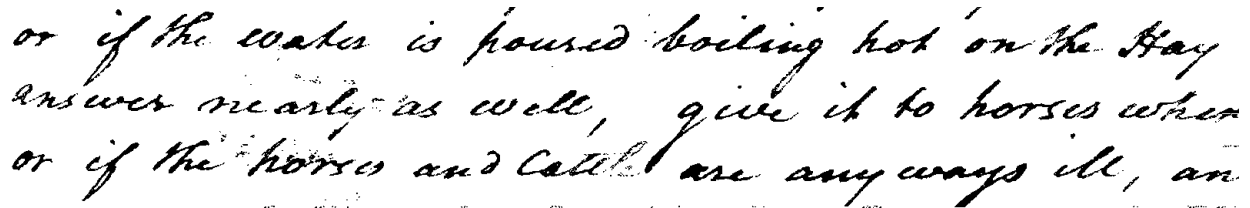


Figure. 7.10 Binarized Document Image for Figure. 7.9 taken from DIBCO hand written 2011 dataset

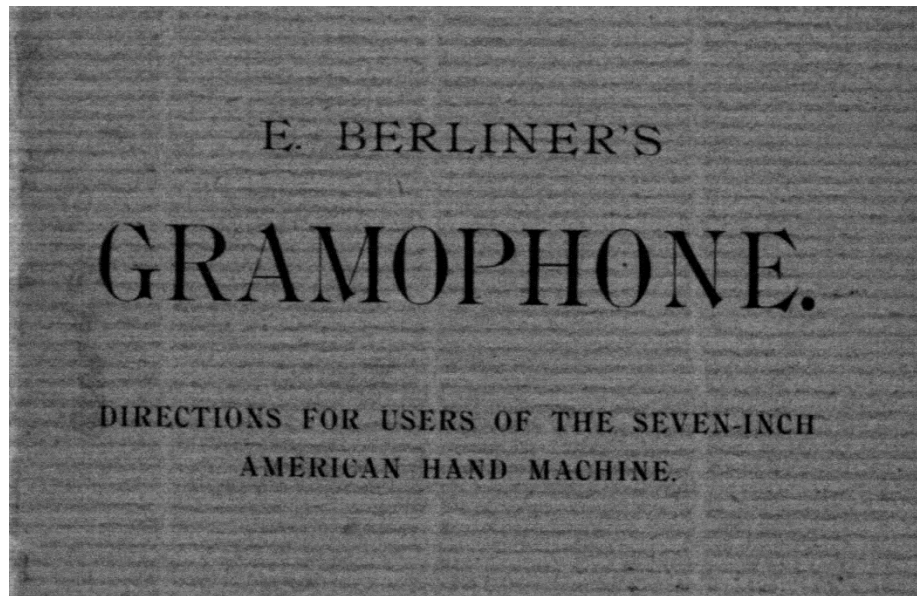


Figure. 7.11 Degraded Document Image taken from DIBCO 2011 machine printed dataset

E. BERLINER'S  
**GRAMOPHONE.**

**DIRECTIONS FOR USERS OF THE SEVEN-INCH  
AMERICAN HAND MACHINE.**

Figure. 7.12 Binarized Document Image for Figure. 7.11 taken from  
DIBCO 2011 machine printed dataset

Lured my <sup>1922</sup> guests were here 15 days. Left at 12 noon to go on "Lily of Calcutta" for Hong Kong. Left \$5 each for servant \$20 in all. Gave them basket of pineapples, oranges, bananas & candy. Linen to the 3 ladies. Hard thunderstorm last night sultry & rainy today. The Calcutta of German Line was a miserable dirty boat. Dr. Hargreaves preached 5:30 & dined with us, leaving to preach again at 8 P.M.

MARCH 26

Monday - Had done & Mrs. Hollington called at 11 o'clock after seeing the schools with Pa. They went <sup>1923</sup> to Raffles to dinner & at 3 o'clock several of us went to boat to fare them off. Only had time to see Had done cabin or suite of 4 rooms - luxurious \$15,000 for the trip. Mr. Peabody sat on the rail & did antics as long as he could see him. Very tame departure compared to that of Laconia - no band music today - very few Americans at wharf.

Wednesday - Mrs. Maynard came & took in shopping. <sup>1924</sup> While I packed. Went to train at 11:30 a.m. at Naples at 5 P.M. & went to linen store, where we got a few pcs. Had room with bath, so were very comfortable & retired early.

Thursday - Fine cool day at St. Petersburg. Washed my hair in the morning & spent day on the porch talking, sewing & reading & watching the crowds go by this busy corner. <sup>1925</sup>

Friday. Rained very hard early this morning & is showery yet. Wrote letters & read a newsy one from motel here. <sup>1926</sup> Mezzie, Betty & I went to 7 P.M. show, to see "The Keeper of the Keys" - we had just read it - it was interesting, even tho' it differed in many details. More people than at any time since we came.

