A New and Efficient Technique to Remove Back-to-Front Interference in Historical Document Images

Thesis submitted in partial fulfillment of the requirements for the award of degree of

Master of Technology

in

Computer Science and Applications

Submitted By

Bhuvnesh Malik

(Roll No. 601403010)

Under the Supervision of:

Dr. Rajiv Kumar
Assistant Professor

COMPUTER SCIENCE AND ENGINEERING DEPARTMENT
THAPAR UNIVERSITY
PATIALA-147004

JUNE, 2016

Certificate

I hereby certify that the work which is being presented in the thesis entitled, **"A New and Efficient Technique to Remove Back-to-Front Interference in Historical Document Images"**, in partial fulfillment of the requirements for the award of degree of Master of Technology in Computer Science and Applications submitted in Computer Science and Engineering Department of Thapar University, Patiala, is an authentic record of my own work carried out under the supervision of *Dr. Rajiv Kumar* and refers other researcher's work which are duly listed in the reference section.
The matter presented in the thesis has not been submitted for award of any other degree of this or any other University.

Bhuvnesh Malik

Bhuvnesh Malik

This is to certify that the above statement made by the candidate is correct and true to the best of my knowledge.

Dr.Rajiv Kumar
Assistant Professor
CSED

Countersigned by:

Dr. Maninder Singh
Head
Computer Science and Engineering Dept.
Thapar University
Patiala

Dr. S.S.Bhatia
Dean (Academic Affairs)
Thapar University
Patiala

Acknowledgment

First and foremost, I would like to thank my supervisor **Dr. Rajiv Kumar** for his guidance and encouragement through out my work. He have set a high academic and professional standard for me to follow. This proposition work was empowered and supported by his vision and thoughts. I have been amazingly lucky to have a counsel like him who gave me the flexibility to investigate new ideas on my own and in the meantime he guided me to recoup when my steps faltered. His understanding and backing dependably helped me overcome many crisis situations and effectively finish this dissertation.

I would like to thank **Dr. S.S.Bhatia**, Dean of Academic Affairs, for giving provisions of the entire required infrastructure such as computer labs, library facilities, immensely useful for learners to equip themselves with the latest in the field.

I take this opportunity to express my appreciations towards the **Dr. Maninder Singh**, Head CSED, for his kind help and cooperation, and other faculty members for their constructive suggestions through out the research. I am also thankful to all my respected teachers in the Department and all my friends, for their direct or indirect help, inspiration and motivation.

I want to express my greatest gratitude to my dear parents for their endless love, constant support.

Bhuvnesh Malik

Bhuvnesh Malik

Abstract

The study of historical records is a topic that presents significant difficulties for specialists from different fields, for example, history, political science, brain research, software engineering, among others. Historical archives contain significant information about cultural and scientific value. Historical artifacts comprise of archives, letters, daily papers, pictures, maps, and so on. A large number of these are put away in libraries, historical centers or government files. In any case, because of the conservation, few individuals have admittance to this material. Also such documents are frequently degraded over time. The main aim of preserving historical archives is to reestablish the corrupted content containing data. Digital preservation is the foremost requirement of today to preserve the historical documents. Digital preservation is the most promising method for preserving the historical documents. But if the document is old age document then this method may not give the best results and may give results which might be difficult to read. Now this further enhances the problem if the document is written on both sides, there may be some impression of back image to the front image, which reduces the readability of the document. This phenomenon is known as "back-to-front interference" also called "bleeding" or "show-through". The thresholding of such records with conventionally used algorithm produces unreadable documents. Generally numerous binarization strategies are implemented in the literature for various sorts of binarization issues. The few simple available binarization strategies cannot be applied to numerous binarization issues. Keeping in mind the end goal to enhance the quality of historical archives pictures, a joined methodology based on mathematical morphology, global and local binarization strategies is applied. This technique at initial step removes the background (noise) and adjusts the intensity values of pixels of the historical image, which enhances the performance of global thresholding. After that, the mathematical erosion operation is applied on this binarize image, then local binarization is applied to binarize the image. The proposed technique effectively denoises different sorts of corrupted records, enhancing textures with clear background. The present study proposed new and efficient technique to eliminate back-to-front type of interference in such type of historical artifacts.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Overview

Scientific and military purposes were the main reasons for the invention of computers, where less data had to be entered and more computations were performed. On the other hand, in business applications, the data to be entered is huge and computations to be performed are low. The exchange of data between humans and computers is a challenging problem in business applications. Even today, direct keyboard entry is the most commonly used method by an operator. This process of data entry is very slow and has a possibility of introducing human errors. It can also slow down the process of data acquisition. Therefore, a feasible outcome would be computers doing this job for humans where computers shall transform the raw document into some intermediate form and process it within lesser time which in turn will have fewer errors. The presence of human operator would only be necessary if the system has problems with recognition, for correction purpose. Evolution of Optical Character Recognition took place at this time. OCR is the system of transforming scanned documents or pictures into machine readable characters. The input to OCR can be scanned pictures of handwritten or printed text in any language or images captured by digital camera or even PDF files. The OCR recognition mechanism converts this text into editable text.

Eyes are optical mechanism in case of human beings where input to the brain is the image as seen by the eyes. There are many factors that vary from person to person which affect the human capability to identify these inputs. OCR is a mechanism that matches the human capability of recognizing inputs. While human recognition capability cannot be matched by OCR, printed text and handwritten characters can be recognized by it. The quality of input documents directly affects the OCR performance.

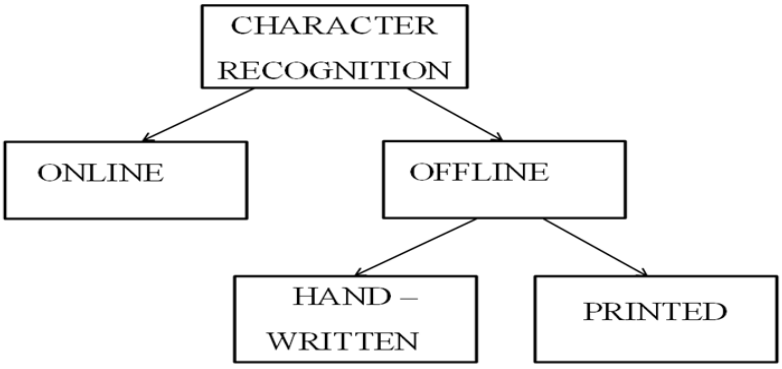


Figure 1.1: Classification of Character Recognition System.

Character acknowledgment is also called as optical character acknowledgment which is currently a subject of study and has vast possibility in future also where we would need to track and find all bit of data that is being traded around the world. There are various problems with handwritten text recognition due to variation in calligraphy, similarity in text patterns, and variation in writing styles. Also, if the images are captured by a digital camera rather than scanning, often suffer from distorted edges and illumination variations, cause difficulties for the OCR application in recognizing the text correctly. OCR makes it easy for us to interact with the computers easily making our tasks efficient.

1.2 Types of Handwriting Input: Offline and Online

The process of associating a meaning with the components of an image like letters, numbers and symbols written or printed on it is called Character Recognition. Character recognition mechanism takes scanned data as an input and then applies various pre-processing, classification and recognition techniques to process the image and detect the components. When writing or printing is complete and the document is ready for scanning, offline optical character recognition is performed whereas on-line recognition is performed when the computer detects the components instantly as and when they are written or caligraphed. OCR recognizes both printed and handwritten characters but the quality of the input directly determines its performance.
We can classify character recognition system in the subsequent types on the base of data acquisition as shown in figure 1.1 :

The performance of OCR is satisfactory if the input is constrained. OCR machines still have a long way for reading as well as humans for unconstrained inputs. We can broadly classify character recognition into the categories as:

- Online Handwriting Recognition
- Offline Handwriting Recognition

The process of identifying words that are obtained by scanning documents like a paper and storing them in digital format is called offline handwriting recognition. After storing, further processing is performed for recognition.

When the data is captured and stored directly in digital form, it is called online handwriting recognition. The data is acquired by various mediums such as writing on a tablet, PC or digitizer which converts the characters automatically into digital format. Pen-tip motion and strokes are identified by a sensor in these devices. Usually, an electronic surface is accompanied by a pen specifically designed for this purpose. The coordinates of consecutive points in two-dimensions are symbolized as a task of time and are then kept in successive order as the pen moves on the electronic surface.

Better results can be achieved by online character recognition than offline recognition. The fact behind this is that more information is captured if data acquisition is online such as order, direction and speed of strokes of the handwriting. Online Character Recognition has constant relevant data however offline data does not, a major variance between online and offline recognition. This variation generates a major deviation in processing methods.

1.3 Historical Background of OCR

Initially in 1929, Tauschek got a patent on OCR in Germany, trailed by Handel who got a US patent on OCR in USA in 1993 (U.S Patent 1,915,993). Tauschek was additionally allowed a US patent(U.S.Patent 2,026,329) on his technique in 1935. Tauschek machine was a mechanical gadget that utilized templates. A photo detector was put so that when the format and the character to be perceived were lined up for a definite match, and a light was coordinated towards it, no light would reach the photo detector. The United States Postal Services has been using OCR machine to separate mail since 1965 based on engineering devised primarily by the productive inventor Jacob Rabinow.

The modern version of OCR is said to have originated in 1951 with David Shepard's invention: GISMO-A Robot Reader Writer. This invention was closely followed by Jacob Rabinow's prototype machine in 1954, which was able to read upper case, printed output at the sluggish speed of one character per minute [1]. David Shepard's company Intelligent Machine Research (IMR) is said to have been the first apply optical reading technique in a commercial situation. They installed a system at Reader's Digest in 1955 [7]. A number of other companies, including IBM, were conducting research into OCR throughout the early 60's culminating in the first marketable commercial OCR system: IBM's 1418 [8]. Early systems, such as the one mentioned above, were very constrained in the sense that they were bound to read special, artificial fonts. These types of OCR systems are usually associated with the first generation of OCRs. The second generation, on the other hand, was characterized by hand-printed character recognition capabilities [8]. In the early stages, only numerals could be recognized by this pioneering machine. One such device was IBM's 1287 OCR system. This was the first of the second generation machines, and one of the most famous ones. It was originally exhibited at the world's fair in 1965 [8]. It has been approximately 40 years since the first automatic handwritten readers were suggested. From that point forward, fantastic advancement has been made to empower PCs to perceive, translate and distinguish machine printed and hand written text. Broad exploration has been completed as far as specialized papers and reports by different scientists around the world.

They have divided the commercial OCR framework into four generations relying upon flexibility, robustness and proficiency. First generation includes commercialized OCR, IBM 1418, while the second generation includes famous OCR, IMB 1287. Third generation has OCR's for hand printed and poor print quality characters developed during 1975-1985 [1]. Fourth generation has OCRs of documents with text, graphics, tables and mathematical symbols etc. The OCR of this generation generally has an accuracy of less than 85%. Research on complex documents is in progress [9]. Research into handwriting recognition still keeps on being serious during all these years. The inspiration might be ascribed to the testing way of the character acknowledgment issue and the endless number of the business applications that it might be applied [10].

1.4 Significance of OCR and Its Usage

Optical character recognition (OCR) mechanism acquires input by scanning a document and converts the text in the image into digital text that is editable. There are innumerable remunerations of using OCR software, from saving space

to speedy searches:

- No retyping: If we accidentally delete or lose a vital digital file, but still have hard copy of it, you can easily get it back by using OCR to scan the document.
- Quick searches: Scanned text is converted into a readable and editable file by OCR, allowing us to search for a keyword or phrase in the document.
- Edit text: We have the choice to edit text in any word editor once we have a scanned copy of a document due to required updating needed with time.
- Save space: We can easily free all the space occupied for storing documents by scanning all the documents. This can turn a cabinet filled with documents into editable files on a CD which are worthy of vital information.
- Accessibility: "Ease of Access tool" or "Accessibility" can be termed as OCR software. Books, magazines, faxes, mails, or other documents can be scanned into word processing programs to be used along with text to speech utility.

1.5 Applications of Offline Handwriting Recognition

Offline handwritten recognition has the following application areas:

- Banking
  OCR has the capability to process bank checks without human intervention which makes it widely used in banks. Bank checks can be easily acquired into an image by a phone camera followed by scanning the text on it and then transfer the amount of money successfully. This mechanism gives almost cent percent accuracy in printed checks while it is fairly accurate for hand written checks. It is implemented in hand written checks by occasional human confirmation before transfer of money. This mechanism speeds up the banking process.
- Legal Industry
  Legal industries are also digitizing papers and documents now-a-days. This allows them to save spaces required to store documents. It also makes it simple to search any file once the documents are scanned and a database is maintained. Quick and easy access is therefore available to legal professionals and a huge database which is stored in digital format and can look up for them easily as they are text-searchable.

- Other Industries
  Numerous different fields, for example, education, finance, and government organizations likewise utilize OCR. It has prepared innumerable text accessible online for researchers plus students saving their money and making a large source of knowledge to be shared just with a few clicks and search keywords.

- Vocal Monitoring
  Audio information is more effective than written text in some cases especially for the visually impaired humans. This appeal is strong enough while we may still focus on other visual sources of information. Henceforth the vision of text to speech came up.

1.6 Optical Character Recognition

OCR is a procedure which relates a typical importance with items (literatures, images and figures) drawn on a picture, i.e. OCR method relates a symbolic identity with the picture of a character [8]. OCR can likewise be characterized as the procedure of changing over scanned pictures of machine printed or handwritten manuscript (numbers, letters and images), into a PC procedure capable arrangement. A definitive objective of an OCR is to mimic the human capacity to peruse at a much quicker rate by associating symbolic identities with picture of characters. The pragmatic significance of OCR applications, and in addition the intriguing way of the OCR issues has prompted extraordinary interest and quantifiable development in this field [1] .

1.7 Components of an OCR System

All the types of character recognition in various applications can be summed up by a single term “Character Recognition” which is processing of input patterns based on textual content to create meaningful outputs by machines. The input may come from On-line devices like tablets, stylus based devices or Off-line devices like scanners. Output may be a sequence of symbols like “Y”, “E” , “S” or a date on cheque like “Nov 12,2015” or validation result of a signature.

Hierarchical tasks grouped into stages of the character recognition are included in character recognition mechanism as preprocessing, segmentation, feature extraction, classification and post-processing.

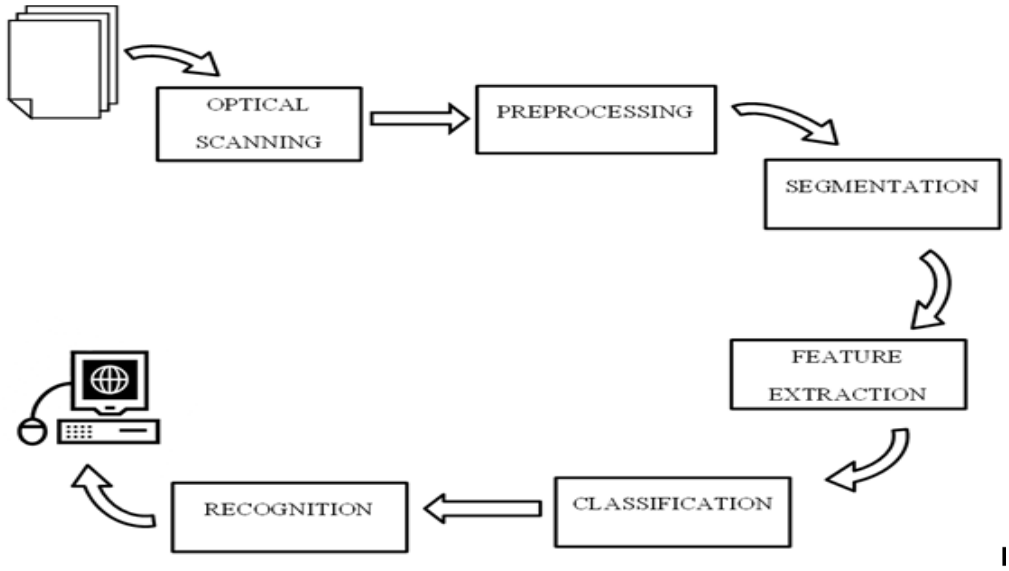


Figure 1.2: Phases of OCR.

Digitization of the document using an optical scanner is the first step in character recognition. After scanning, the regions having text are located in the document and segmentation techniques are applied to extract each symbol. The extracted symbols are then processed further, noise elimination, to ease the extraction of features in future. Each symbol is then identified by comparing it with the features of symbol classes obtained by machine learning phase. The derived information is used to reconstruct the words and numbers of the original text at last.

Some of the methods involved in the above process are described in more details in figure 1.2 .

1.8 Digitization

It refers to the procedure of changing over a paper or film based record into electronic form. The electronic transformation is refined through imaging a procedure whereby a report is scanned and an electronic representation of the first, as a bitmap picture, is created. The imaging procedure includes recording changes in light intensity reflected from the archive as a matrix of dots. The light/color estimations of every dot are stored in binary digits. One bit would require for every dot in a binary output, though up to 32 bits could be required per dot for a color scan. Digitization creates the computerized picture, which is sustained to

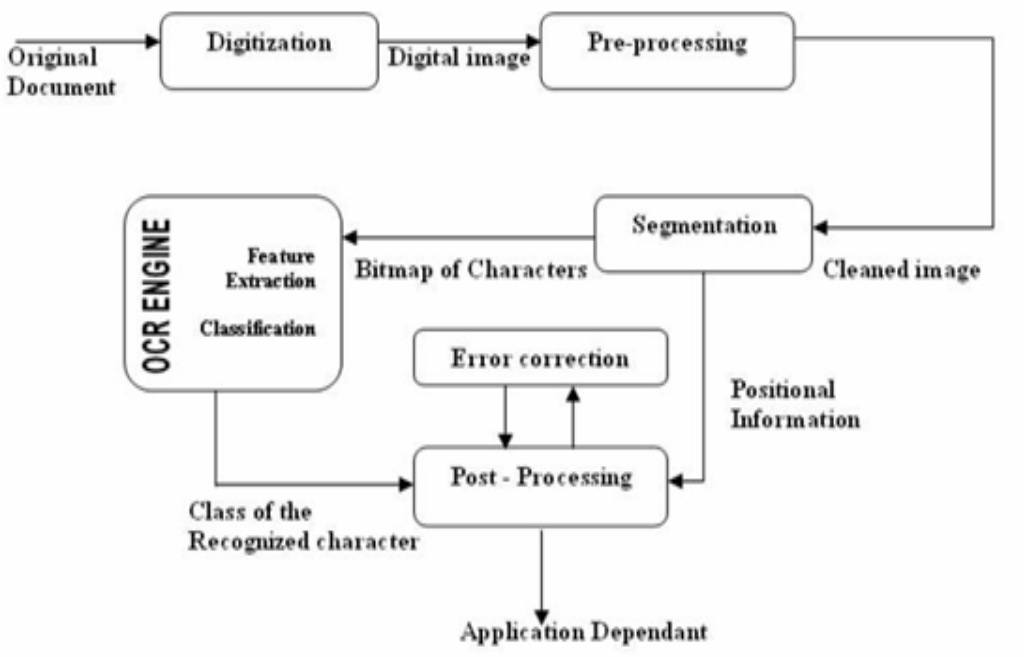


Figure 1.3: Components of OCR [1]

the pre-processing stage [11].