Figure. 7.13 Badly Degraded Document Image taken from Bickley Diary dataset

Lunch any <sup>1923</sup> ~~lunch~~ <sup>lunch</sup> were for 15 days. Left at 12 noon to  
 go on ship <sup>1923</sup> ~~Calcutta~~ <sup>Calcutta</sup> for Hong Kong. Left \$5 each for servant  
 \$20 in all. Gave them each a bundle of oranges, bananas & candy  
 & pa. Linen to the 3 ladies. Hard thunderstorm last night sultry &  
 rainy today. The Calcutta of Ellerman Line was a miserable  
 dirty boat. Dr. Hargreaves preached 5:30 & dined with us, leaving  
 to preach again at 8 P.M.

Monday - Had Mrs. & Mrs. Hargreaves called  
 at 11 o'clock after seeing the school with Pa. They went  
 to Raffles Ho. Sigin & at 3 o'clock several of us  
 went to boat to see them off. Only had time to see  
 Haddons cabin or suite of rooms. - Immense \$15, or for  
 the trip. Mr. Peabody sat on the rail & did antics as long  
 as he could all time. Very tame departure compared  
 to that of Laccina - no band music today - very few  
 Americans & 1 other.

Wednesday - Mrs. Maynard came & took in shopping.  
 while I packed. Went to train at 11:30 arr. at hotel  
 at 5 P.M. & went to linen store, where we got a few pcs.  
 Had room with bath, so were very comfortable & re-  
 tired early.

Thursday - Fine cool day at St. Petersburg. Washed  
 my hair in the morning & spent day on the  
 porch talking, sewing & reading & watching  
 the citinas so in this busy corner.

Friday - Rained very hard early this morning & it cleared  
 up. Little & cold & very one after another.  
 1923 <sup>1923</sup> ~~Friday~~ <sup>Friday</sup> ~~at 7:30~~ <sup>at 7:30</sup> ~~to see~~ <sup>to see</sup>  
 the folks "as it had just had it" - it was  
 raining, but the air is agreeable in many details.  
 more people than at any time since we

Figure. 7.14 Binarized Document Image for badly degraded document image in Figure. 7.13  
 taken from Bickley Diary dataset

The above results are the binarized document images for badly degraded document images taken from a series of datasets as DIBCO 2009, H-DIBCO 2010, DIBCO 2011, Bickley diary. These binarized document images are obtained using the adaptive image contrast method. The adaptive image contrast which is tolerant to different types of document degradations. Two images are taken from each dataset where one image is taken from machine printed dataset and other image from handwritten dataset for the datasets DIBCO 2009, DIBCO 2011. And one image is taken from Bickley diary dataset and two images are taken from H-DIBCO 2010 handwritten dataset. So, the binarization results are shown for those images as mentioned which are obtained by using the proposed adaptive image contrast method. Therefore, the proposed method produces good binarization results for the degraded document images that suffer from several common document degradations such as smear, smudge, bleed-through and low contrast.

## **7.2 Analysis of the results obtained**

In the result analysis section different evaluation metrics are used to evaluate the proposed algorithm with the state of art techniques and the analysis is shown below. The proposed algorithm achieves highest scores in F-measure and PSNR values compared with the state of art techniques. This means that the proposed method produces a higher overall precision and preserves the text strokes better. On the other hand, the proposed method produces a binary result with better visual quality and contains most of the text information. In addition, the computation time of the proposed method and other state of the art techniques implemented in matlab are tested. The average execution time of the proposed method over DIBCO's test dataset is around 148.6 seconds. The execution time of OTSU [1], BERN [3], NIBL [5], SAUV [18], GATO [26], BE [43] and LMM [39] methods are around 0.5 seconds, 18 seconds, 27 seconds, 28 seconds, 100 seconds, 24 seconds and 20 seconds respectively. The proposed technique is comparable to the state of art adaptive document thresholding methods.

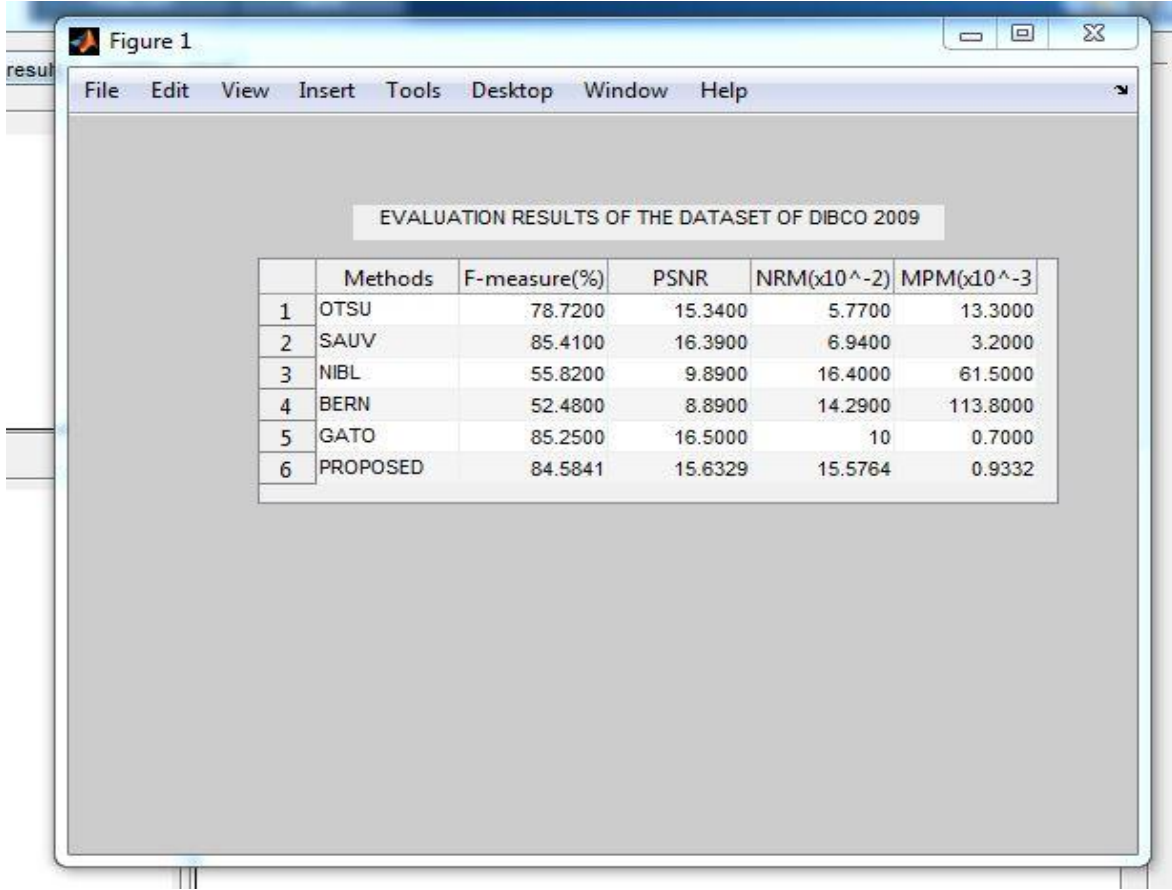
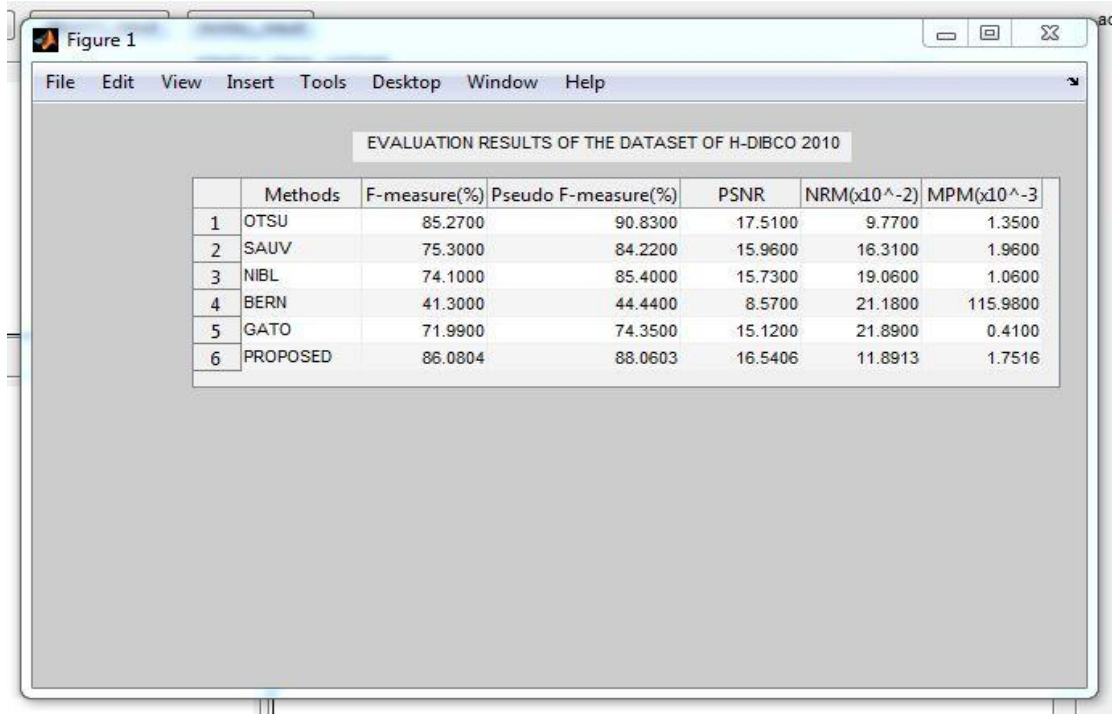