1.9 Preprocessing

Depending on its type of data acquisition, the raw information is subject to a lot of primary image processing steps in order to make it functional in the following stages of character recognition. The aim of preprocessing is to yield data which is easy for the character recognition systems for accurate results of recognition. Preprocessing involves a family of procedures for filtering, smoothing, cleaning-up, enhancing and creating an image so that following algorithms in the subsequent stages can perform accurately to allow easy final classification. Preprocessing methods as shown in figure 1.4 include noise removal, normalization, compression, smoothing, skew correction and thinning. Removing any undesirable bit-patterns, which don't have importance in the output, is the main objective of noise removal. It also simplifies pattern acknowledgement process without losing any important data. It diminishes any conflicting information, if present. It also improves the picture for the afterward stages.

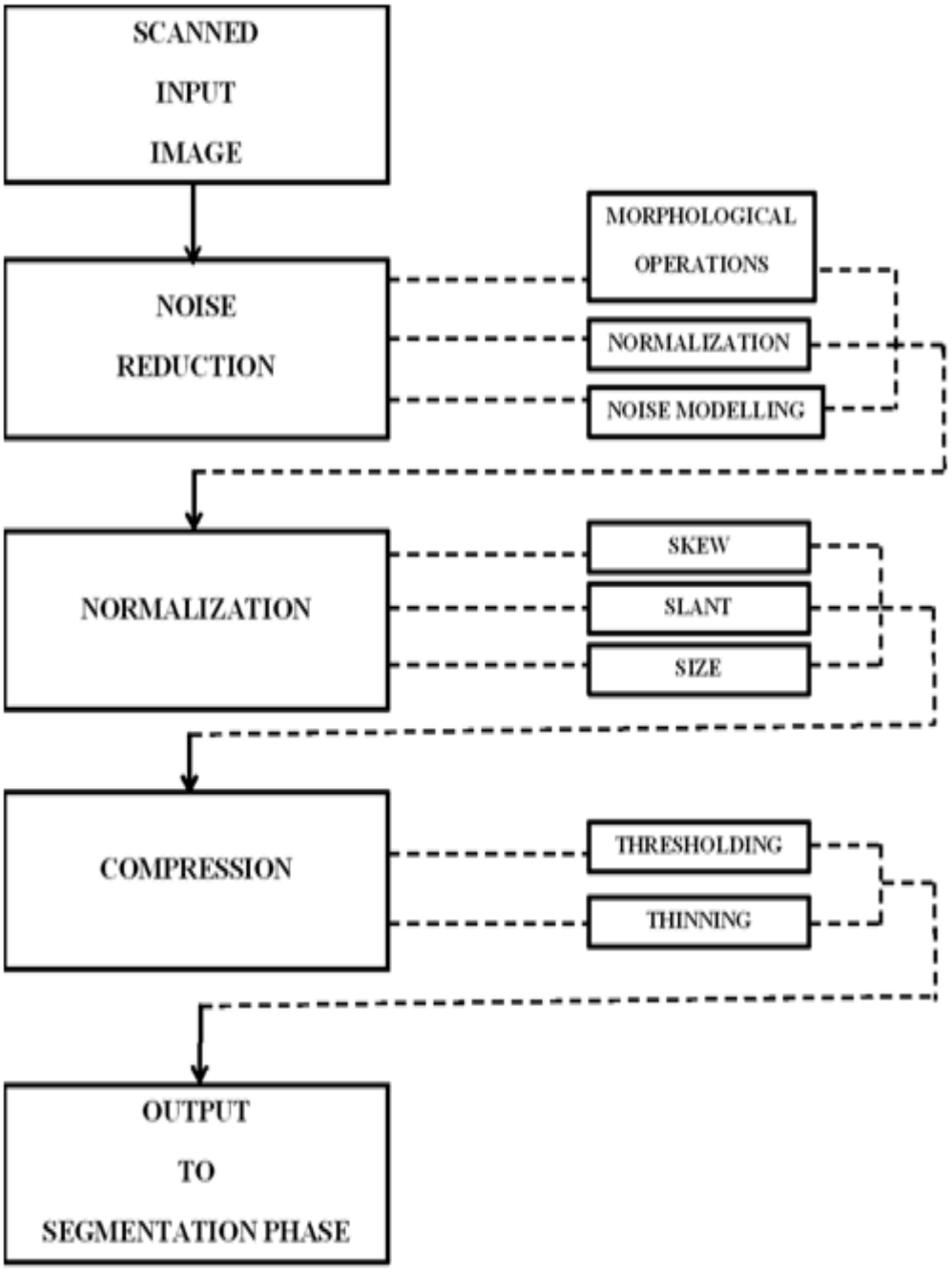


Figure 1.4: Various Stages in Preprocessing.

1.9.1 Overview of Preprocessing Techniques

The preprocessing phase can be further subdivided into three main categories: Noise reduction or removal, normalization of text and compression of the image. Each of these sub phases have various techniques and algorithms that are applied to prepare the image for subsequent phases. Each of these are described in detail below:

Noise Reduction

The scanning device or the writing instrument usually introduce some noise, which causes distortion, including local variations, line segments get disconnected, bumps and gaps are introduced in lines, dilation, and erosion occurs etc. It is essential to eliminate these inadequacies before the Character Recognition.

There are three major groups of noise reduction techniques:-

- Filtering: To eliminate noise and reduce unnecessary points (shown in figure 1.5) typically presented by irregular writing platform and/or poor rate of sampling of information acquisition device is called filtering. A value is assigned to every pixel which is a function of the gray values of its neighbouring pixels. Various filters are planned for contrast adjustment purposes, smoothing, sharpening, thresholding and eliminating slightly textured or colored background.

- Morphological Operations: We perform these operations to replace the convolution operations by logical operations in order to filter the document image. Morphological operations are designed for connecting the broken stokes and decompose the strokes which are connected, contour smoothing, thinning of characters and boundary extraction. So, we can say that morphological procedures may be used for removal of noise in the artifacts.

- Noise Modelling: Some standard techniques can be used to remove noise if there was a model available for it. But it is not always possible to remove the noise in most applications as not much work has been done in noise modelling. The noise which arises due to optical distortion like blur, skew or speckle should be modelled. However, we can still remove noise by assessing the quality of documents to a certain degree.

Normalization of Data

Removing the variations in writing and obtaining standardized data is the main objective of normalization. The basic methods for normalization are as follows:

Figure 1.5: Image with Salt and Pepper Noise and Image after Noise Removal.

- Skew Normalization and Baseline Extraction: Sometimes different writing styles or errors in scanning process may cause the writing to appear curved or slightly tilted in the image as shown in figure 1.6. The effectiveness of successive algorithms is decreased by this. Therefore, detection and correction of these inaccuracies is necessary. The relative position w.r.t. the baseline (for example "9" and "g") helps us differentiate various characters from each other.

- Slant Normalization: The average angle of near-vertical elements gives slant estimation. The slant angle between the longest stroke in a word and the vertical direction is one of the most considerable factor in different handwriting styles. All the characters are normalized to a predefined standard using slant normalization as shown in figure 1.7 .

- Size Normalization: Adjusting the size of characters to a standard size is called size normalization. The character is distributed into a number of regions where every region is scaled independently and hence normalization is performed both horizontally and vertically.

- Contour Smoothing: Elimination of all the errors introduced in the document due to inconsistent handwriting styles is called contour smoothing. It improves the efficiency of the successive steps in preprocessing as it reduces the sample points required in-order to represent the document.

Figure 1.6: Skewed Image [2]

Figure 1.7: Image before and after Slant Normalization [3]

Compression

Image compression technique is used to reduce the unnecessary bits in the image due to redundancy. This technique reduces the overall size of the image.
Thresholding and thinning are the two most popular compression techniques described in detail as follows:

- Thresholding: Thresholding is used to decrease the storage necessities and increase the processing speed of the images where gray scale images as represented as binary pictures based on a threshold value. Thresholding can be majorly classified into two types based on the threshold value: global and local.
  When a single threshold value is picked for the whole archive image by taking an approximation of the background intensity level from the intensity histogram of the picture is called Global Thresholding.
  If different values are assigned to each pixel according to the local neighbourhood of the pixel then it is called Local or adaptive thresholding. We can define the neighbourhood according to our need.

- Thinning: The extraction of information about the shape of the characters is called Thinning. It is the conversion of offline data to almost online like data. Thinning has two main approaches: pixel wise and non-pixel wise.
  If the image is processed iteratively using local information till one pixel wide skeleton is left, it is called Pixel wise thinning. These are quite sensitive to noise and may cause deformation in shape of characters. If global information of characters is used then it is called Non-pixel wise thinning.

1.9.2 Thresholding

One of the techniques used for image compression is thresholding. It is performed to reduce storage and increase the performance by converting gray scale pictures into binary pictures based on a threshold value: global or local.

The information of the image is binary. The data which the image carries is not likely to have only two levels of intensities. It can have a range of intensities. This occurs both due to non-uniform printing and non-uniform illumination of the image. It results into intensity transitions at the edges.

The main objective of thresholding is to identify the pixels that belong to foreground region with a single intensity (“on”) and background region with a different intensity (“off”). This separates the regions of an image corresponding to the objects of our interest. To distinguish the pixels we need to break down from

the remaining, a comparision of each pixel intensity regarding a threshold value is performed. When we have isolated the required pixels, we can allocate them with an altered quality to classify them which means a value of 0 (representing black), 255 (representing white) or any other value according to our needs can be assigned.

Global Thresholding

Global Thresholding selects an optimal or several optimal threshold values automatically in order to separate the items of our importance in a picture from the background based on their gray-level histogram distribution. It assigns a single unique threshold value to the whole image on basis of an approximation of the background intensity level from the intensity histogram of that picture. If the pixel intensity values of the components can be distinguished easily from the background for the entire image then a global threshold value can be chosen appropriately. The concept of using a single or global value for thresholding is called as global thresholding.

The histogram for an image which has well-distinguished background and foreground intensities will have two distinct peaks. The intensity value is chosen at the valley between the two peaks that best separates the two peaks as it is the minimum between the two maxima as shown in figure 1.8.

The thresholding technique is demonstrated below in the figure 1.11 when global thresholding is applied. The low threshold gives an unclear image while the high threshold removes some of the important details from the image. Hence, the selection of good threshold is necessary for thresholding technique to achieve its objective.

Local Thresholding

It is noticed that due to noise and poor contrast, the images do not always have a well differentiated background and foreground. Therefore, a single value for thresholding these images is not a feasible approach. So we divide the picture into several sub pictures and then threshold each of these sub picture separately. Since each pixel defines its threshold value based on its placement in the image, we can also call this technique as adaptive thresholding.

The gray-level intensities are analyzed within the local windows across the entire image in-order to determine the local thresholds.Image histogram for local thresholding is shown in figure 1.12. The local area information of the pixel determines

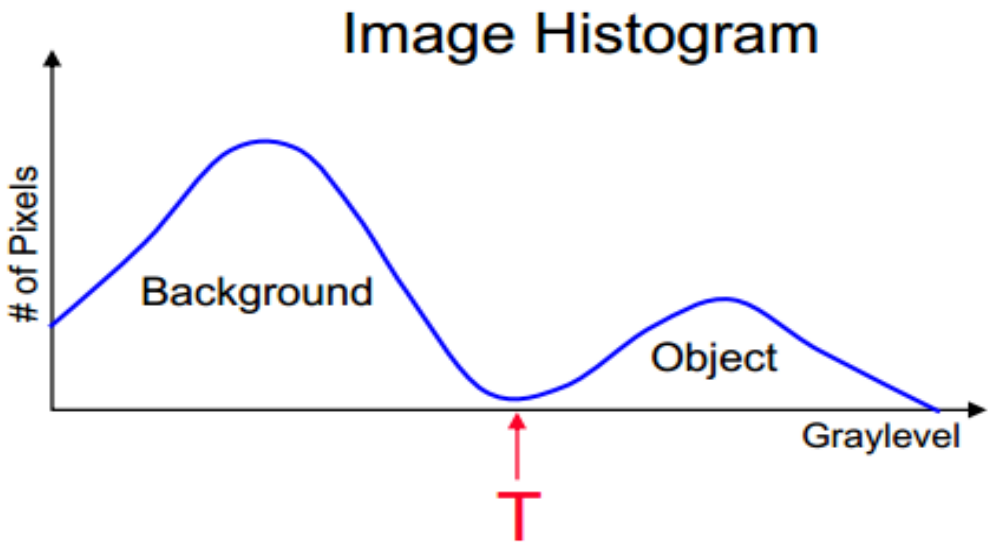


Figure 1.8: Image Histogram for Global Thresholding.

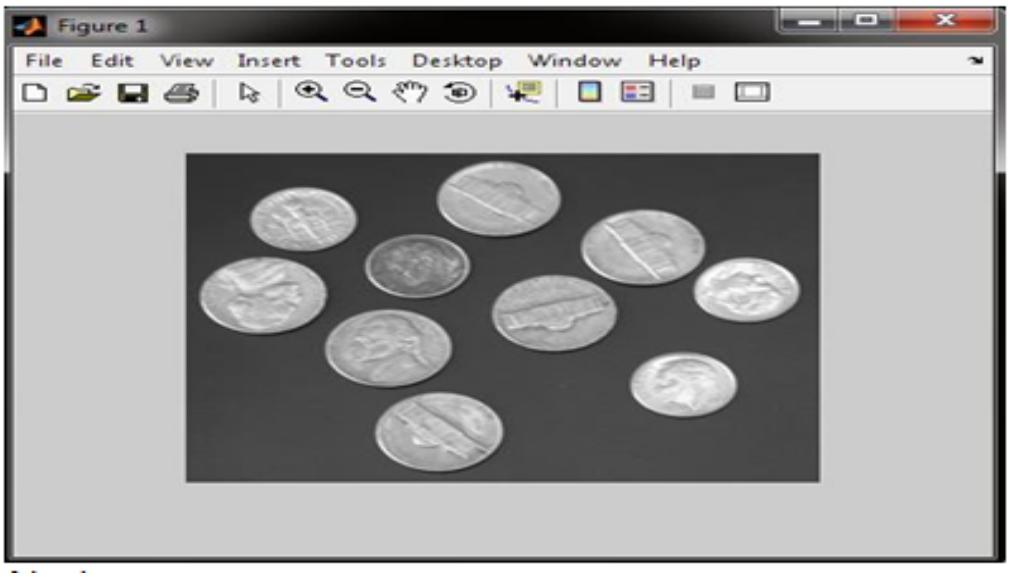


Figure 1.9: Grayscale Image for Global Thresholding.

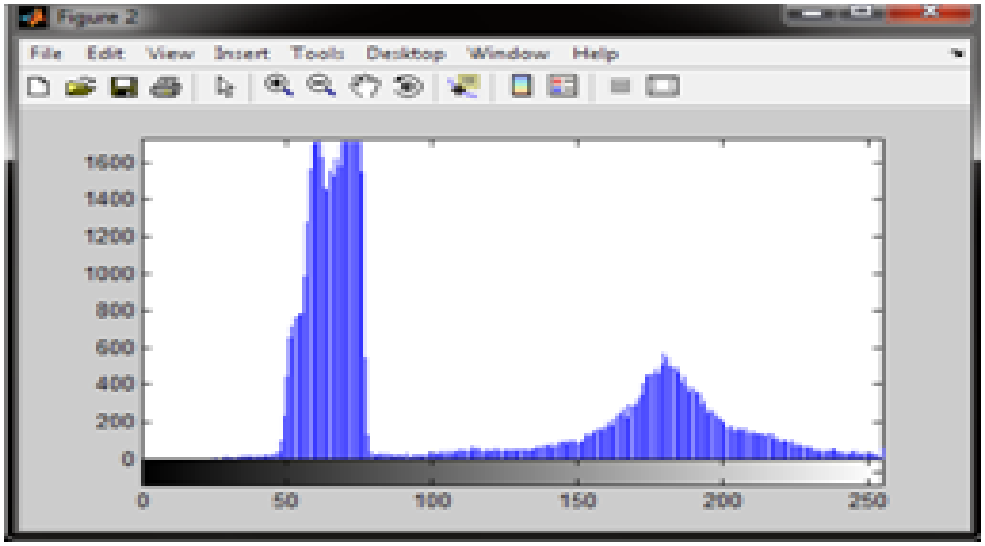


Figure 1.10: Histogram for Image in Figure.1.9

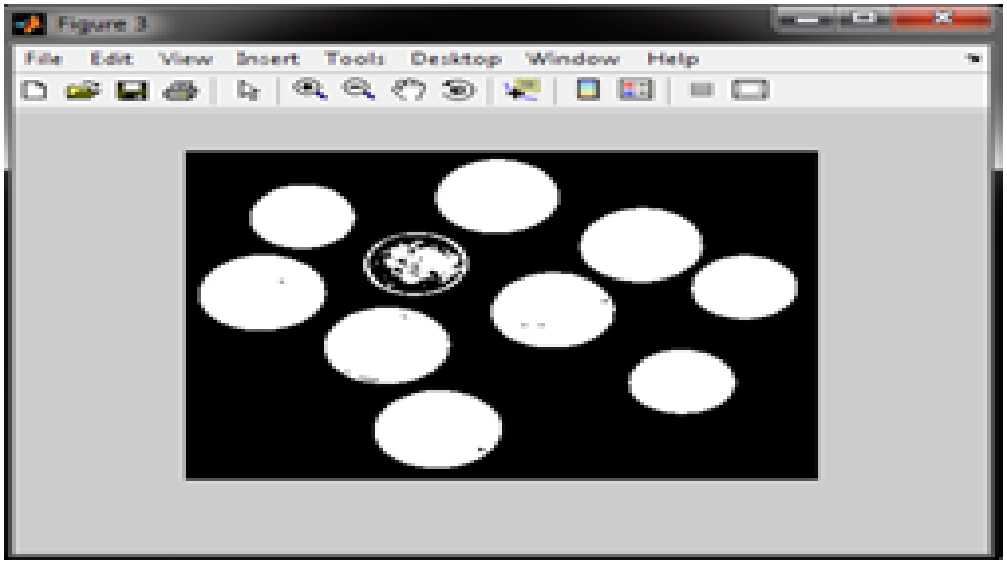


Figure 1.11: Image after Global Thresholding.