Figure. 7.15 Testing on competition dataset DIBCO 2009

The above figure. 7.15 shows the evaluation results for the competition dataset DIBCO 2009. The evaluation metrics that are used to evaluate the DIBCO 2009 dataset are F-measure, PSNR (Peak Signal to Noise Ratio), NRM (Negative Rate Metric), Misclassification Penalty Metric (MPM). The proposed method is quantitatively compared with other state of the art techniques on DIBCO 2009 dataset. These methods include OTSU's method (OTSU) [1], Sauvola's method (SAUV) [18], Niblack's method (NIBL) [5], Bernsen's method (BERN) [3], GATO's method (GATO) [26] are compared with the proposed method. The DIBCO 2009 dataset contains ten testing images that consist of five degraded handwritten documents and five degraded printed documents. The proposed achieves highest score in F-measure and PSNR values. This means that the proposed method produces a higher overall precision and preserves the text strokes better. Therefore, the proposed method extracts the text strokes better than the other comparison methods.



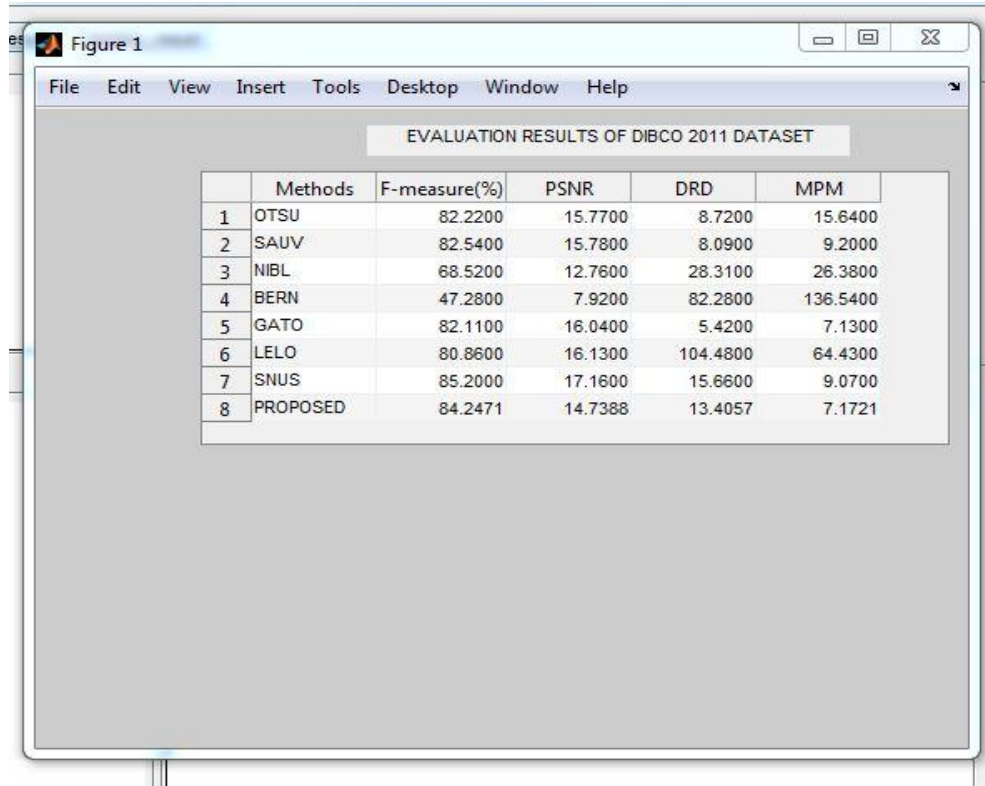


	Methods	F-measure(%)	Pseudo F-measure(%)	PSNR	NRM( $\times 10^{-2}$ )	MPM( $\times 10^{-3}$ )
1	OTSU	85.2700	90.8300	17.5100	9.7700	1.3500
2	SAUV	75.3000	84.2200	15.9600	16.3100	1.9600
3	NIBL	74.1000	85.4000	15.7300	19.0600	1.0600
4	BERN	41.3000	44.4400	8.5700	21.1800	115.9800
5	GATO	71.9900	74.3500	15.1200	21.8900	0.4100
6	PROPOSED	86.0804	88.0603	16.5406	11.8913	1.7516

Figure. 7.16 Testing on competition dataset H-DIBCO 2010

The above figure 7.16 shows the evaluation results for the competition dataset H-DIBCO 2010. The evaluation metrics that are used to evaluate the H-DIBCO 2010 dataset are F-measure, pseudo F-measure, PSNR (Peak Signal to Noise Ratio), NRM (Negative Rate Metric) and MPM (Misclassification Penalty Metric). The proposed method is compared with the state of art techniques on H-DIBCO 2010 dataset. These methods are OTSU's method (OTSU) [1], Sauvola's method (SAUV) [18], Niblack's method (NIBL) [5], Bernsen's method (BERN) [3], GATO's method (GATO) [26] are compared with the proposed method. The H-DIBCO 2010 dataset consists of ten degraded handwritten documents. The proposed method achieves highest score in F-measure, pseudo F-measure and the PSNR values. This means that the proposed method produces a higher overall precision and preserves the text stroke better for the H-DIBCO 2010 dataset. The proposed method extracts the text strokes better than the other comparison methods for the test images in H-DIBCO 2010 dataset.





	Methods	F-measure(%)	PSNR	DRD	MPM
1	OTSU	82.2200	15.7700	8.7200	15.6400
2	SAUV	82.5400	15.7800	8.0900	9.2000
3	NIBL	68.5200	12.7600	28.3100	26.3800
4	BERN	47.2800	7.9200	82.2800	136.5400
5	GATO	82.1100	16.0400	5.4200	7.1300
6	LELO	80.8600	16.1300	104.4800	64.4300
7	SNUS	85.2000	17.1600	15.6600	9.0700
8	PROPOSED	84.2471	14.7388	13.4057	7.1721

Figure. 7.17 Testing on competition dataset DIBCO 2011

The above figure 7.17 shows the evaluation results of DIBCO 2011 dataset. The evaluation metrics that are used to evaluate the DIBCO 2011 dataset are F-measure, PSNR (Peak Signal to Noise Ratio), DRD (Distance Reciprocal Distortion) and MPM (Misclassification Penalty Metric). The DIBCO 2011 dataset consists of eight degraded handwritten documents and eight degraded printed documents. The proposed method is compared with the state of art techniques on DIBCO 2011 dataset. These methods include OTSU's method (OTSU) [1], Sauvola's method (SAUV) [18], Niblack's method (NIBL) [5], Bernsen's method (BERN) [3], GATO's method (GATO) [26] are compared with the proposed method. Besides these comparison methods the proposed method is also compared with top algorithms namely Lore et al.'s method (LELO) [37] and the method submitted by a team (SNUS) for the DIBCO 2011 dataset. The proposed method achieves good score in F-measure and PSNR values. The proposed technique performs good score in terms of DRD and MPM, which means that the proposed technique maintains good text stroke contours and provides best visual quality. The proposed method produces good results on all the testing images which is reflected on the good

F-measure score. BERN, NIBL and LELO methods fail to produce reasonable results for some test images from DIBCO 2011 dataset. SNUS method instead remove too much character strokes. The proposed method produces quite reasonable results with little noise for more challenging test images taken from DIBCO 2011 dataset where BERN, NIBL methods fail to produce good results. The proposed method produces a binary result with better visual quality and contains most of the text information. However, the proposed method produces the binary result with a little over binarization for the test images with high text stroke variation of the input test image taken from DIBCO 2011 dataset.