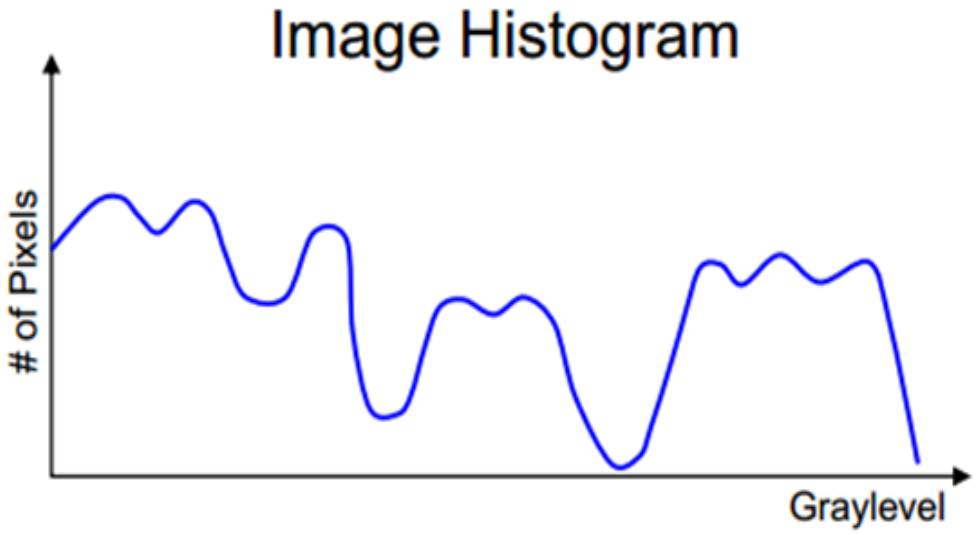


Figure 1.12: Image Histogram for Local Thresholding.

the threshold value to be assigned to it. But the choice of the window size is a tricky problem. The window size should be chose in a manner that it is large enough to allow a satisfiable number of pixels from the background in-order to obtain a good estimate of the average values but not too large to pick over average non- uniform intensity of background. As shown in figures 1.13 and 1.14.

1.10 Segmentation

We obtain a "clean" document image from the preprocessing stage which has extracted sufficient information about the shape of characters, has high compression and low noise. Segmentation is dividing the document image into sub components. It is important to segment properly because the recognition rate of characters in successive stages is directly affected by the extent of separation reached in lines, words and characters. It decides the outcome of recognition phase. If this stage yields incorrect output, the final outcome will not be desirable. So, it is the decision process for desirable outcome in OCR.

Segmentation is mainly of two types: external and internal. The isolation of writing units like paragraphs, sentences and words is called external segmentation whereas isolation of characters especially in cursive handwriting is called internal segmentation.

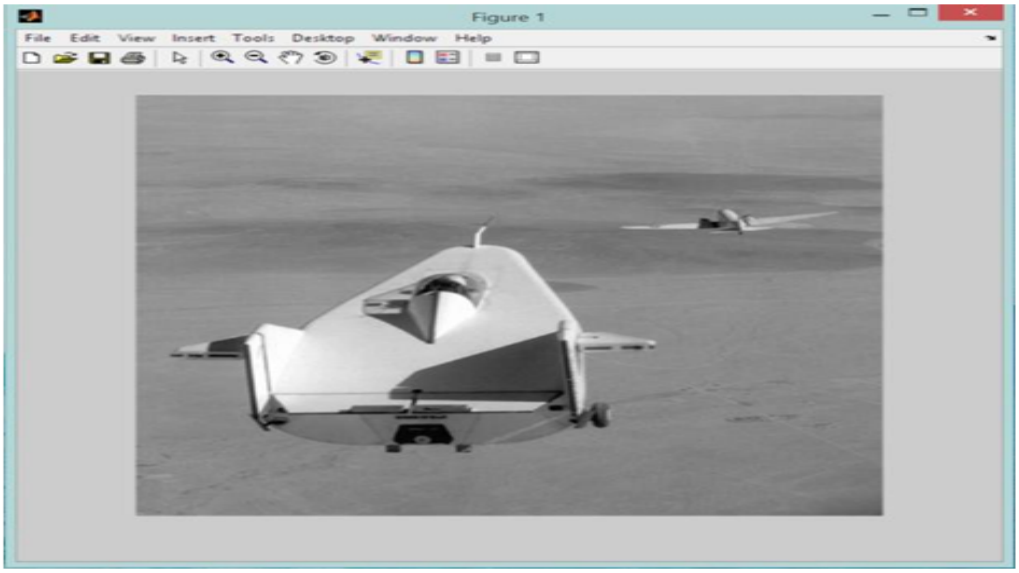


Figure 1.13: Grayscale Image for Local Thresholding.

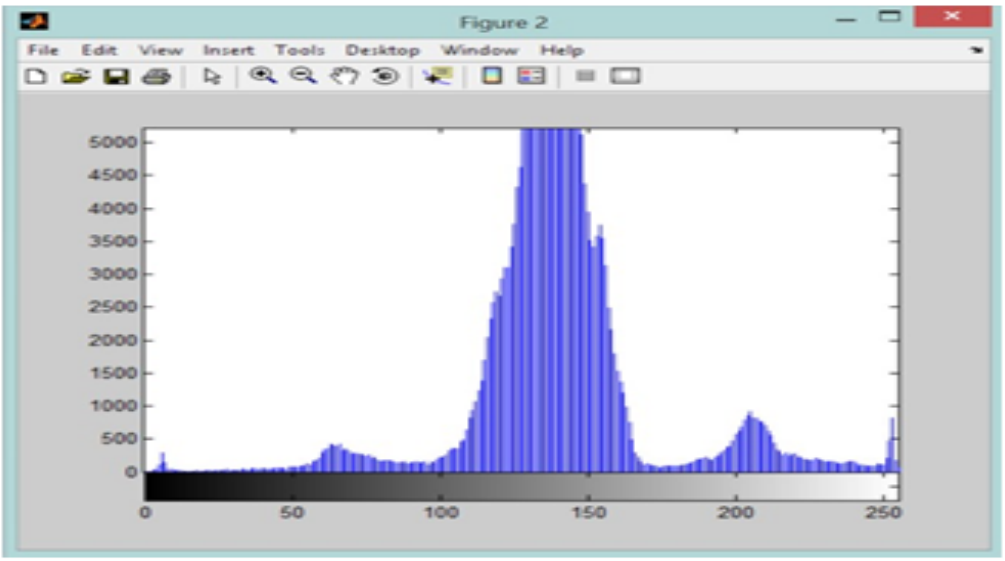


Figure 1.14: Histogram for Image in Figure 1.13.

The need of segmentation arises due to the fact that handwritten characters interfere with each other frequently. Some ways in which they can interfere are overlapping each other, touching or intersecting, connected letters etc. We also need text-graphics segmentation in-order to isolate text from images, lines and graphs because we desire an output containing text only. It is a major step in OCR especially for cursive handwritten documents where the letters are connected together. After segmentation, the characters are size normalized for improving accuracy. We can then extract the features from the characters which are of the same size to maintain uniformity in data.

The characters which are broken or have separate parts are grouped. A bounding box is used to completely enclose the region. If the bounding box of one region completely encloses another region for any two regions then the enclosed region is labelled to the value of enclosing region. Therefore, the final region contains two disjoint sub regions.

The most common problem in segmentation occurs due to confusion in text and graphics when the document contains joints and split characters. These joints and splits occur due to scanning process. Joints occur when the document is scanned at low threshold and splits occur if the document is scanned at high threshold. Joints can be spotted in dark photocopy and splits occur in light photocopy. So, OCR gets confused while segmenting characters connected to graphics.

1.10.1 Overview of Segmentation Techniques

Image segmentation is an area of research which requires a high degree of attention. There are a lot of different segmentation techniques that can be applied but there is no single technique which can be used for all types of images. Different techniques suit different types of images and various image segmentation techniques are shown in figure 1.15 . The technique developed for a particular type of image might not be useful for the other types of images. Hence, there is lot of difficulty to develop a universal segmentation approach that can be used to segment all types of images. The selection of a particular segmentation approach for a particular type of image is also not an easy task.

We can broadly classify image segmentation techniques into the following categories: According to two aspects of images:

1. Identifying Discontinuities: Partitioning an image on the basis of sudden changes in intensity i.e. edge detection.

Main Categories	Sub Classes	
Edge Base segmentation	Grey Histogram Technique	
	Gradient Based	Differential coefficient technique
		Laplacian of a Gaussian
		Canny Technique
		Watershed technique
Region Based	Thresholding	Global Thresolding
		Local Thresolding
		Dynamic Adaptive Thresolding
	Region Operating	Region growing
		Region Splitting and Merging
Special Theory Based	Clustering	K-means
		Fuzzy
	Neural Network	

Figure 1.15: Organization of Picture Segmentation Techniques.

2. Identifying Similarities: Partitioning an image into different regions on the basis of similarity according to a predefined criterion i.e. binarization, area growing, area splitting and merging.

1.10.2 Segmentation Based on Edge Detection

Edge detection aims to identify all the points in an image document at which there is a sudden change in intensity or has discontinuity or when there is a jump in intensity from one pixel to the neighbouring pixel. It is of huge importance for image analysis. Edges define the image boundaries which are very helpful for segmentation. There are a lot of ways to accomplish edge detection; but they can be assembled into two principle classifications:

1. Gray histogram method: In this method, histogram is calculated firstly on the basis of the color or intensity of the all pixels in the image, and then valleys and edges in image are found. Here, the segmentation is dependent on the threshold value that is chosen. The efficiency of this technique is more in comparison to other techniques. But the method cannot be used if the valleys and edges detected are large.

2. Gradient based method: Gradient based methods are best suited for images with sudden changes in intensity near edges and images having little

noise. It is calculated as the first derivative f(x,y) of the image. Convolving gradient operators are used in this technique. If the value of gradient magnitude is high, it indicated that there is rapid conversion between two regions. This identifies edge pixels which should be linked in-order to form closed boundary of regions. Sobel, laplacian of gaussian (log), laplace and sobel operators are some of the commonly used edge detection operators. The best out of these operators is canny but it is more time consuming that sobel. A balance must be maintained between noise immunity and detecting accuracy while practise of edge detection. In the event that the level of precision is too high, then noise can raise fake edges which make unreasonable picture diagram. In the event that the level of noise immunity accomplished is too high, then some parts of picture layout may stay undetected and object position cannot be right. Therefore, this method can generate equally good results on complex and noisy images along with simple and noiseless images.

3. Watershed segmentation method: Watershed Segmentation is a segmentation technique in which a "watershed" is formed when an image is "flooded" by its local minima and "dams" are formed wherever waterfronts meet. All the dams form a watershed together when the image is fully flooded. This watershed of edgness image can then be used for segmentation. The idea behind this method is visualization of the edgeness image as a three- dimensional landscape where the objects are the catchments basins. Object boundaries are formed in the watershed of edgeness image which mark catchment basins. There can be some defects due to image artefacts but they do not affect the watershed segmentation much. We need both preprocessing and postprocessing of the watershed image to achieve good segmentation results but over segmentation should be avoided. Postprocessing of watershed image includes the filling of boundaries to obtain solid segments.

1.10.3 Region Based Segmentation Methods

Region based segmentation is applied to partition an image in regions which have some similarity based on some predefined criterion. Region splitting and merging, region growing, thresholding etc. are some of the techniques which belong to this category.

1. Thresholding Method: Thresholding is applied to images when we need to speed up processing and reduce the requirements for storage. It is done by converting grayscale images into binary images based on a threshold value. These methods works according to picture space segmentation of regions. The basic idea behind this method is based on the features of a picture.

Thresholding is applied to select a threshold value which divides the image into various classes and therefore separates the objects which belong to the background. If a single threshold value is picked for the whole picture then we can say that any pixel intensity value which is greater than the threshold value belongs to the object and the pixel intensity values which are smaller than the threshold value belong to the background.

2. Region Operating Methods: Image segmentation segments an image into similar or homogenous regions. The methods we discussed above performed segmentation based on threshold values which were chosen based on the pixel information. Whereas region based methods obtain the complete required region directly. The only limitation of this method is that it is more time consuming.

   (a) Region Growing In this method, pixels having similar properties are grouped together to create a region. It is performed as follows:

   - Find a seed pixel as an underlying point for each of required segmentation.
   - Join together the comparable or like properties of pixel (Based on a foreordained developing or comparative recipe to decide) with the seed pixel around the seed pixel area into the space of seed pixel.
   - These new pixels go about as another seed pixel to proceed with the above procedure until no more pixels that fulfill the condition can be incorporated.

   (b) Region Splitting and Merging In this method, the picture is subdivided into an arrangement of irregular disjoints areas and after that union and/or split the region as per the given condition for segmentation. Region based segmentation is to a great extent affected by splitting. This splitting method can be represented to as quad trees in which each hub has precisely four branches. It is performed as:

   - Split the area into the four disjoint branches.
   - When no further splitting is possible, combine any area.
   - When no further combining is possible, then stop.

In this manner, the segmentation can be done by splitting and merging method. The limitation of this method is that it is difficult and time-consuming technique.

1.10.4 Theory Based Segmentation

This type of segmentation has a lot of segmentation techniques which are derived from various different fields and are quite imperative for segmentation. Some the algorithms based on this technique are wavelet based, neural network based, fuzzy based, clustering based, genetic algorithms etc.

1. Clustering Techniques: Grouping of similar images in a database is called as clustering. The basic idea behind this method is to increase the effectiveness of storage, quickly retrieval and to get desirable results. Various properties of an image like size, colour, texture etc. are considered while performing clustering.

    - Fuzzy c means clustering : This technique identifies the natural groups of data in-order to form a large set of data which produces a brief representation of the systems behaviour. Fuzzy c-means is one of the information clustering techniques in which datasets are gathered into n groups where each information point from the dataset belongs to each group to a specific degree.
    - K-Means Algorithm : This calculation groups information vectors into a predefined number of groups. The centroids of the predefined groups are instated randomly at first. The measurements of the centroids are same as the measurements of information vectors. Euclidian distance measure is used to determine the proximity of pixels and to assign them to clusters. Mean of each cluster is re-calculated after assigning all the pixels to respective clusters. This is repeated either until no progressions are seen for all cluster implies or for a fixed number of iterations.

2. Neural Network-based segmentation: The picture is mapped to a neural network in this algorithm. Each pixel of the image is identified as a neuron of the neural network. After this, the edges are found by applying dynamic equations for directing the state of each neuron to minimal energy defined by neural network.

    There are three basic characteristics of neural network based segmentation

    - To a great degree parallel capacity and quick processing capacity make it well-suited for real-time application.
    - Unrestricted nonlinear degree and high collaboration among handling units makes this technique skilled to set up displaying for any strategy.
    - Reasonable robustness makes it unaffected to noise.

Various limitations of neural network based segmentation are:

- We must have segmentation information of various kinds beforehand.
- The result of segmentation can be manipulated by initialization.
- Prior learning processes are needed by neural networks.
- The training period must not be too long and overtraining should also be avoided.

1.10.5 Model Based Segmentation

The above segmentation methods use only local information about the objects or pixels. Since humans have the capability to distinguish items even if they are not separated or represented completely, we can derive the conclusion that the information only about the local neighbourhood is insufficient for performing segmentation. Instead highly accurate and detailed information about the geometrical form of items is mandatory. This information can then be matched with the local information. This is the basic concept behind model based segmentation.

1.11 Feature Extraction

One of the most important roles in recognizing an image is played by feature extraction. Here, a binary or grayscale image is fed to a recognizer in the simplest case. But for most recognition systems, an additional compact and distinctive image is obligatory for avoiding complexity and increasing accuracy of the algorithms. In-order to accomplish this, a set of features are extracted from objects/letters to form a feature vector for each class. This feature vector helps the recognition system to distinguish an object from different classes while remaining insusceptible to the typical differences inside a class. The classifier recognizes the input units with the target output units by using these feature vectors. This makes it easy for the classifier to classify among various classes by considering these features.

1.12 Classification

The CR procedure allots a character picture to a class by utilizing an order calculation taking into account the features extracted and the connections among the components. Since individuals from a character class are equal or comparable in as much as they share characterizing qualities, the estimation of similitude, either unequivocally or certainly, is key to any classifier. In this stage, we prepare the

neural net utilizing the feature vectors acquired amid feature extraction strategy against the required targets.

Feature extraction is concerned with recouping the characterizing attributes covered up by imperfect estimations. To represent a character class, either a model or an arrangement of tests must be known. The feature selection process endeavors to recoup the pattern attributes characteristic for every class. The classification stage recognizes every information character picture by considering the distinguished features.

1.13 Postprocessing

The consolidation of setting and shape data in all the phases of character acknowledgment frameworks is fundamental for significant enhancements in acknowledgment rates. This is done in the post-preparing stage with an input to the early phases of character acknowledgment. The most straightforward method for fusing the setting data is the usage of a word reference for amending the minor errors of the character acknowledgment frameworks. In post-processing, a dictionary can be used to restrict the character combinations. This can be implemented as a grammar that specifies all possible combinations of characters. The essential thought is to spell check the character acknowledgment yield and give a few other options to the yields of the recognizer that don't occur in the dictionary.

1.14 Historical Documents

Historical documents are of great importance to us due to their cultural and scientific value. Therefore, the study of historical artifacts is a huge challenge among scholars from different research areas like political science, history, psychology, computer science, and many others. Historical artifacts consist of archives, maps, pictures, newspapers, letters, etc. Most of these are put away in museums, libraries, and/or government archives. In any case, few individuals have access to this material because of the preservation. So as to profit simpler access to this rich source of information and facts about the history of a society, digitization of these is one of the possible solutions. Once these documents are digitized, they can be made available in digital libraries or on the Internet for wider dispersion. In-order to achieve high paperwork processing efficiency and utilization of data content, the paper documents must be converted into document images in electronic form which can be later converted to computer understandable format. However, the documents must be handled delicately during digitization. Now, to improve read-

Figure 1.16: Historical Documents [4]

ability and remove noise from these digitized historical document images, we need specialized processing techniques. Also, it is important to ensure that these documents are carefully processed and the information in them is recognized correctly so that the content in them is correctly accessible around the world. Also such documents are frequently degraded over time. Some types of degradation which appear in such documents frequently are paper deterioration and discoloration, presence of smear, smudges, or ink, non-uniform intensity, poor contrast due to humidity, etc. Therefore, specialized thresholding techniques are required for these document images to remove noise while keeping essential textual information. So we need to make the computers do this work automatically rather than humans doing this manually to attain historical and financial benefits along with preservation. Figure 1.16 shows some historical documents that are not in good condition [12].