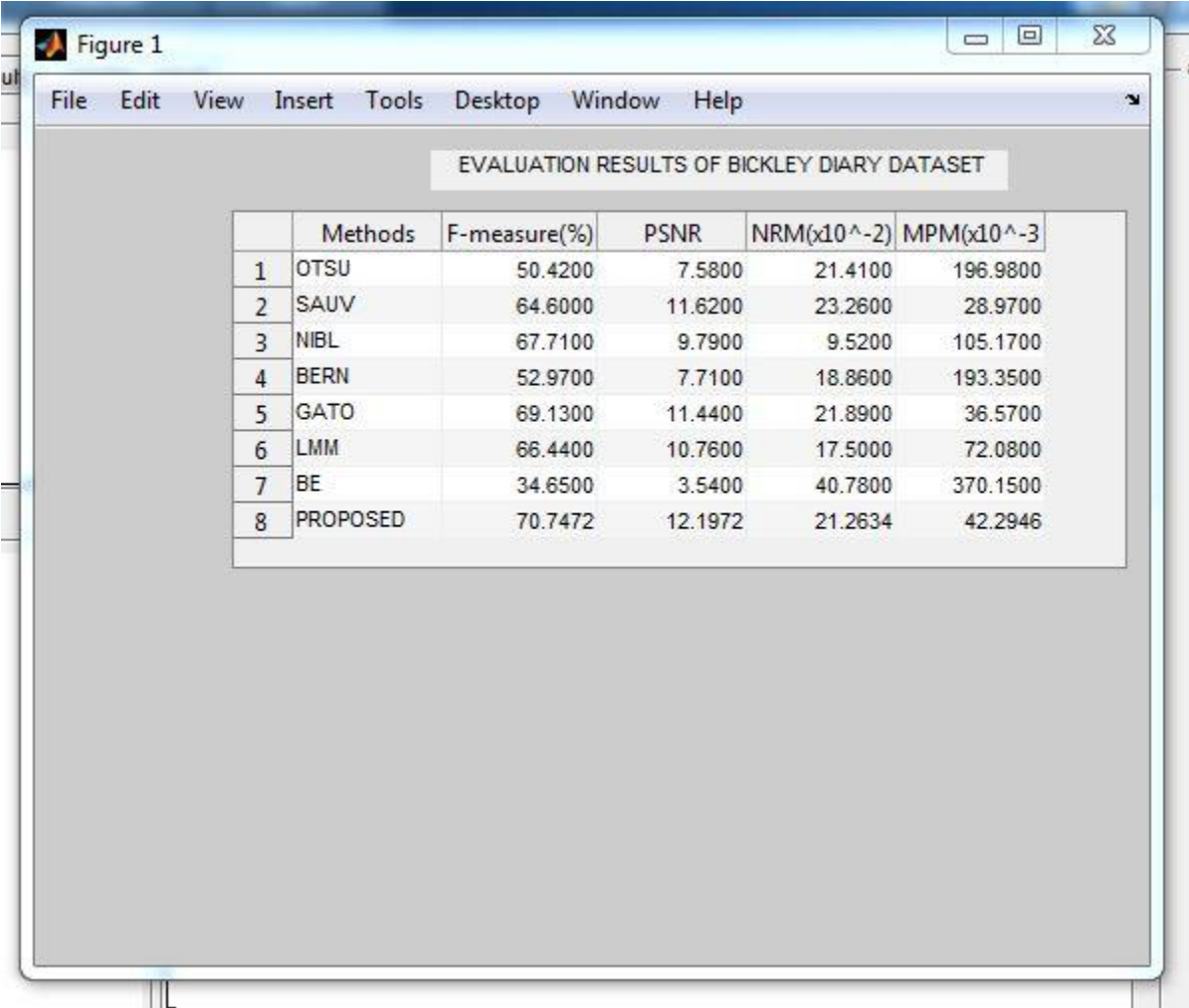


Figure 1

File Edit View Insert Tools Desktop Window Help

EVALUATION RESULTS OF BICKLEY DIARY DATASET

	Methods	F-measure(%)	PSNR	NRM( $\times 10^{-2}$ )	MPM( $\times 10^{-3}$ )
1	OTSU	50.4200	7.5800	21.4100	196.9800
2	SAUV	64.6000	11.6200	23.2600	28.9700
3	NIBL	67.7100	9.7900	9.5200	105.1700
4	BERN	52.9700	7.7100	18.8600	193.3500
5	GATO	69.1300	11.4400	21.8900	36.5700
6	LMM	66.4400	10.7600	17.5000	72.0800
7	BE	34.6500	3.5400	40.7800	370.1500
8	PROPOSED	70.7472	12.1972	21.2634	42.2946

Figure. 7.18 Testing on bickley diary dataset

The above figure 7.18 shows the evaluation results of bickley diary dataset. There are seven ground truth images to evaluate the proposed method with other methods. The proposed method achieves average 70.74% accuracy in terms of F-measure which is at least 10% higher than the other seven methods (OTSU's method (OTSU) [1], Sauvola's method (SAUV) [18], Niblack's method (NIBL) [5], Bernsen's method (BERN) [3], GATO's method (GATO) [26], Local maximum and minimum method (LMM) [39] and Background Estimation (BE) [43]). Therefore, the proposed method performs better than other methods in terms of PSNR, NRM, MPM values also by preserving most textual information and producing least noise for the images taken from bickley diary dataset.

## **CHAPTER 8**

### **CONCLUSION AND FUTURE WORK**

#### **Conclusion**

The proposed method presents an adaptive image contrast based document image binarization technique that is tolerant to different types of document degradation such as uneven illumination and document smear. The proposed technique is simple and robust, only few parameters are involved. Moreover, it works for different kinds of degraded document images. The proposed technique makes use of the local image contrast that is evaluated based on the local maximum and minimum. The proposed method has been tested on the various datasets. Experiments show that the proposed method out performs with other document binarization methods in term of the F-measure, pseudo F-measure, PSNR. The proposed method involves several parameters, most of which can be automatically estimated based on the statistics of the input document image. This makes the proposed technique more stable and easy-to-use for document images with different kinds of degradation.

#### **Future work**

The superior performance of the proposed method can be explained by several factors. First, the proposed method combines the local image contrast and the local image gradient that help to suppress the background variation and avoid the over-normalization of document images with less variation. Second, the combination with edge map helps to produce a precise text stroke edge map. Third, the proposed method makes use of the text stroke edges that help to extract the foreground text from the document background accurately. But the performance on Bickley diary dataset and some images of DIBCO contests still needs to be improved, and they may be explored it in future.

## CHAPTER 9

### REFERENCES

- [1] N. Otsu, "A threshold selection method from gray level histogram," *IEEE Trans. Syst., Man, Cybern.*, vol. 19, no. 1, pp. 62–66, Jan. (1979).
- [2] J. Kittler and J. Illingworth, "On threshold selection using clustering criteria," *IEEE Trans. Syst., Man, Cybern.*, vol. 15, no. 5, pp. 652–655, Sep.–Oct. (1985).
- [3] J. Bernsen, "Dynamic thresholding of gray-level images," in *Proc. Int. Conf. Pattern Recognit.*, Oct. (1986), pp. 1251–1255.
- [4] J. Canny, "A computational approach to edge detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 8, no. 6, pp. 679–698, Jan. (1986).
- [5] W. Niblack, *An Introduction to Digital Image Processing*. Englewood Cliffs, NJ: Prentice Hall, (1986).
- [6] L. Eikvil, T. Taxt, and K. Moen, "A fast adaptive method for binarization of document images," in *Proc. Int. Conf. Document Anal. Recognit.*, Sep. (1991), pp. 435–443.
- [7] A. Brink, "Thresholding of digital images using two-dimensional entropies," *Pattern Recognit.*, vol. 25, no. 8, pp. 803–808, (1992).
- [8] M. van Herk, "A fast algorithm for local minimum and maximum filters on rectangular and octagonal kernels," *Pattern Recognit. Lett.*, vol. 13, no. 7, pp. 517–521, Jul. (1992).
- [9] J. Parker, C. Jennings, and A. Salkauskas, "Thresholding using an illumination model," in *Proc. Int. Conf. Doc. Anal. Recognit.*, Oct. (1993), pp. 270–273.
- [10] J.-D. Yang, Y.-S. Chen, and W.-H. Hsu, "Adaptive thresholding algorithm and its hardware implementation," *Pattern Recognit. Lett.*, vol. 15, no. 2, pp. 141–150, (1994).
- [11] N. Papamarkos and B. Gatos, "A new approach for multithreshold selection," *Comput. Vis. Graph. Image Process.*, vol. 56, no. 5, pp. 357–370, (1994).
- [12] O. D. Trier and A. K. Jain, "Goal-directed evaluation of binarization methods," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 17, no. 12, pp. 1191–1201, Dec. (1995).
- [13] O. D. Trier and T. Taxt, "Evaluation of binarization methods for document images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 17, no. 3, pp. 312–315, Mar. (1995).

- [14] Y. Liu and S. Srihari, "Document image binarization based on texture features," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 5, pp. 540–544, May (1997).
- [15] D. Ziou and S. Tabbone, "Edge detection techniques—An overview," *Int. J. Pattern Recognit. Image Anal.*, vol. 8, no. 4, pp. 537–559, (1998).
- [16] M. Cheriet, J. N. Said, and C. Y. Suen, "A recursive thresholding technique for image segmentation," in *Proc. IEEE Trans. Image Process.*, Jun. (1998), pp. 918–921.
- [17] T. Lindeberg, "Edge detection and ridge detection with automatic scale selection," *Int. J. Comput. Vis.*, vol. 30, no. 2, pp. 117–156, (1998).
- [18] J. Sauvola and M. Pietikainen, "Adaptive document image binarization," *Pattern Recognit.*, vol. 33, no. 2, pp. 225–236, (2000).
- [19] R. Cao, C. L. Tan, Q. Wang, and P. Shen, "Double-sided handwritten archival documents," in *Proc. Int. Workshop Doc. Anal. Syst.*, (2000), pp. 147–158.
- [20] C. Wolf and D. Doermann, "Binarization of low quality text using a markov random field model," in *Proc. Int. Conf. Pattern Recognit.*, (2002), pp. 160–163.
- [21] I.-K. Kim, D.-W. Jung, and R.-H. Park, "Document image binarization based on topographic analysis using a water flow model," *Pattern Recognit.*, vol. 35, no.1, pp. 265–277, (2002).
- [22] G. Leedham, C. Yan, K. Takru, J. Hadi, N. Tan, and L. Mian, "Comparison of some thresholding algorithms for text/background segmentation in difficult document images," in *Proc. Int. Conf. Document Anal. Recognit.*, vol. 13. (2003), pp. 859–864.
- [23] M. Sezgin and B. Sankur, "Survey over image thresholding techniques and quantitative performance evaluation," *J. Electron. Imag.*, vol. 13, no. 1, pp. 146–165, Jan. (2004).
- [24] S. Kumar, R. Gupta, N. Khanna, S. Chaudhury, and S. D. Joshi, "Iterative multimodel subimage binarization for handwritten character segmentation," *IEEE Trans. Image Process.*, vol. 13, no. 9, pp. 1223–1230, Sep. (2004).
- [25] Y. Chen and G. Leedham, "Decompose algorithm for thresholding degraded historical document images," *IEE Proc. Vis., Image Signal Process.*, vol. 152, no. 6, pp. 702–714, Dec. (2005).
- [26] B. Gatos, I. Pratikakis, and S. Perantonis, "Adaptive degraded document image binarization," *Pattern Recognit.*, vol. 39, no. 3, pp. 317–327, (2006).