1.14.1 Problems in Historical Documents

Paper is extremely delicate and vulnerable to aging still it is the essential medium to store information. There are lots of precious historical records preserved in libraries everywhere throughout the world. These important document are presented to progressive decay even with all the watchful treatment that the libraries offer. Earlier the paper was crafted with high amount of chemicals which speed up the degradation process. There are many degradation problems in historical

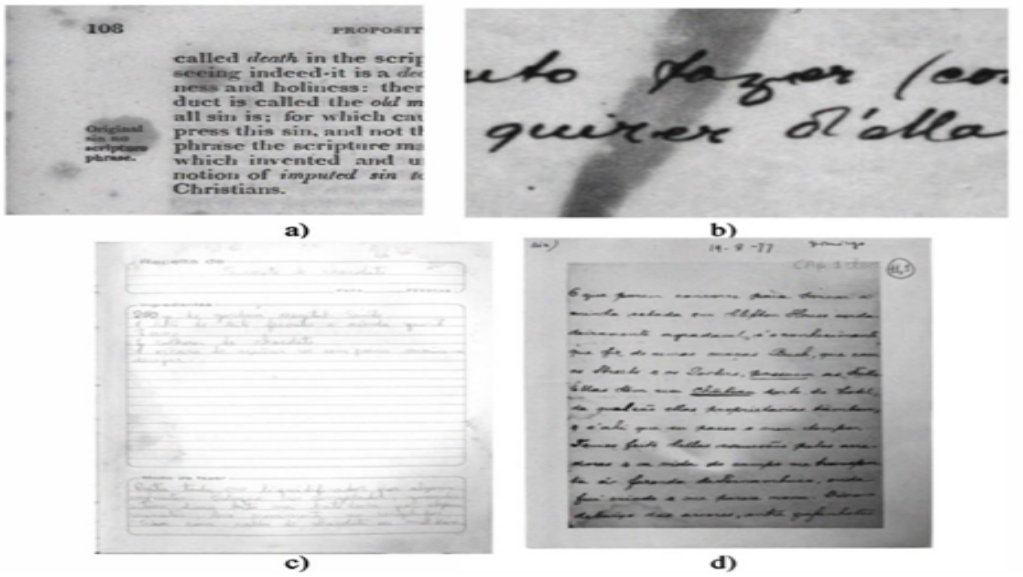


Figure 1.17: Sample Historical Documents with (a) Smudges (b) Marks of Adhesive Tape (c) Faded Ink and Different Levels of Degradation and (d) Non-Uniform Illumination in a Two-Color Paper.

documents which occur due to various different reasons [13] as shown in figure 1.17 .

Some of the problems found in historical artifacts are:

- dark or light brown tones comes in paper
- damages in the paper due to deterioration
- smudges, blemishes or dirt marks because of man-handling
- presence of smear, ink
- folding marks
- adhesive tapes marks
- non uniform intensity
- poor contrast due to humidity

The term degradation can be defined as “By degradation (or defects), we mean every sort of less-than ideal properties of real document images” as suggested by Henry S. Baird [14] .

1.15 Motivation behind Research

It is very important to preserve the historical stained documents which reveal the information about the civilized past. The importance of the information retrieval from these documents cannot be denied. These documents contain invaluable knowledge about our civilized past. Paper is a very important storage medium for storing information and distribution but it is very delicate and susceptible to aging. The information written in the paper can be machine printed and handwritten. With the time the information written in it is degraded, so it is very important to preserve such invaluable information. It is found from the archeological department that many historical documents belonging to 3rd century BC are also available and maintained. But not all available ones are in Good condition. To preserve this information for longer period, we require modern tools and techniques. There are extensive collection of ancient historical documents in libraries and museum all across the world. These documents are either printed or hand written in their native language. Typically only few people are allowed to access such collections, because preserving the documents is of great concern. Nowadays, libraries have started to digitize the historical documents that are of great attention to a huge number of persons, with the aim of preserving the content of the historical documents. Historical collections are of excessive attention to a huge number of people like students, scholars, historians etc. These collections are very important for them to study about our civilized past. It is found that the documents information is not completely visible because of some noises. There are different types of noises observed in the historical documents like smudges, marks of sticky tape, blurred ink and distinctive levels of deprivation, non-uniform illumination in a two-shading paper, back-to-front interference. All these noises makes difficult to access the necessary information from the documents. The most common noise found in the historical archive is back-to-front impedance. The present work is to remove the back-to-front impedance in the historical stained archive.

1.16 Objective of Thesis

The historical documents are the main source from where we can get the information about civilized past but these documents are affected with a various kind of noises and the information stored in them is losing day by day which is big matter of concern. This invaluable knowledge about our civilized past must be saved. In the ancient time paper was the main source for storing information but it is very fragile and susceptible to aging. The information stored in the paper is degraded with time. There are various kinds of noises which are responsible for losing the information but one of the most common noises found in the his-

Figure 1.18: Historical Artifacts with Back-to-Front Impedance [5].

torical stained document is “back-to-front interference” also known as “bleeding” or “show-through” [15]. This noise decreases the readability of the front text because the back printing causes interference in the front text of document. It is very important to preserve such type of documents so that they can be used in future to get some knowledge about our civilization. Document images - acquired either by scanners or digital cameras - almost always present these kinds of noisy artifacts [5].

Figure 1.18 shows some historical documents which are affected from the back to front interference. As we have a record composed on both sides of the paper and the ink has transposed from one side to the next inciting a hard impedance. Human visual system can separate both sides but this is not a simple task for computing systems. The main objective of this thesis is to implement such type of an efficient algorithm which can remove back-to-front interference from the historical stained documents. So that, it is easy to extract the information stored in such type of documents.

1.17 Thesis Organization

The thesis is organized in a sequential way starting with the description of historical documents and problems associated with them. Chapter 2, this chapter includes the description of the literatures that reveals the different techniques used so far to remove the back–to-front interference in the historical documents. Chapter 3 , this chapter reveals the problem definition. Chapter 4 , in this chapter the authors implemented new and efficient technique to eliminate the back-to-front type of interference problem in historical stained documents. Chapter 5,discussed the results of the proposed algorithms with other existing algorithms. Chapter 6, thesis is concluded with a summary and discussion on future research directions.

Chapter 2

Literature Review

The inexplicable research work done in character recognition and restoration of historical documents has led to the development of innumerable approaches to deal with the various aspects of the character recognition. The approach for the restoration process may vary according to the type of document under consideration. In order to understand the current state of art in this area, a survey of work done related to segmentation of degraded historical documents has been presented in this chapter.

2.1 Back to Front Interference

Whenever the artefacts are written on both the sides of translucent paper, the print of one side is visible on the other one called "back–to-front interference" or "bleeding" later called "show- through" [15]. The thresholding of artefacts containing interference of back-to-front by conventionally used algorithm produce unreadable documents. Figure 2.1 shows some Historical stained documents affected with bleeding.

Preprocessing is a phase in which the document image is cleaned and sent for the next phase. Thresholding is one of the techniques used in preprocessing phase [9]. The next section present the brief introduction of various thresholding techniques used to eliminate interference of back-to-front in the artefacts.

Figure 2.1: Historical Stained Documents Affected with Bleeding [6].

2.2 Techniques to Eliminate Back-To-Front Interference

This section presents some of the available techniques used to eliminate back-to-front interference in the historical artefacts. All techniques are based on a gray-scale image histogram threshold. The true-color to gray-scale conversion is done by:

$$G_{val} = 0.587G + 0.114B + 0.299R \tag{2.1}$$

Where G_{val} is the new pixel intensity and G, B and R are the green, blue and red estimations of the original pixel. The procedures displayed in this area take the picture histogram and standardize each of its entrances by the aggregate number of pixels in the picture, yielding a probability distribution gave by relative frequencies. In this manner,

$$p_a = \frac{m_a}{M}, 0 \leq a \leq 255 \tag{2.2}$$

$$p_b = \sum_{a=0}^{b} p_a \tag{2.3}$$

where m_a represent the quantity of pixels with gray level 'a' (where a=0 to 255), M represent the aggregate number of pixels in the picture, the set $p_0, p_1, \ldots, p_{255}$

is the probability distribution of the pixel gray-levels considering their relative frequencies, and P_b is the expansion of all probabilities from 0 to b.

Otsu's [1979] algorithm does not have a place with the class of algorithms taking entropy into account [16]. It is incorporated here on the grounds that it is a standout amongst the frequently utilized calculations as a part of picture division. Otsu's calculation creates discriminator examination by characterizing if a gray level 'b' will be mapped to item or background data. This procedure attempts to amplify the between-class variance $\sigma_B^2(b)$ given by:

$$\sigma_B^2(b) = p_b\bigg(\mu_\beta(b) - \mu_b\bigg)^2 + \bigg(1 - p_b\bigg)\bigg(\mu_\omega(b) - \mu_\tau\bigg)^2 \tag{2.4}$$

where

$$\mu_\beta(b) = \sum_{a=0}^{b} a\frac{p_a}{p_b} \tag{2.5}$$

$$\mu_\omega(b) = \sum_{a=b+1}^{255} a\frac{p_a}{1 - p_b} \tag{2.6}$$

$$\mu_\tau = \sum_{a=0}^{255} ap_a \tag{2.7}$$

The argument $'b'$ is the threshold that enlarges the between-class variance $\sigma_B^2(b)$.

Kapur et al. [1985] considers the frontal area and foundation pictures as different sources with following distributions [12]:

$$\beta : p(a) = \frac{p_a}{p_b}, 0 \leq a \leq b \tag{2.8}$$

$$\omega : p(a) = \frac{p_a}{1 - p_b}, b + 1 \leq a \leq 255 \tag{2.9}$$

The entropy of the two sources is calculated as, Abramson [17]

$$E_\beta = -\sum_{a=0}^{b} p(a)log(p(a)) \tag{2.10}$$

$$E_\omega = -\sum_{a=b+1}^{b} p(255)log(p(a)) \tag{2.11}$$

Where, $p(a)$ is dictated by Equations 2.8 and 2.9. The ideal threshold esteem that expands the entirety of the two entropies is figured as:

$$E(b) = E_{\beta}(b) + E_{\omega}(b) \tag{2.12}$$

Niblack's method [1986] computes the thresholding estimations of every window over the picture independently by the accompanying equation [18]:

$$T = M + C.S \tag{2.13}$$

Here M is the mean quality and S is the standard deviation estimation of the pixels inside the window. The estimation of C is in general 0.2.

Yen et al. [1995] follows the same thought as the one by Kapur and his associates in admiration to the forefront and foundation distributions [19]. An entropic relationship is characterized by:

$$TC(b) = C_{\beta}(b) + C_{\omega}(b) \tag{2.14}$$

$$TC(b) = -log\sum_{a=0}^{b}(\frac{p_a}{p_b})^2 - log\sum_{a=b+1}^{b}(\frac{p_a}{p_b})^2 \tag{2.15}$$

Furthermore, the threshold is the contention that boosts that expression. Where $C_{\beta}(b)$ and $C_{\omega}(b)$ are Ranyi entropy, with $= 2$.

Casey and Lecolinet [1996] gave a review of various techniques and methodologies used in character segmentation. The importance of segmentation in the recognition process and various steps of classical optical character recognition process have been discussed. They proposed the idea of dividing the segmentation strategies into three approaches: classical, recognition based and holistic. The classical approach which identifies the segments based on character like properties is discussed. The recognition based approach that finds those components in image that matches the classes in alphabet and the holistic method recognizes the word as a whole is explained. The dissection techniques like projection analysis, white space and pitch approach and connected component processing have been discussed [20].

Ha and Bunke [1997] presented a new approach to offline handwritten numeral recognition. They developed a recognition method which can account for a variety of distortions due to eccentric handwriting. The technique for the perturbation

based recognition system has also been discussed. The key idea of the perturbation approach, which lies in the process of reversing an input image back to one of its standard forms, has also been stated. The methodology for the parameterization process based on four geometric transformations, namely, rotation, slant, perspective view and shrink has also been explained. The approach to replace normalization by a set of perturbation processes modeling writing habits and instruments has also been discussed [21].

Mo and Mathews [1998] proposed an adaptive filter to be used prior to binarization for the edge enhancement of the characters. The importance of edge enhancement and noise reduction in the document image for the recognition process has also been explained. The paper also stated the similarity of proposed approach with the equalization of the binary communication channels. An introduction to the quadratic filter for the removal of the noise from the document acquired during image acquisition process has also been given. The mathematical description for the quadratic filter and its application has also been given. The design and implementation issues with the proposed algorithm have also been stated in the paper. The low pass and high pass filters and their application for binarization process has also been summarized [22].

Wu et al. [1998] calculated the same entropies evaluated by the Kapur, Sahoo and Wong's algorithm. But, instead of maximizing the addition of these, Wu, Songde and Hanqing minimize the difference given by [23]:

$$f(b) = |E_\beta(b) - E_\omega(b)| \tag{2.16}$$

Cai and Liu [1999] proposed an approach that integrates the statistical and structural information for unconstrained handwritten numeral recognition. The approach that uses the state duration adapted transition probability to improve the modeling of state duration in conventional markov models and uses macro states to overcome the difficulty in modeling pattern structures by markov models has also been explained. The technique for the encoding of the orientations into discrete codebooks and the distributions of locations are modeled by joint Gaussian distribution functions has also been discussed. The preprocessing methods for the disjoint connection region, slant correction, size normalization have also been dealt with [24].

Mello [2000] proposed the utmost frequent gray level of the picture and takes it like beginning limit to assess the values E_β, E_ω and E by conditions 2.10,2.11

and2.12 respectively,[25] however the entropies must be ascertained with the logarithm to the base M. The entropy E decides the estimation of weights m_β and m_ω If $E0.25$, then $m_\beta = 3$ and $m_\omega = 2$. If $0.25 < E < 0.30$, then $m_\beta = 2.6$ and $m_\omega = 1$. If $E0.30$, then $m_\beta = 1$ and $m_\omega = 1$. Furthermore, the threshold is straightforwardly figured by

$$T = 256(m_\beta E_\beta + m_\omega E_\omega) \tag{2.17}$$

Sauvola's technique [2000] was produced from Niblack's strategy. It intends to take care of the issue of black noise relying upon the effect on the standard deviation esteem by utilizing a scope of gray-level values in the pictures [26]. The binarization equation is:

$$T = M(1 + C(\frac{S}{D-1})) \tag{2.18}$$

Here, D is the dynamic scope of standard deviation and the parameter C gets positive values ($C = 0.5, D = 128$).

Arica [2001] proposed a guide for the researchers working in the CR area. He presented the historical evolution of CR systems and the techniques available for CR with detailed discussion of their pros and cons. His main focus was offline handwritten recognition as this is a popular area of research and advancements are required in this field. Therefore, he reviewed all the significant approaches which can be used in CR [11].

Mello and Lins [2002] designed a system for storage, indexing and network transmission of historical artefacts. For this purpose, he first decomposes the documents into their features such as texture of paper, colors, text is classified into printed and handwritten parts, pictures, etc. Common features or components are factored. After segmentation, paper and ink of the images are processed separately to extract their main features. The system generates a final synthetic version of images of historical documents which are good by both qualitative and quantitative measures [6].

Kasturi et al. [2002] gave a detailed description of the document image analysis process. The sequence of steps starting from data capture, pixel level processing, feature level analysis until text recognition and analysis has been elaborated. A brief analysis of graphical documents has been presented. The techniques for noise reduction and binarization have been discussed. The techniques for thinning

and region detection, chain coding and vectorization have also been explained. The techniques for line and curve fitting, critical point detection, skew estimation, layout analysis have also been discussed. The strategy for feature extraction and classification based on template matching and contextual processing has been explained. The various OCR's for Indian languages and document analysis in multilingual context has also been stated [15].

Feldbach and Tonnies [2003] proposed a new approach for segmentation of dates in historical documents of church registers from 18th and 19th century on the basis of predicting possible boundaries of a word along with the analysis of distance between different text objects. The algorithm uses a priori knowledge of semantic information and hypothesis of potential boundaries that can be generated. The problems in segmentation of such documents arises if the lines are not straight or if the words are touching or crossing each other. But their algorithm had an accuracy of 97% for correctly identifying objects. This was achieved by analysing the positions of boundaries between the words in word sequences with a limited number of variations. At present, the algorithm uses only parts of a line and can be extended to consider parts of the line close to potential parts of line [27].
Mello [2004] proposed a system for complete generation of synthetic historical document images that can be used for efficient storage and network transmission. This is done by segmenting the images into two classes, namely, paper and ink. The information in them is then re-assembled and a document close to the original document is synthesized. Texture is created and colourised automatically and an image is created using text from a text file with the help of OCR. This algorithm gives quantitatively and qualitatively good synthesized images [28].

Feng and Manmatha [2005] proposed a study to solve the historical handwritten manuscript recognition problem based on the comparison of support vector machines, conditional maximum entropy models and Naive Bayes with kernel density estimates. They focussed on whole word problem to avoid character segmentation. The results show that Naive Bayes with Gaussian kernel density estimates signicantly outperforms the other models and prior work using hidden Markov models on this heavily unbalanced dataset [29].

Gatos et al. [2005] proposed a novel segmentation free approach for keyword look in historical typewritten reports consolidating picture preprocessing, synthetic information creation, word spotting and client's feedback techniques [30]. The strategy seeks keywords typed by the client in a substantial accumulation of digitized typewritten historical archives. The strategy depends on: (i) picture preprocessing for picture binarization and improvement, noisy border and frame removal, orientation and skew correction; (ii) production of manufactured picture

words from keywords typed by the client; (iii) word segmentation utilizing dynamic parameters; (iv) proficient feature extraction for every picture word and (v) a recovery methodology that is enhanced by client's feedback.

Drira [2006] suggested a typology for various types of degradation of old text pictures. His typology is based on the type of image processing undertaken in course of virtual restoration and is made according to the future treatments that will be applied to restore the document to its original state. He also gave a restoration method treating specific document degradation: "ink bleed-through". This approach combines both Principal Component Analysis (PCA) and K-means. These procedures are applied recursively to separate original text from interfering and overlapping areas of text [31].

Silva et al. [2006] consider the histogram distribution as the 256-symbol source (a priori source) distribution [4]. Further, assumed the hypothesis that all symbols are statistically independent. They exhibited a statistical conform, utilizing the normalized entropy, between the distributions of the gray-scale and the black-and-white adaptations of the archive picture. Initially, the entropy E of the gray-scale picture histogram is computed as:

$$E = -\sum_{a=0}^{2} 55 p_a log(p_a) \tag{2.19}$$

Where, $p_0, p_1, \ldots, p_255$ is the priori probability distribution provided by Equation 2. The posteriori probability distribution $P_t, 1 - P_b$ is calculated by scanning b levels for each value of b. Where, $Pb0.5$ is the entropy associated with that distribution:

$$E'(b) = e(p_b) \tag{2.20}$$

Here, $e(p) = -plog_2(p) - (1-p)log_2(1-p)$ is the entropy function (pal and choudhary) and Pb is given by Equation 3. At long last, one decides the optimal limit that minimizes the $|e(b)|$ esteem given by:

$$|e(b)| = |\frac{E'}{\frac{E}{log(256)}} - L(\frac{E}{log(256)})| \tag{2.21}$$

Where, L is a loss factor, tentatively decided, given by:

$$L(\frac{E}{log(256)}) = \begin{cases} \frac{-3}{7}\frac{E}{log(256)} + 0.8, & \text{when } \frac{E}{log(256)} < 0.7 \\ \frac{E}{log(256)} - 0.2, & \text{when } \frac{E}{log(256)} \geq 0.7 \end{cases} \tag{2.22}$$

Makridis et al. [2007] proposed a new approach for word segmentation in historical and degraded machine-printed documents. The major difficulties with their algorithms were having different sizes of text, having text and non-text areas lying close to each other and having non-linear and warped text lines [33]. It depends on:

- a dynamic run length smoothing calculation that helps gathering together homogeneous content regions,
- removal of noise and punctuation marks,
- detection of obstacle to simplify the segmentation process and
- draft content line estimation process that guides the final word segmentation result.

This methodology performs better compared to previous word segmentation methods for historical and degraded machine-printed artefacts.

Martin and Thonnat [2007] presented a way to deal with perform undertaking focused segmentation based on segmentation calculation parameter tuning and learning strategies. They proposed a plan that, for every segmentation calculation to test, first concentrates ideal parameters and second takes in the region labeling as indicated by the segmentation task. This administered approach utilizes two sorts of ground-truth information: manual region based segmentation and semantic region labels. The initial step comprises in separating ideal segmentation calculation parameters by utilizing a closed loop enhancement method, an assessment metric and ground-truth (manual region segmentation). During the second step, region classifiers are prepared taking into account ground-truth explanations (semantic region labels) to permit segmentation labeling. This information (i.e. ideal parameters and learned region classifiers) is then used to produce an upgraded class-based segmentation of new pictures [34].

Zahour et al. [2007] presented a content line segmentation strategy for printed or handwritten historical Arabic reports. Archives are initially characterized into two classes utilizing a K-means scheme. These classes correspond to archive complexity (simple or difficult to segment). At that point, an archive which incorporates overlapping and touching characters, is partitioned into vertical strips. The extracted content blocks acquired by horizontal projection are characterized into three classes: little, normal and large content blocks. Subsequent to segmenting the large content blocks, the lines are obtained by matching adjacent blocks inside two progressive strips utilizing spatial relationship. The archive without covering

or touching characters is segmented by making abstraction on the segmentation module of the large content blocks [35].

Vamvakas et al. [2008] gave a complete OCR strategy for perceiving authentic reports (printed and written by hand) with no earlier information of text style. This approach comprises of three stages: The initial two stages allude to making a database for training utilizing an arrangement of reports, while the third one alludes to recognition of new archive pictures. Initial, a pre-preparing step that incorporates picture binarization and enhancement happens. At the second step, a top-down segmentation methodology is utilized in order to recognize content lines, words and characters. A clustering scheme is then adopted keeping in mind the end goal to gathering characters of similar shape. At that point, a database is made in order to be utilized for recognition. At last, for each new archive picture the above segmentation approach happens while the recognition depends on the character database that has been created at previous stage. It is a self-loader technique since the client can communicate at any time to correct conceivable mistakes of clustering and assign an ASCII label [36].

Alves et al. [2008] proposed a method for handwritten digit segmentation in recorded archive pictures. It depends on one-class classifiers, which are utilized to recognize confined characters from touching characters. The technique does not require negative information in the training stage. Three techniques for feature extraction and five one class classifiers are considered and have their performance compared. Exploratory results on an data set of handwritten digits separated from an accumulation of historical reports demonstrate the effectiveness of the proposed technique [37].

Yosef et al. [2009] gave another methodology for content line segmentation in view of versatile local projection proles for corrupted archives with content lines written in extensive skew. This methodology is quick and it applies the local calculation in an incremental way which adjusts to the skew of every content line as it advances. It gives profoundly precise results for degraded records with lines written in various skew angles and curvatures [38].

Gatos and Pratikakis [2009] proposed another productive word spotting technique that can be applied to historical printed records without requiring any previous block or word segmentation step. They tended to an approach which is segmentation-free since in many cases of authentic archives, the segmentation procedure does not deliver important results because of unconstraint design, a few corruptions or typesetting flaws. The technique depends on block based record picture descriptors that are utilized at a format coordinating procedure fulfilling

invariance as far as translation, rotation and scaling. Improvement in terms of time expense is obtained by applying the coordinating procedure just on salient regions of the picture [39].

NICK technique [2009] is an enhancement of Niblack's strategy. It plans to take care of the issue of black noise in Niblack's technique and low contrast issue in Sauvola's strategy by moving the thresholding esteem downward [40]. The binarization equation is:

$$T = M + C\sqrt{\frac{\sum P_t^2 - M}{X}} \tag{2.23}$$

Where C is a variable in the reach [-0.2, - 0.1], Pi is the gray level estimation of the pixel, and the aggregate number of pixels is X.

Mello [2010] has suggested another way to deal with section pictures of historical reports. These documents are difficult to segment due to their natural characteristic of paper degradation, ink fading, gaps in text etc. His concept was to decrease the forefront information (the ink) by recreating the data we see when we go far from the record picture. As we remain back, the content has a tendency to vanish. Just the main colors from the foundation remain. This strategy is very proficient in records with different sorts of degradation despite the fact that it is not appropriate for little noises. For stained paper, his strategy accomplished preferable results over 24 established thresholding algorithms (counting histogram-based, entropy-based and versatile calculations) [41].

Silva et al. [2010] presented another procedure to expel such noise in color archives which makes utilization of neural classifiers to assess the level of intensity of the obstruction what's more that to show the presence of blur. Such classifier permits tuning the parameters of a calculation for back-to-front impedance and record upgrade [42].

Tian et al. [2011] suggested that the Back-to-Front noises may influence the feature extraction and arrangement of content when utilizing ORC to distinguish. They exploit dynamic threshold technique to handle the vision archive picture of Back-to-Front noise with unimodal or bimodal histogram characteristic [43].

Sangsawad et al. [2011] presented another methodology for segmenting text lines on Thai handwritten historical reports. The procedure depends on an Adaptive Local Connectivity Map idea utilizing Piece-Wise Separating Lines. The technique tackle issues in transcribed archives, for example, fluctuating content lines.

In addition, local maxima projection profile is utilized for upgrading the speed of extraction. The method comprises of four stages. Firstly, Otsu calculation is utilized to binarize the source picture. Second, Piece-Wise Separating Lines is applied to infer the Adaptive Local Connectivity Map to show mask content lines. In the third step, local maxima projection profile is utilized as a rule for extracting content lines. At long last, contour calculation is utilized to distinguish the intrigued mask content line. The interested mask text is utilized to map with content picture keeping in mind the end goal to remove the content lines [44].

Rao et al. [2011] described that a gray-scale picture of the noisy artefacts is commonly denoted as $J(i, j)$.

$$J(i, j) = R, R \in [0, 1] \tag{2.24}$$

Where i and j are the horizontal and vertical coordinates of the picture $J(i, j)$ and R can take any value somewhere around 0 and 1 where $R = 1$ stands for white and $R = 0$ remains for black. There are two sections in the proposed Modified IGT. In the initial segment the level shifting of the pixels of a picture is assessed, while the second part of the calculation, decides the relative significance of pixels as for item data [45]. After every emphasis, some measure of pixels will be moved from fuzzy region to foundation. The emphasis procedure will proceed as long as the accompanying standard is fulfilled is communicated by the condition given underneath

$$|Th_i - Th_{i-1}| = t_p \tag{2.25}$$

Where, Th_i is the threshold used in the i^{th} iteration, Th_{i-1} is the threshold before the i^{th} iteration and t_p is used as a sensitivity parameter of threshold.

Clausner et al. [2012] exhibited a hybrid content line segmentation technique that uses a novel data structure and a rule base to join the qualities of top-down and bottom-up methodologies while minimizing their shortcomings. The strategy utilizes a combination of rule based gathering of associated parts (bottom-up) and projection profile examination (top-down). The strategy works with bitonal pictures. Both input and output are represented to utilizing the PAGE design. Content regions, depicted by their diagrams, are populated with identified content lines. They took after the strides , Connected part examination, Rule-based gathering of associated segments to content line candidates, Splitting of substantial segments in under segmenting lines utilizing local projection profile and repeat, Merging little line contender to their closest neighbour, Creating last content lines taking into account the candidates [46].

Fernandez and Terrades [2012] evaluated the utilization of Relative Location Features (RLF) on an authentic archive segmentation task, and think about the nature of the outcomes acquired on organized and unstructured reports utilizing RLF and not utilizing them. They demonstrated that utilizing these features enhance the final segmentation on records with a solid structure, while their application on unstructured archives does not indicate huge change [47].

Garz et al. [2012] proposed a novel binarization free line segmentation technique that is robust to noise and adapts to covering and touching content lines. To start with, interest focuses representing parts of characters are separated from gray-scale pictures. Next, word clusters are recognized in high density regions and touching segments, for example, ascenders and descenders are isolated utilizing seam carving. At last, text lines are produced by connecting neighboring word clusters, where neighborhood is characterized by the common introduction of the words in the archive [48].

Gatos et al. [2014] gave a text zone detection and text line segmentation technique for historical archives in order to achieve accurate text recognition performance. They confronted a few difficulties, for example, horizontal and vertical standard lines covering with the content, two column reports and characters of various text lines touching vertically. Powerful and effective page division is important for historical handwritten document images. For text zone detection, they analysed vertical rule lines, associated parts and in addition vertical white runs while for content line division. They upgraded a current methodology based on Hough transform keeping in mind the end goal to better treat instances of vertical connected characters. Their strategy demonstrated gave promising results after an assessment of a set of historical handwritten documents [50].

Kale et al. [2015] gave a hybrid binarization way to enhance the quality for the old reports utilizing a combination of global and local thresholding strategies. At first, global thresholding is applied to the entire picture. The picture areas that in any case have background noise are identified and the technique is again re-applied to each area separately. This accomplishes a superior versatility for the algorithm where various types of noise re-exist in various regions of same picture. Global thresholding avoids the computational and time cost of applying a local thresholding in the whole picture which is main advantage of utilizing this strategy. Consequently it is effective in removing background noise and enhancing the quality of degraded images [51].

From this vast study of literature survey it is found that many different types of algorithm were given by many researchers to eliminate back-to-front type of in-

terference in the historical stained artifacts. But there are implementation gaps and none of the algorithm produces satisfactorily results for all kinds of historical documents. Hence, it is required to fill these gaps. Therefore, a study has been proposed where efficient algorithms were developed and implemented which can remove back to front interference in the historical stained documents.

Chapter 3

Problem Formulation

Historical documents often suffer from various types of problems due to environmental and human factors. These problems need to be removed for recognition by OCR. Various algorithms that have been implemented to remove these problems have been discussed in chapter 2. An attempt to describe the problem of back-to-front interference and the gaps in its implementation by existing algorithms has been made in this chapter.

3.1 Problem Definition

Historical artifacts are a significant source of knowledge about our history and civilization. There are many historical artifacts like documents, pictures, maps etc. which are preserved in museums and libraries all over the world. But these documents get degraded over time due to various environmental factors like climatic conditions of humidity and because the chemicals used in the composition of the paper used for writing these documents. Hence, to make this precious source of information accessible to students, teachers and researchers all over the world and avail it to the future generations, we need to digitize them. Digitization allows these documents to be available over the Internet so that they can be studied around the world. OCR scans a document and applies various preprocessing and segmentation algorithms to extract the features of the text on the documents for accurate recognition and provides desirable output that contains clearly visible text that is in editable electronic format. But the process of digitization may introduce some noise along with the existing noises in the paper and text due to degradation with time. There are several different types of problems seen in historical documents like smudges, smears, blemishes, bleed through, gaps in text due to environmental damage and strokes of ink, dirt due to man handling. Variance in illumination and warping of text may occur during digitization. “Back-to-front

interference" or "bleed-through" or "show-through" is one of the most common problems faced by historical documents which occurs when the document is written on both sides on a translucent paper and the ink from the back side of the paper seeps through and shows up on the front side of paper causing difficulty in recognition of text by OCR. Several developments have been made in this field and various algorithms have been developed by different researchers to remove "back-to-front interference" but no single algorithm provides satisfactory results for all types of historical documents. Therefore, there are implementation gaps in the previous algorithms. We need to find the gaps, analyse them to provide a feasible solution and compare it with the existing algorithms.

3.2 Gap Analysis and Objective

The study of various types of historical documents and the problem of bleeding noise is the evidence that there are gaps in the implementation of algorithms to remove bleed through in historical documents. Many algorithms were given by many authors, but none of them provide satisfactorily results. There are lots of thresholding methods implemented in the literature for various types of thresholding problems. The few simple available binarization techniques cannot be applied to many thresholding problems. The main objectives of this thesis are:

- To develop a new and efficient algorithm to remove the back-to-front interference in the historical stained documents.
- To test the algorithm on historical stained documents affected from back-to-front interference.
- To compare the proposed algorithm with other best known algorithm. The next chapter includes the development and implementation of the algorithms which can remove back-to-front interference in the historical documents which are not in good condition.

Chapter 4

Development and Implementation of Proposed Algorithm

The back-to-front interference noise in the historical document is a big matter of concern. The documents affected from this noise are losing their content day by day. It is very necessary to preserve such type of documents and implement an algorithm which can help to preserve such invaluable historical documents. Many authors worked on this problem but none of them provide the satisfactorily solution. In the present work two new and efficient algorithms are developed which can effectively remove back-to-front interference in the historical stained documents.

4.1 Preprocessing

The processing phase is used to filter the image such that any noise due to digitization or environmental factors is removed. This makes it easier to attain suitable results while segmenting the image.

4.1.1 Normalization

The original grayscale image is first normalized. Normalization corrects the variance in intensity or to reduce non-uniform illumination. Normalization is sometimes also called as "histogram stretching".

Algorithm for Normalization:

1. Let minimum intensity of the original grayscale image be (Omin).
2. Let maximum intensity of the original grayscale image be (Omax).

3. Find the range of intensity for the grayscale image.
4. Assign desired minimum intensity (Dmin) as 0.
5. Assign desired maximum intensity (Dmax) as 1.
6. Find the desired range of intensity.
7. Subtract Omin from original grayscale image.
8. Multiply the above difference with desired range.
9. Divide the above product with the original range.

4.1.2 Morphological Operations

Morphological Operations are used in image processing to extract image components such as shape and boundaries. There are two basic operations: dilation and erosion. Dilation thickens an image while erosion thins an image. Morphological closing operations were used which is dilation followed by erosion. It tends to smooth the contours of an object by filling and joining any holes or breaks.

4.1.3 Filtering

Filtering is required as the images captured by us are not fit to be provided to the OCR. Any kind of variance in intensity or illumination and poor contrast must be removed. There are various types of filtering techniques which can be used to remove noise such as liner filtering, median filtering or adaptive filtering. Since, adaptive filtering produces best results as it more selective and preserves the edges and other high-frequency parts of the image, it is best suited for historical documents. An effort has been used to apply the same for removing noise.

4.2 Proposed Algorithms

Algorithm-1
In order to improve the quality of historical documents images, a combined approach based on mathematical morphology, global and local thresholding methods is applied. This method at initial step removes the background (noise) and adjusts the intensity values of pixels of the historical image, which enhances the performance of global thresholding. After that, the mathematical erosion operation is applied on this binarize image, then local binarization is applied to binarize the image.

Pseudo-code:

Input: A grayscale image

1. Background removal is performed by using median filter.

2. Adjust image intensity values (increasing contrast) of image resulted from step 1.
   $J = image_{adjust}(I)$.

3. A global threshold (T) applied to this image (J).
   Where $T = average(J)$,
   $IfJ(m,n) > T : Out(m,n) = 255$
   $ElseOut(m,n) = 0$
   End

4. Apply morphological erosion operation by using a disk of radius 1 on the binary image (Out), we get after global threshold.
   $Im_{morph} = image_{erosion}(Out, disk(1))$.

5. Add images J and Im_{morph}.
   $Im_{add} = J + Im_{morph}$.

6. Apply adaptive threshold method on the intermediate image Im_{add}.

Algorithm-2

This method at initial step removes the background (noise) by using the distance perception and after that adjusts the intensity values of pixels of the historical image, which enhances the performance of global thresholding. After that, the mathematical erosion operation is applied on this binarize image, then local binarization is applied to binarize the image.

Pseudo-code:

Input: A grayscale image

1. Background removal is performed by using distance perception.
   $Img = inputgrayscaleimage$;
   $[p,q] = resolution(Img)$;
   $Img_{down} = image_{resize}(Img, [p * 0.10, q * 0.10]$; *decrease the resolution of image by 10% of row and column.*
   $Img_{normal} = image_{resize}(Img_{down}, [p,q])$; *restore the resolution of image.*
   $Img_{sub} = Img_{normal}–Img$;
   $Img_{comp} = imagecompliment(Img_{sub})$;

2. Enhance the contrast of this picture by extending the scope of intensity values it contains to traverse a desired scope of values. Let $L = 255$ and $H = 0$;
   $ForI = starting_{pixel}$ to $ending_{pixel}$.
   $If(L > pixel[I]_{intensity})$
   $L = pixel[I]_{intensity}$;
   $ElseIf(H < pixel[I]_{intensity})$
   $H = pixel[I]_{intensity}$;
   End
   At that point every pixel P is scaled utilizing the following function:
   $ForI = starting_{pixel}$ to $ending_{pixel}$.

   $$Pixel[I]_{out} = (Pixel[I]_{in} - L)(\frac{255}{(H-L)})$$

   Let the image after increasing contrast be $Img_{adjusted}$.