- [27] Blayvas, A. Bruckstein, and R. Kimmel, "Efficient computation of adaptive threshold surface for image binarization," *Pattern Recognit.*, vol. 39, no. 1, pp. 89–101, (2006).
- [28] A. Dawoud, "Iterative cross section sequence graph for handwritten character segmentation," *IEEE Trans. Image Process.*, vol. 16, no. 8, pp. 2150–2154, Aug. (2007).
- [29] E. Badekas and N. Papamarkos, "Optimal combination of document binarization techniques using a selforganizing map neural network." *Eng. Appl. Artif. Intell.*, vol. 20, no. 1, pp. 11–24, Feb. (2007).
- [30] S. Kumar, R. Gupta, N. Khanna, S. Chaudhury, and S. D. Joshi, "Text extraction and document image segmentation using matched wavelets and MRF model," *IEEE Trans. Image Process.*, vol. 16, no. 8, pp. 2117–2128, Aug. (2007).
- [31] S. Nicolas, J. Dardenne, T. Paquet, and L. Heutte, "Document image segmentation using a 2D conditional random field model," in *Proc. Int. Conf. Doc. Anal. Recognit.*, Sep. (2007), pp. 407–411.
- [32] B. Gatos, I. Pratikakis, and S. Perantonis, "Improved document image binarization by using a combination of multiple binarization techniques and adapted edge information," in *Proc. Int. Conf. Pattern Recognit.*, Dec. (2008), pp.1–4.
- [33] G. Kuk, N. I. Cho, and K. M. Lee, "Map-MRF approach for binarization of degraded document image," in *Proc. Int. Conf. Image Process.*, (2008), pp. 2612–2615.
- [34] Q. Chen, Q. Sun, H. Pheng Ann, and D. Xia, "A double-threshold image binarization method based on edge detector," *Pattern Recognit.*, vol. 41, no. 4, pp. 1254–1267, (2008).
- [35] B. Gatos, K. Ntirogiannis, and I. Pratikakis, "ICDAR 2009 document image binarization contest (DIBCO 2009)," in *Proc. Int. Conf. Document Anal. Recognit.*, Jul. (2009), pp. 1375–1382.
- [36] E. Saund, J. Lin, and P. Sarkar, "Pixlabeler: User interface for pixel-level labeling of elements in document images," in *Proc. Int. Conf. Document Anal. Recognit.*, Jul. (2009), pp. 646–650.
- [37] T. Lelore and F. Bouchara, "Document image binarisation using Markov field model," in *Proc. Int. Conf. Doc. Anal. Recognit.*, Jul. (2009), pp. 551–555.
- [38] B. Su, S. Lu, and C. L. Tan, "A self-training learning document binarization framework," in *Proc. Int. Conf. Pattern Recognit.*, Aug. (2010), pp. 3187–3190.

- [39] B. Su, S. Lu, and C. L. Tan, "Binarization of historical handwritten document images using local maximum and minimum filter," in Proc. Int. Workshop Document Anal. Syst., Jun. (2010), pp. 159–166.
- [40] F. Deng, Z. Wu, Z. Lu, and M. S. Brown, "Binarizationshop: A userassisted software suite for converting old documents to black-and-white," in Proc. Annu. Joint Conf. Digit. Libraries, (2010), pp. 255–258.
- [41] H. Yi, M. S. Brown, and X. Dong, "User-assisted ink-bleed reduction," IEEE Trans. Image Process., vol. 19, no. 10, pp. 2646–2658, Oct. (2010).
- [42] Pratikakis, B. Gatos, and K. Ntirogiannis, "H-DIBCO 2010 handwritten document image binarization competition," in Proc. Int. Conf. Frontiers Handwrit. Recognit., Nov. (2010), pp. 727–732.
- [43] S. Lu, B. Su, and C. L. Tan, "Document image binarization using background estimation and stroke edges," Int. J. Document Anal. Recognit., vol. 13, no. 4, pp. 303–314, Dec. (2010).
- [44] N. Howe, "A Laplacian energy for document binarization," in Proc. Int. Conf. Doc. Anal. Recognit., Sep. (2011), pp. 6–10.
- [45] Pratikakis, B. Gatos, and K. Ntirogiannis, "ICDAR 2011 document image binarization contest (DIBCO 2011)," in Proc. Int. Conf. Document Anal. Recognit., Sep. (2011), pp. 1506–1510.
- [46] T. Lelore and F. Bouchara, "Super-resolved binarization of text based on the fair algorithm," in Proc. Int. Conf. Document Anal. Recognit., Sep. (2011), pp. 839–843.