3. A global threshold (T) applied to this image $(Img_{adjusted})$.
   Where $T = average(Img_{adjusted})$,
   $ForI = starting_{pixel}$ to $ending_{pixel}$.
   $If(pixel[I]_{value} > T)$
   $Pixel[I]_{value} = 255$;
   $Else\ Pixel[I]_{value} = 0$;
   End

4. Apply morphological erosion operation by using a disk of radius 1 on this binary image, we get after global threshold.

   $$Img_{morph} = image_{erosion}(image, disk(1)).$$

5. Add images $Img_{adjusted}$ and Img_{morph}.
   $Img_{add} = Img_{adjusted} + Img_{morph}$.

6. Apply adaptive threshold method on the intermediate image Img_{add}.

The proposed algorithms are applied to many historical stained documents and the algorithms provides better results. In the next chapter, results obtained from the proposed algorithms are shown and discussed.

Chapter 5

Results and Comparison

The proposed algorithms are implemented in MATLAB R2014a version and applied to various historical stained documents affected from the back-to-front interference noise. The proposed algorithms yields the best results as compared to the others works discussed in the literatures. Two well-known assessment measurements are utilized for comparing performance of proposed procedure and already existing procedures for text segmentations.One of them is called as SSIM and other as RMS.

1. SSIM: The Structural Similarity index (SSIM) is a technique used for calculating the similitude among two pictures proposed by Wang et al. [52]. SSIM is characterized as,

$$SSIM(A, B) = \frac{(2\mu_A\mu_B + C_1)(2\sigma_{AB} + C_2)}{(\mu_A^2 + \mu_B^2 + C_1)(\sigma_A^2 + \sigma_B^2 + C_2)} \tag{5.1}$$

   Where A and B implies mean intensities of two images A and B respectively, AB is the covariance of A and B. A and B are the standard deviation of image A and B respectively. C1 and C2 are constants. Here C1 = C2 = 0. The estimation of SSIM(X, Y) can be any value from interim [-1, 1]. At the point when the value is equivalent to 1 implies the pictures are totally comparable and when equivalent to - 1 implies the pictures are totally unique.

2. RMS: Root Mean Square (RMS) is a strategy designed for computing divergence among two matrices. This is also called the root mean square error (RMSE) or root mean square deviation (RMSD). In this technique, the distinction between real (unique) picture and the changed (by model) picture is ascertained.RMS is expressed as

$$RMS(A,B) = \sqrt{\frac{\sum_{i=1}^{N}(a_i - b_i)^2}{M}} \tag{5.2}$$

Where, ai and bi are pixels of images A and B individually, and M is the aggregate number of pixels. at the point when the estimation of RMS(A, B) is equivalent to zero means two images have most extreme closeness and as the quality increases divergence increments.

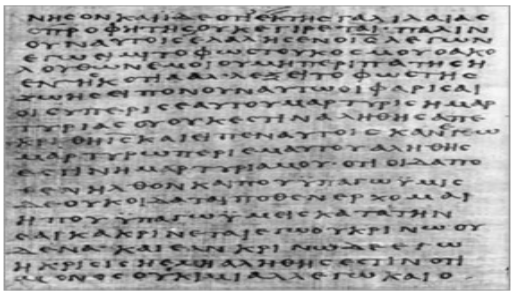
5.1 Results of Algorithm-1

Algorithm-1 was applied on number of historical documents and obtained the good results. It is not possible to show all the results here. Therefore, one historical document is taken and figure 5.1 shows the various changes done to that documents by algorithm-1 when the document passes through the various phases of algorithm.

5.1.1 Comparision of Results of Algorithm-1 with other Algorithms

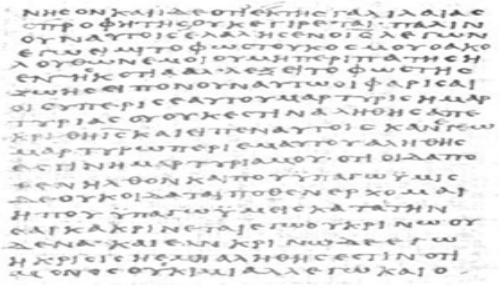
To evaluate the proposed algorithm in comparison to other algorithms the authors implemented various most common algorithm used in image segmentation, in MATLAB R2014a version and applied to many historical documents affected from the back-to- front interference. The proposed algorithm gives best result as compared to other algorithms. Out of the results obtained, one is shown in figure 5.2 to 5.9 :

To evaluate the proposed method experimentally (with other existing techniques), a clean image (without any impurity) is required. But it was very difficult to get the clean image of historical document. Finally, it was decided to add impurities to a clean image to compare results given by proposed method and results given by existing techniques. A ground truth handwritten document image is shown in figure 5.10a. Background of a degraded historical document is extracted, shown in figure 5.10b . This background is added to the ground truth image to acquire a corrupted picture as shown in figure 5.10c . Diverse strategies are applied on the corrupted picture. Resultant pictures are compared against ground truth utilizing assessment measurements RMS and SSIM as appeared in table 5.1 .

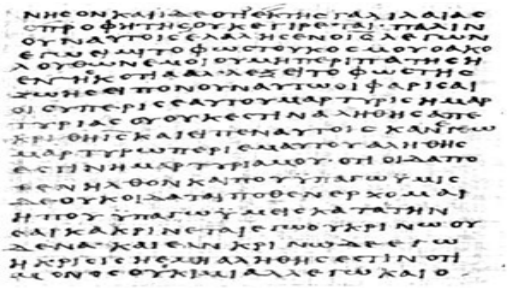


(a) Grayscale Image

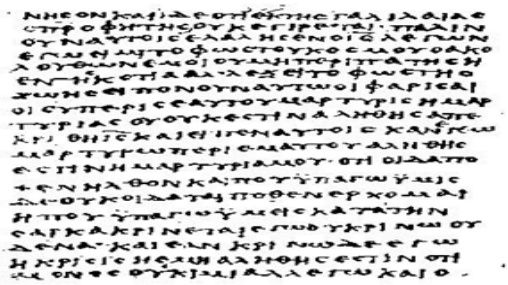
(b) Image after Applying Median Filter (background extraction)



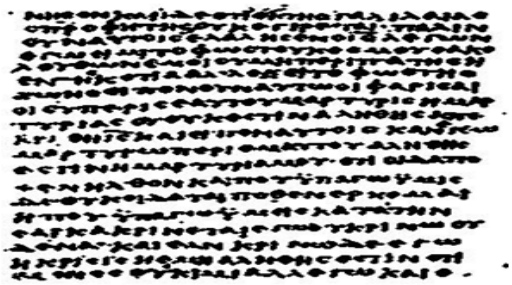
(c) Image after Removal of Background



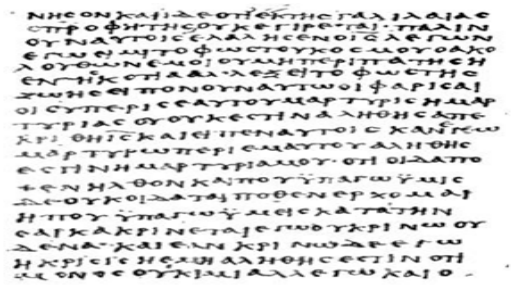
(d) Image after Adjusting Intensity of Pixels



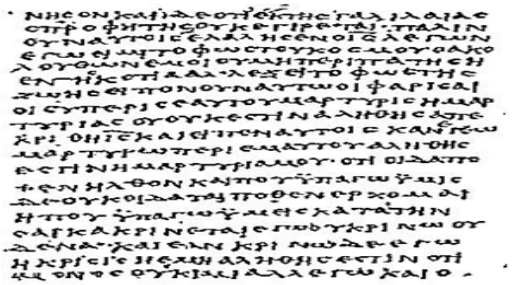
(e) Image after Applying Global Binarization



(f) Image after Erosion Operation



(g) Image after the Addition of $IntensityAdjustedImage(d) and\ ErodedImage(f)$



(h) Binarized Image using Adaptive Threshold.

Figure 5.1: Various Steps of Proposed Algorithm-1

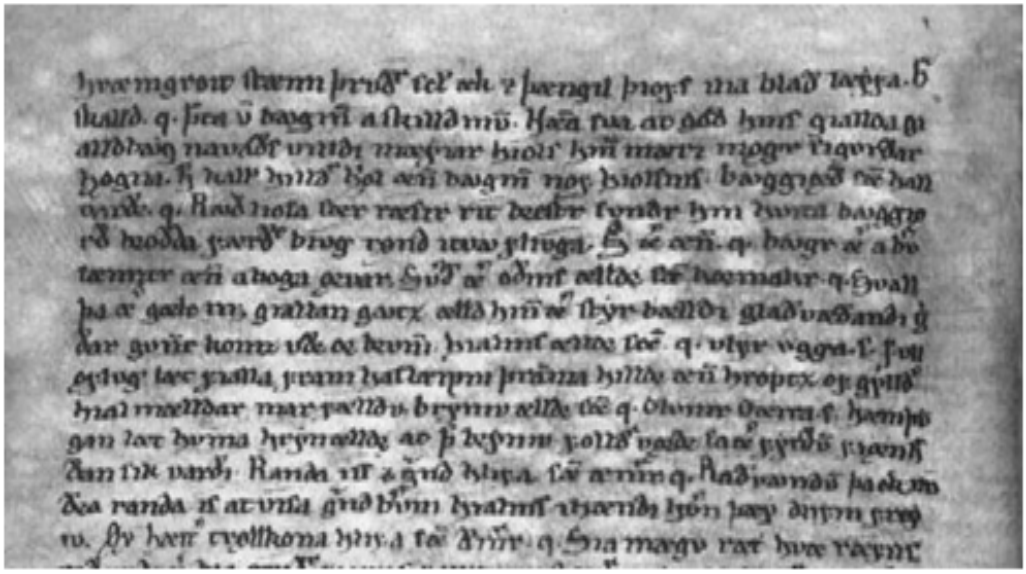


Figure 5.2: Original Historical Document

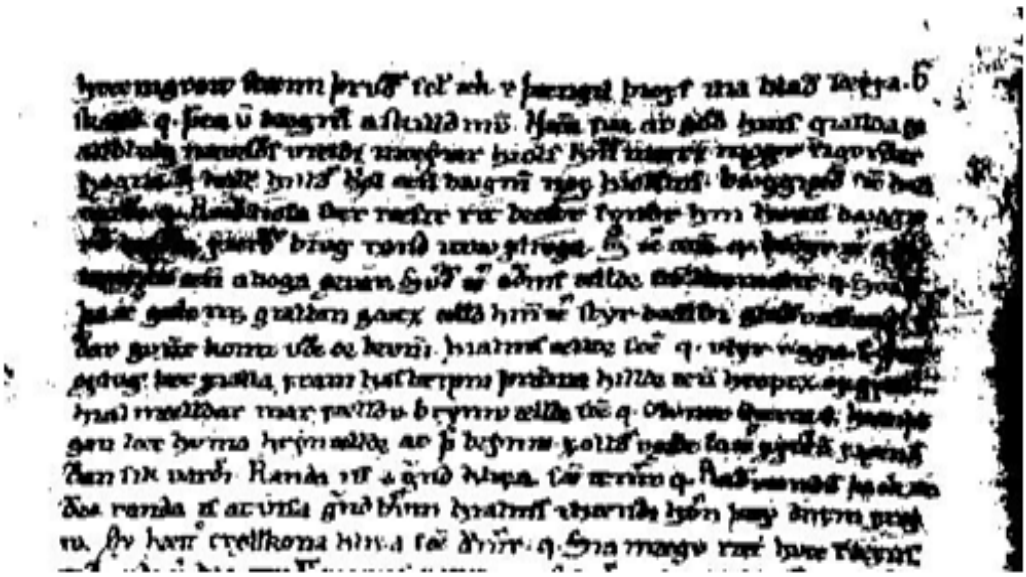


Figure 5.3: Otsu's Method

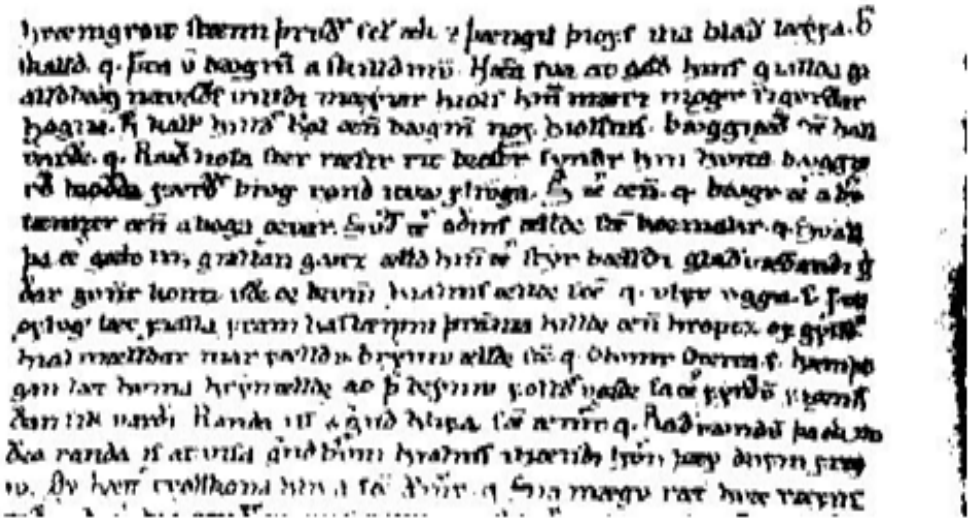


Figure 5.4: Iterative Method

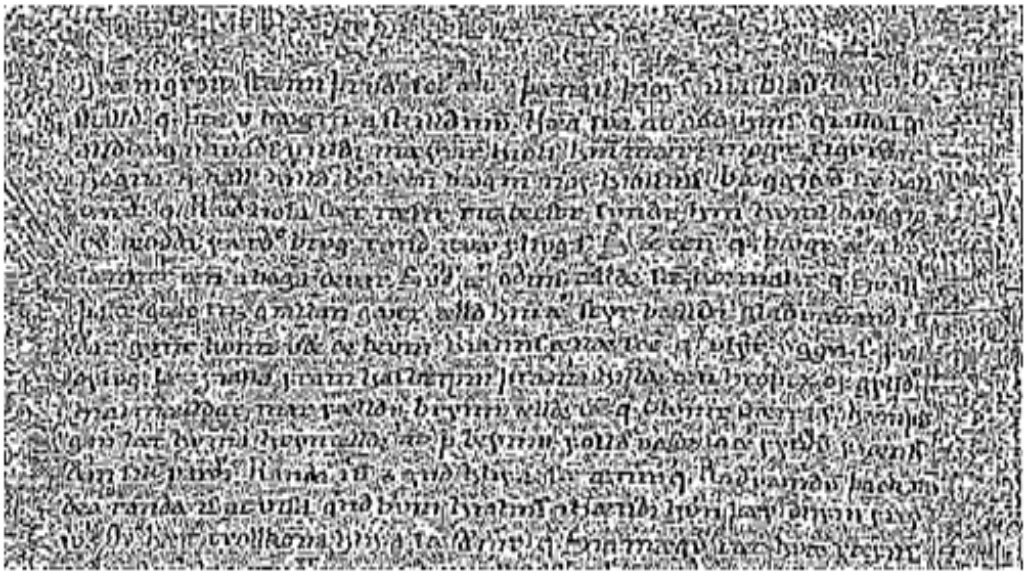


Figure 5.5: Niblack's Method

Figure 5.6: Sauvola's Method

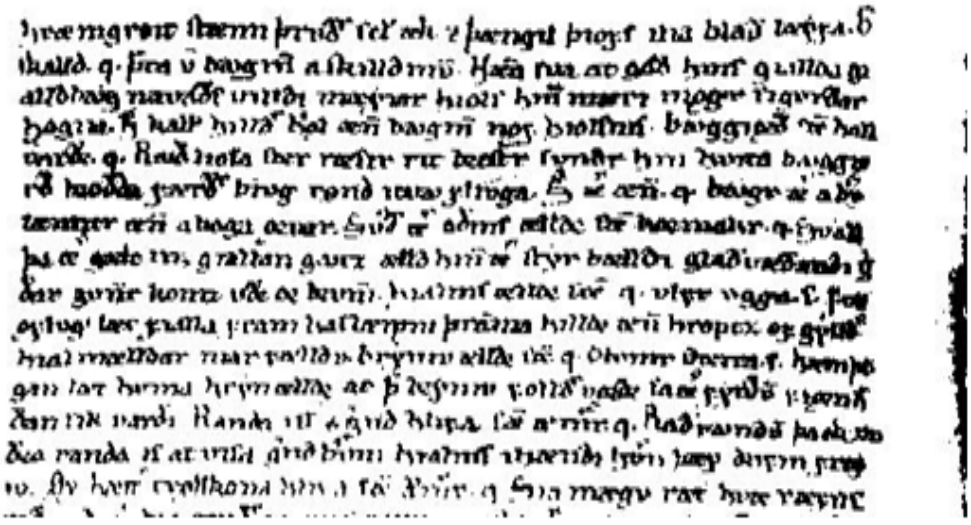


Figure 5.7: Kittler's Method

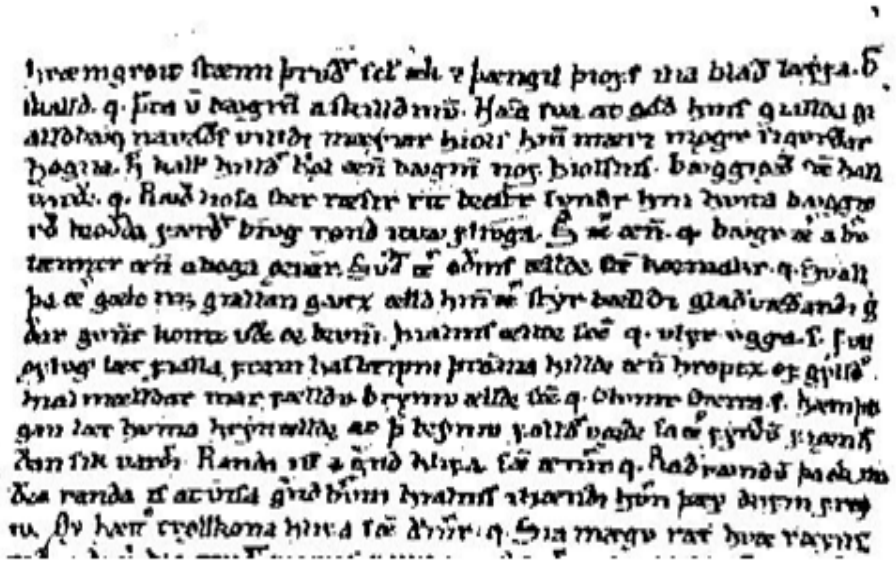


Figure 5.8: NICK's Method

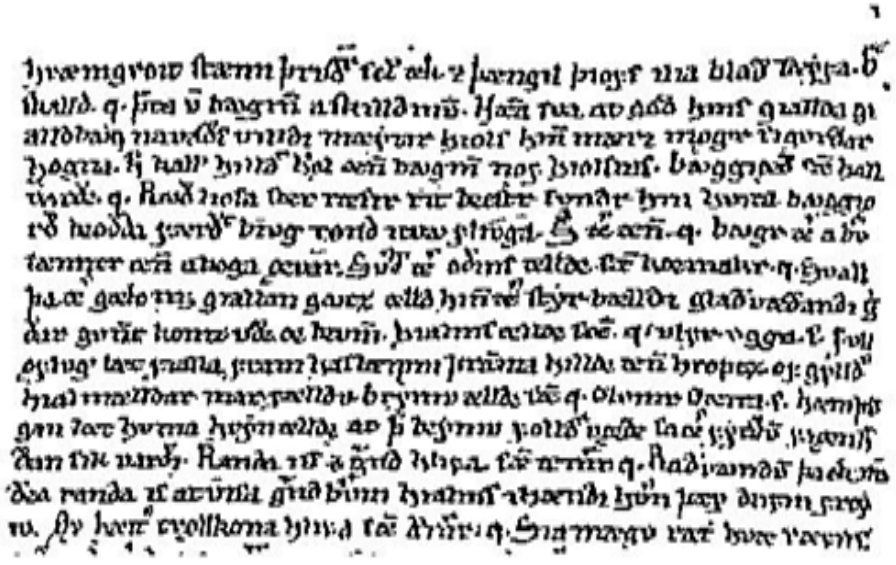
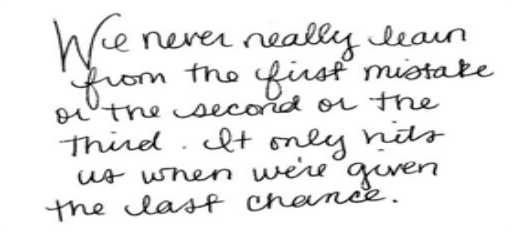
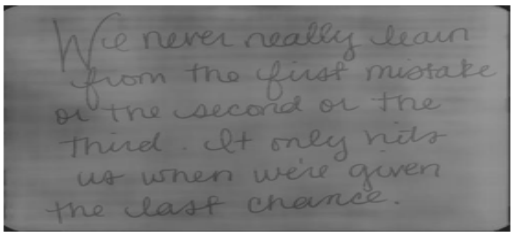


Figure 5.9: Proposed Method

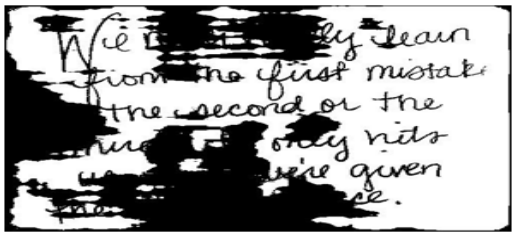


(a) Original Image

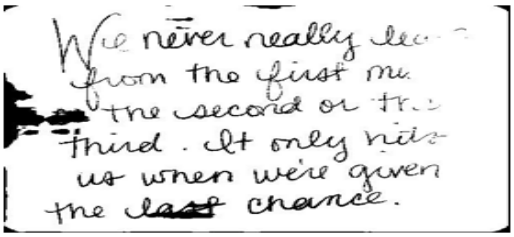
(b) Noisy Image



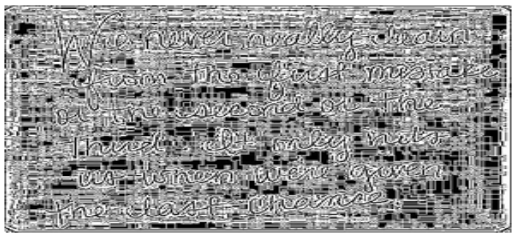
(c) Image with Noise (noise added to the original image)



(d) Otsu's Method

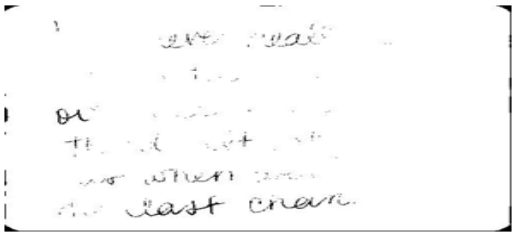


(e) Iterative Method

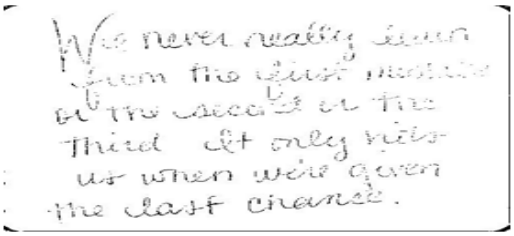


(f) Niblack's Method

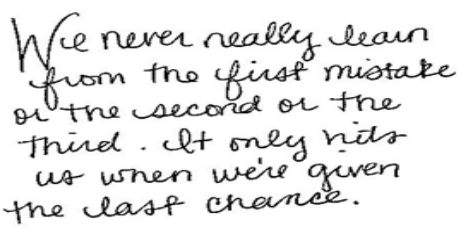
(g) Sauvola's Method



(h) Kittler's Method



(i) NICK's Method



(j) Proposed Method

Figure 5.10: Results of Various Algorithms.

Fig. 5.10d shows binarized picture utilizing otsu's global thresholding technique with adequate loss of information. Fig. 5.10e shows the result obtained by applying iterative method. This method also could not recover text. Fig. 5.10f indicates resultant picture utilizing Niblack's thresholding technique, it is totally misshaped. Fig. 5.10g shows resultant image using Sauvola's local binarization technique, which calculated threshold subject to standard deviation and mean however the extra parameter is steady everywhere throughout the picture which distorts contour. It doesn't improve textures of low intensity. Fig. 5.10h indicates consequence of applying Kittler's strategy with adequate loss of information. Fig. 5.10i shows the result of NICK method, which also could not recover text. Fig. 5.10j demonstrates the aftereffect of the proposed algorithm, which gives superior result. Similar investigation of these strategies utilizing assessment measurements is given as a part of Table 5.1.

Table 5.1: Comparative Performance of Various Thresholding Techniques.

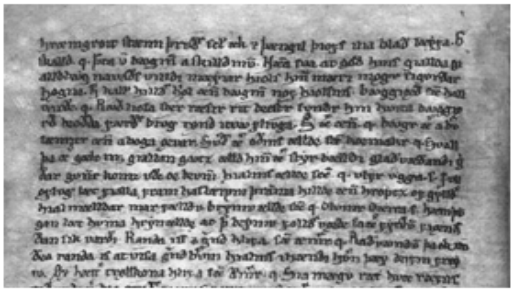
Thresholding Technique	***RMS Evaluation Metrics***	***SSIM Evaluation Metrics***
Otsu's method	2.8348	0.5025
Iterative method	4.3045	0.8376
Niblack	4.1614	0.1225
Sauvola	6.0655	0.6700
Kittler	5.8194	0.6865
NICK	5.4186	0.7738
Proposed method	2.8272	0.9296

5.2 Results of Algorithm-2

The step by step changes in one of the historical documents affected from the back-to-front noise by applying the algorithm-2 is shown in figure 5.11 :

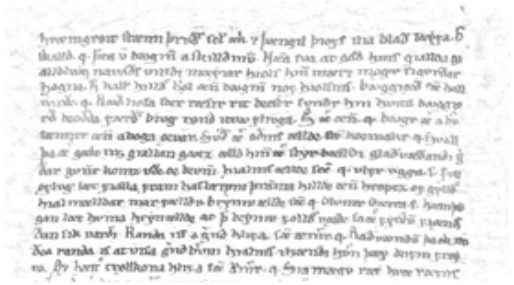
5.2.1 Comparision of Results of Algorithm-2 with other Algorithms

To evaluate the proposed algorithm in comparison to other algorithms the authors implemented various most common algorithm used in image segmentation, in MATLAB R2014a version and applied to many historical documents affected from the back-to- front interference. The proposed algorithm gives best result as

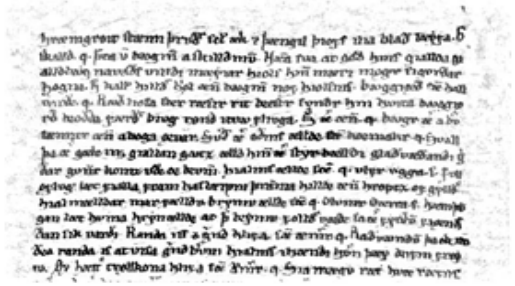


(a) Grayscale Image

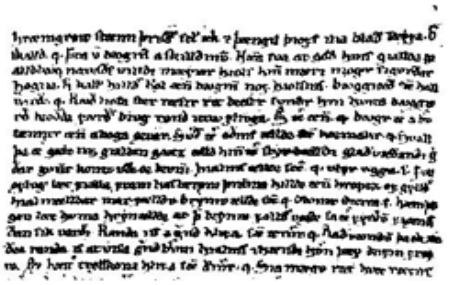
(b) Image after Applying Distance Perception (i.e. background extraction)



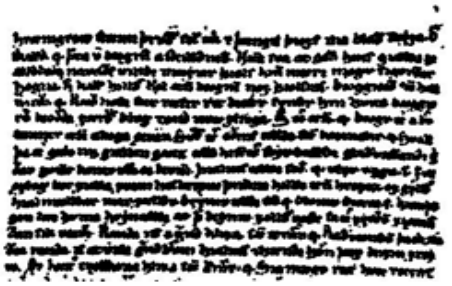
(c) Image after Removal of Background



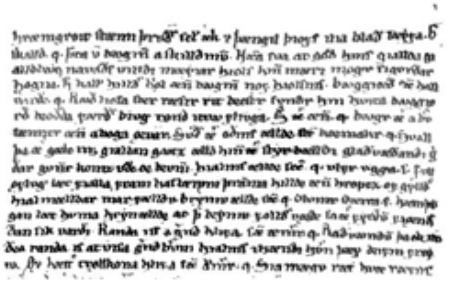
(d) Image after Improving Contrast of Pixels



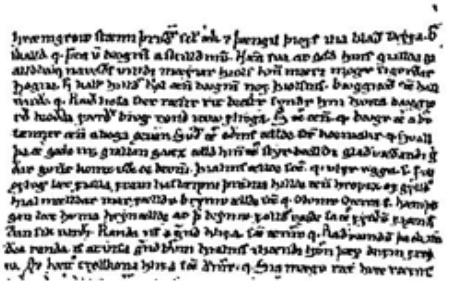
(e) Image after Applying Global Binarization



(f) Image after Erosion Operation



(g) Image after the Addition of $ContrastImprovedImage(d)$ and $ErodedImage(f)$



(h) Binarized Image Using Adaptive Threshold.

Figure 5.11: Various Steps of Proposed Algorithm-2

Figure 5.12: Original Historical Document

compared to other algorithms. Out of the results obtained, one is shown in figure 5.12 to 5.19 :

To evaluate the proposed method experimentally (with other existing techniques), a clean image (without any impurity) is required. But it was very difficult to get the clean image of historical document. Finally, it was decided to add impurities to a clean image to compare results given by proposed method and results given by existing techniques. A ground truth handwritten document image is shown in fig. 5.20a . Background of a degraded historical document is extracted, shown in fig. 5.20b . This background is added to the ground truth image to acquire a corrupted picture (Fig. 5.20c). Diverse strategies are applied on the corrupted picture. Resultant pictures are compared against ground truth utilizing assessment measurements RMS and SSIM as appeared in table 5.2 .

Fig. 5.20d shows binarized picture utilizing otsu's global thresholding technique with adequate loss of information. Fig. 5.20e shows the result obtained by applying iterative method. This method also could not recover text. Fig. 5.20f indicates resultant picture utilizing Niblack's thresholding technique, it is totally misshaped. Fig. 5.20g shows resultant image using Sauvola's local binarization technique, which calculated threshold subject to standard deviation and mean however the extra parameter is steady everywhere throughout the picture which distorts contour. It doesn't improve textures of low intensity. Fig. 5.20h indicates

Figure 5.13: Otsu's Method

Figure 5.14: Iterative Method

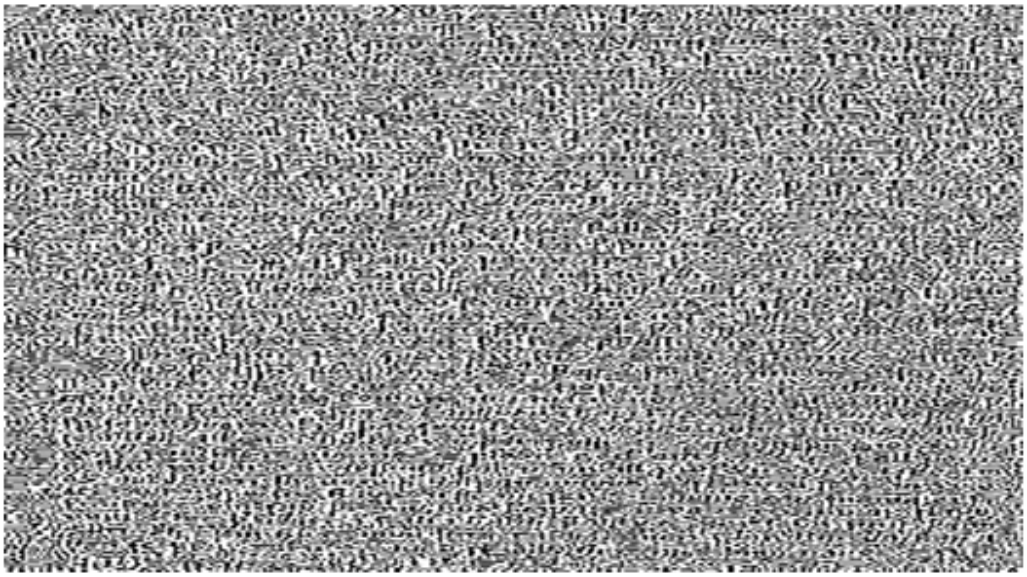


Figure 5.15: Niblack's Method

Figure 5.16: Sauvola's Method

Figure 5.17: Kittler's Method

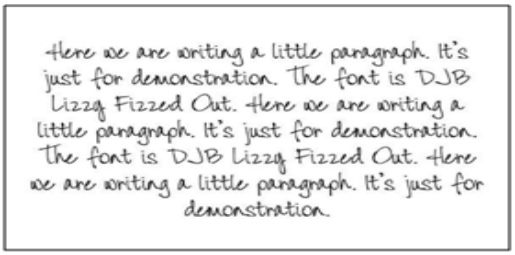
Figure 5.18: NICK's Method

Figure 5.19: Proposed Method

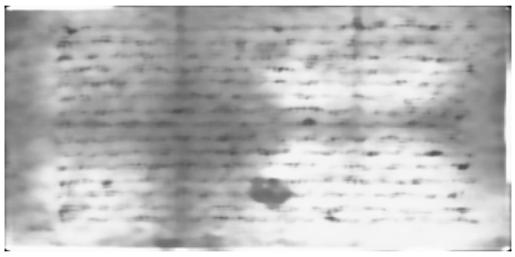
consequence of applying Kittler's strategy with adequate loss of information. Fig. 5.20i shows the result of NICK method, which also could not recover text. Fig. 5.20j demonstrates the aftereffect of the proposed algorithm, which gives superior result. Similar investigation of these strategies utilizing assessment measurements is given as a part of Table 5.2.

Table 5.2: Relative Study of Various Binarization Techniques.

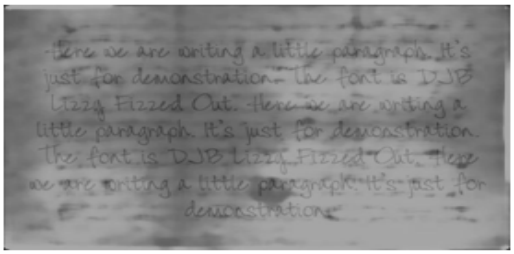
Binarization Method	***RMS Assessment Metrics***	***SSIM Assessment Metrics***
Otsu's method	1.4912	0.3187
Iterative method	2.9449	0.7216
Niblack	2.0081	0.1434
Sauvola	3.9556	0.6751
Kittler	0.9713	0.1197
NICK	3.1280	0.8064
Proposed method	0.9453	0.8568



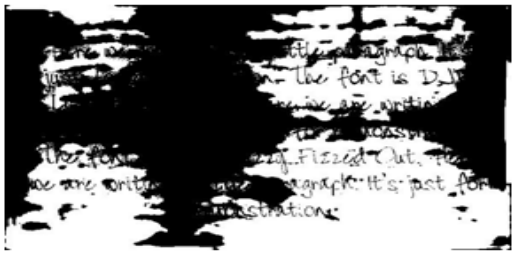
(a) Original Image



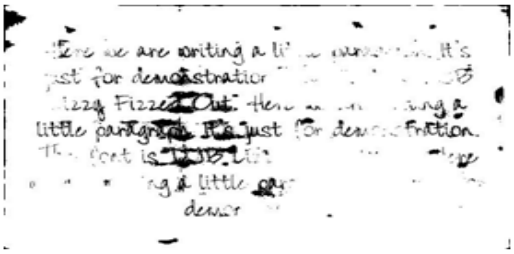
(b) Noisy Image



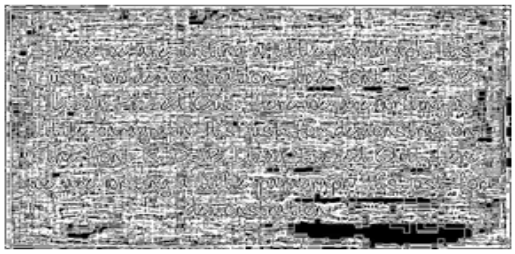
(c) Image with Noise (noise added to the original image)



(d) Otsu's Method



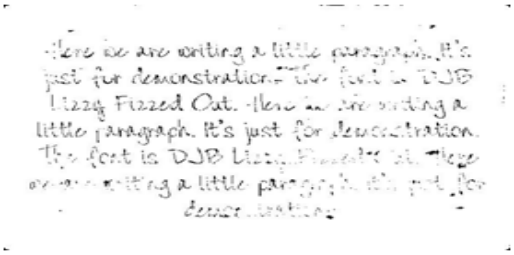
(e) Iterative Technique



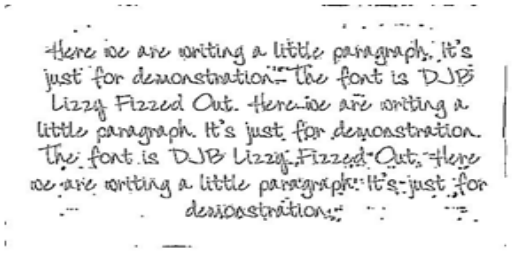
(f) Niblack's Technique

(g) Sauvola's Technique

(h) Kittler's Technique



(i) NICK's Technique



(j) Proposed Method

Figure 5.20: Results of Various Algorithms

Chapter 6

Conclusion and Future Scope

6.1 Conclusion

Historical artefacts are a great source of knowledge about the history and civilization of the past. These documents are conserved in libraries and museums. But keeping them in libraries allows only a few people to access it. Also, they are getting damaged due to environmental condition over time. So, to make them available to a wider audience all over the world and allow access to the future generations, a digital database should be maintained. These documents suffer from various types of degradations among which "bleed through" or "show through" or "back-to-front interference" is the most common. An attempt to solve this problem in historical document images has been made in the proposed work in chapter 4 . The present work introduces two hybrid frameworks for cleansing the historical documents. These frameworks were implemented and compared with other existing methods. As shown in Table 5.1 and Table 5.2 , the proposed methods gives highest SSIM value which being near to 1 proves that the proposed methods works well. On the other hand, RMS value is the lowest which being near to 0 prove that the proposed methods gives result which are close to the original clean image. The proposed method outperforms other existing methods.

6.2 Future Scope

The approach presented can be improvised or extended in the following ways:

1. It can be improved to remove the background (noise) according to the resolution of the image automatically.
2. It can be improved to calculate the radius of the structural element i.e. disk automatically.

References

[1] S. Srihari and S. Lam, "Character recognition, center of excellence for document analysis and recognition (cedar)," tech. rep., Technical Report, 1995.

[2] S. Panwar and N. Nain, "A novel approach of skew normalization for handwritten text lines and words," in *Signal Image Technology and Internet Based Systems (SITIS), 2012 Eighth International Conference on*, pp. 296–299, IEEE, 2012.

[3] P. Slavik and V. Govindaraju, "Equivalence of different methods for slant and skew corrections in word recognition applications," *IEEE transactions on pattern analysis and machine intelligence*, vol. 23, no. 3, pp. 323–326, 2001.

[4] J. M. M. da Silva, R. D. Lins, and V. C. Da Rocha, "Binarizing and filtering historical documents with back-to-front interference," in *Proceedings of the 2006 ACM symposium on Applied computing*, pp. 853–858, ACM, 2006.

[5] G. Johannsen and J. Bille, "A threshold selection method using information measures," in *ICPR*, vol. 82, pp. 140–143, 1982.

[6] C. A. Mello and R. D. Lins, "Generation of images of historical documents by composition," in *Proceedings of the 2002 ACM symposium on Document engineering*, pp. 127–133, ACM, 2002.

[7] M. Stevens, "Automatic character recognition-state-of-the-art report, national bureau of standards & technology, tech," tech. rep., Note 112, Washington, USA, 1961.

[8] S. Mori, C. Y. Suen, and K. Yamamoto, "Historical review of ocr research and development," *Proceedings of the IEEE*, vol. 80, no. 7, pp. 1029–1058, 1992.

[9] R. Plamondon and S. N. Srihari, "Online and off-line handwriting recognition: a comprehensive survey," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 22, no. 1, pp. 63–84, 2000.

[10] C. Y. Suen, R. Legault, C. Nadal, M. Cheriet, and L. Lam, "Building a new generation of handwriting recognition systems," *Pattern Recognition Letters*, vol. 14, no. 4, pp. 303–315, 1993.

[11] N. Arica and F. T. Yarman-Vural, "An overview of character recognition focused on off-line handwriting," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 31, no. 2, pp. 216–233, 2001.

[12] J. N. Kapur, P. K. Sahoo, and A. K. Wong, "A new method for gray-level picture thresholding using the entropy of the histogram," *Computer vision, graphics, and image processing*, vol. 29, no. 3, pp. 273–285, 1985.

[13] M. Cannon, J. Hochberg, and P. Kelly, "Quality assessment and restoration of typewritten document images," *International Journal on Document Analysis and Recognition*, vol. 2, no. 2-3, pp. 80–89, 1999.

[14] H. Baird, "State of the art of document image degradation modelling, invited talk," in *IAPR 2000 Workshop on Document Analysis Systems, Brazil*, 2000.

[15] R. Kasturi, L. O'gorman, and V. Govindaraju, "Document image analysis: A primer," *Sadhana*, vol. 27, no. 1, pp. 3–22, 2002.

[16] N. Otsu, "Thresholds selection method form grey-level histograms," *IEEE Trans. On Systems, Man and Cybernetics*, vol. 9, no. 1, p. 1979, 1979.

[17] N. Abramson, "Information theory and coding," 1963.

[18] W. Niblack, *An introduction to digital image processing*. Strandberg Publishing Company, 1985.

[19] J.-C. Yen, F.-J. Chang, and S. Chang, "A new criterion for automatic multilevel thresholding," *IEEE Transactions on Image Processing*, vol. 4, no. 3, pp. 370–378, 1995.

[20] R. G. Casey and E. Lecolinet, "A survey of methods and strategies in character segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 18, no. 7, pp. 690–706, 1996.

[21] T. M. Ha and H. Bunke, "Off-line, handwritten numeral recognition by perturbation method," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 5, pp. 535–539, 1997.

[22] S. Mo and J. Mathews, "Adaptive, quadratic preprocessing of document images for binarization," *IEEE transactions on image processing*, vol. 7, no. 7, pp. 992–999, 1998.

[23] L. Wu, S. Ma, and H. Lu, "An effective entropic thresholding for ultrasonic images," in *International Conference on Pattern Recognition*, vol. 14, pp. 1552–1554, Citeseer, 1998.

[24] J. Cai and Z.-Q. Liu, "Integration of structural and statistical information for unconstrained handwritten numeral recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, no. 3, pp. 263–270, 1999.

[25] C. A. Mello and R. D. Lins, "Image segmentation of historical documents," *Visual2000, Mexico City, Mexico*, vol. 30, 2000.

[26] J. Sauvola and M. Pietikäinen, "Adaptive document image binarization," *Pattern recognition*, vol. 33, no. 2, pp. 225–236, 2000.

[27] M. Feldbach and K. Tonnies, "Word segmentation of handwritten dates in historical documents by combining semantic a-priori-knowledge with local features," in *Document Analysis and Recognition, 2003. Proceedings. Seventh International Conference on*, pp. 333–337, IEEE, 2003.

[28] C. A. Mello, "Synthesis of images of historical documents for web visualization," in *Multimedia Modelling Conference, 2004. Proceedings. 10th International*, pp. 220–226, IEEE, 2004.

[29] S. Feng and R. Manmatha, "Classification models for historical manuscript recognition," in *Eighth International Conference on Document Analysis and Recognition (ICDAR'05)*, pp. 528–532, IEEE, 2005.

[30] B. Gatos, T. Konidaris, K. Ntzios, I. Pratikakis, and S. J. Perantonis, "A segmentation-free approach for keyword search in historical typewritten documents," in *Eighth International Conference on Document Analysis and Recognition (ICDAR'05)*, pp. 54–58, IEEE, 2005.

[31] F. Drira, "Towards restoring historic documents degraded over time," in *Second International Conference on Document Image Analysis for Libraries (DIAL'06)*, pp. 8–pp, IEEE, 2006.

[32] R. Lins and J. da Silva, "Assessing algorithms to remove back-to-front interference in documents," *ITS-2006*, 2006.

[33] M. Makridis, N. Nikolaou, and B. Gatos, "An efficient word segmentation technique for historical and degraded machine-printed documents," in *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, vol. 1, pp. 178–182, IEEE, 2007.

[34] V. Martin and M. Thonnat, "A cognitive vision approach for image segmentation thresholding images of historical documents with back-to-front interference," in *19th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2007)*, vol. 1, pp. 480–487, IEEE, 2007.

[35] A. Zahour, L. Likforman-Sulem, W. Boussellaa, and B. Taconet, "Text line segmentation of historical arabic documents," in *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, vol. 1, pp. 138–142, IEEE, 2007.

[36] G. Vamvakas, B. Gatos, N. Stamatopoulos, and S. J. Perantonis, "A complete optical character recognition methodology for historical documents," in *Document Analysis Systems, 2008. DAS'08. The Eighth IAPR International Workshop on*, pp. 525–532, IEEE, 2008.

[37] V. Alves, A. L. Oliveira, E. Silva Jr, and C. A. Mello, "Handwritten digit segmentation in images of historical documents with one-class classifiers," in *2008 20th IEEE International Conference on Tools with Artificial Intelligence*, vol. 2, pp. 41–44, IEEE, 2008.

[38] I. Bar-Yosef, N. Hagbi, K. Kedem, and I. Dinstein, "Line segmentation for degraded handwritten historical documents," in *2009 10th International Conference on Document Analysis and Recognition*, pp. 1161–1165, IEEE, 2009.

[39] B. Gatos and I. Pratikakis, "Segmentation-free word spotting in historical printed documents," in *2009 10th International Conference on Document Analysis and Recognition*, pp. 271–275, IEEE, 2009.

[40] K. Khurshid, I. Siddiqi, C. Faure, and N. Vincent, "Comparison of niblack inspired binarization methods for ancient documents," in *IS&T/SPIE Electronic Imaging*, pp. 72470U–72470U, International Society for Optics and Photonics, 2009.

[41] C. A. Mello, "Segmentation of images of stained papers based on distance perception," in *Systems Man and Cybernetics (SMC), 2010 IEEE International Conference on*, pp. 1636–1642, IEEE, 2010.

[42] G. d. F. P. e Silva, R. D. Lins, J. Silva, S. Banergee, A. Kuchibhotla, and M. Thielo, "Enhancing the filtering-out of the back-to-front interference in color documents with a neural classifier," in *Pattern Recognition (ICPR), 2010 20th International Conference on*, pp. 2415–2419, IEEE, 2010.

[43] D.-Z. Tian, C. Wang, and Z.-M. Zhang, "Dynamic threshold algorithm for removal of back-to-front noises of visual document image," in *2011 International Conference on Machine Learning and Cybernetics*, 2011.

[44] S. Sangsawad, R. Chamchong, and C. C. Fung, "Using local maxima profile and piece-wise technique for line segmentation on thai handwritten historical documents," in *Machine Learning and Cybernetics (ICMLC), 2011 International Conference on*, vol. 4, pp. 1862–1866, IEEE, 2011.

[45] N. V. R. A. S. Rao, S. Balaji, and L. P. Reddy, "Cleaning of ancient document images using modified iterative global threshold," *Proceedings of International Journal of Computer Science Issues November*, 2011.

[46] C. Clausner, A. Antonacopoulos, and S. Pletschacher, "A robust hybrid approach for text line segmentation in historical documents," in *Pattern Recognition (ICPR), 2012 21st International Conference on*, pp. 335–338, IEEE, 2012.

[47] F. C. Fernández and O. R. Terrades, "Document segmentation using relative location features," in *Pattern Recognition (ICPR), 2012 21st International Conference on*, pp. 1562–1565, IEEE, 2012.

[48] A. Garz, A. Fischer, R. Sablatnig, and H. Bunke, "Binarization-free text line segmentation for historical documents based on interest point clustering," in *Document Analysis Systems (DAS), 2012 10th IAPR International Workshop on*, pp. 95–99, IEEE, 2012.

[49] A. Fornés, X. Otazu, and J. Lladós, "Show-through cancellation and image enhancement by multiresolution contrast processing," in *2013 12th International Conference on Document Analysis and Recognition*, pp. 200–204, IEEE, 2013.

[50] B. Gatos, G. Louloudis, and N. Stamatopoulos, "Segmentation of historical handwritten documents into text zones and text lines," in *Frontiers in Handwriting Recognition (ICFHR), 2014 14th International Conference on*, pp. 464–469, IEEE, 2014.

[51] P. Kale, G. Phade, S. Gandhe, and P. A. Dhulekar, "Enhancement of old images and documents by digital image processing techniques," in *Communication, Information & Computing Technology (ICCICT), 2015 International Conference on*, pp. 1–5, IEEE, 2015.

[52] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

Publications

1. Rajiv. Kumar, Bhuvnesh. Malik, Animesh. Sharma, A Hybrid Approach to Enhance the Historical Document Images,Journal of Visual Languages and Computing.

   (Communicated)

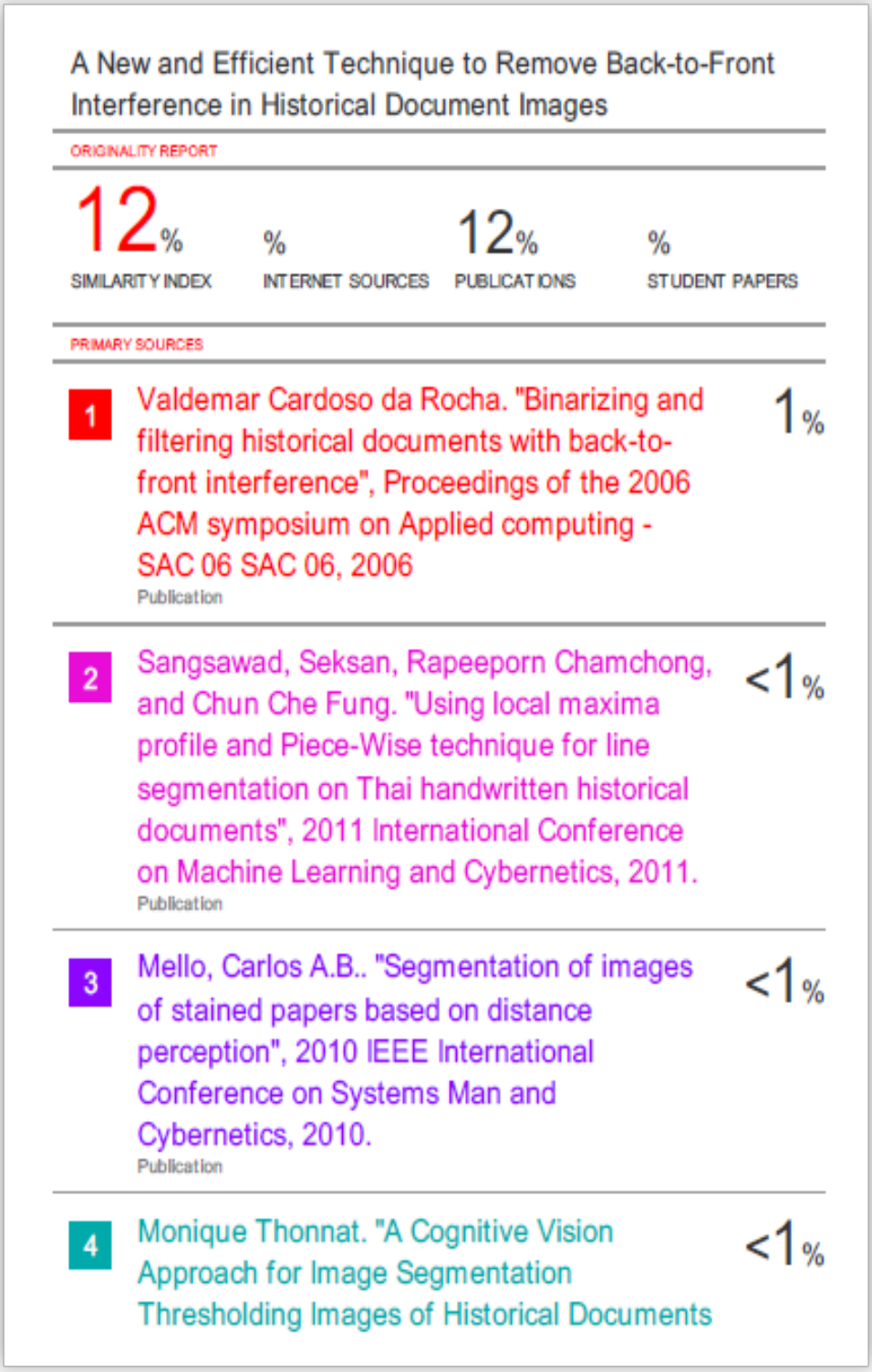
2. Bhuvnesh. Malik, Rajiv. Kumar, An Integrated Approach to Decrease Back to Front Interference in Historical Documents,KSII Transactions on Internet and Information Systems.

   (Communicated)

Video Presentation and Plagiarism Report

The video presentation on the thesis work can be found on following link:
`https://www.youtube.com/channel/UCcEHD4vsoPKgr-oK3Co3liQ`

The generated plagiarism report is attached.



A New and Efficient Technique to Remove Back-to-Front Interference in Historical Document Images

ORIGINALITY REPORT

12%	%	12%	%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Valdemar Cardoso da Rocha. "Binarizing and filtering historical documents with back-to-front interference", Proceedings of the 2006 ACM symposium on Applied computing - SAC 06 SAC 06, 2006 Publication	1%
2	Sangsawad, Seksan, Rapeeporn Chamchong, and Chun Che Fung. "Using local maxima profile and Piece-Wise technique for line segmentation on Thai handwritten historical documents", 2011 International Conference on Machine Learning and Cybernetics, 2011. Publication	<1%
3	Mello, Carlos A.B.. "Segmentation of images of stained papers based on distance perception", 2010 IEEE International Conference on Systems Man and Cybernetics, 2010. Publication	<1%
4	Monique Thonnat. "A Cognitive Vision Approach for Image Segmentation Thresholding Images of Historical Documents	<1%