A Hybrid Technique to Remove Back-to-Front Interference in Historical Document Images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
Arushi Singhal
Roll No. 601303005

Under the Supervision of
Dr. Rajiv Kumar
Assistant Professor

COMPUTER SCIENCE AND ENGINEERING DEPARTMENT

THAPAR UNIVERSITY

PATIALA – 147004

MAY 2015

Certificate

I hereby certify that the work which is being presented in the thesis entitled, **"A hybrid technique to remove back-to-front interference in historical documents"**, in partial fulfillment of the requirements for the award of degree of Master of Technology in *Computer Science and Applications* submitted in Computer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 this report has not been submitted in part or full to any other University or institute for the award of any degree.

Arushi

Arushi Singhal

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

Rajiv Kumar

(Dr. Rajiv Kumar)
Assistant Professor
CSED

Countersigned By:

Deepak Garg

(Dr. Deepak Garg)
Head
Computer Science and Engineering Department
Thapar University
Patiala

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

Acknowledgement

First of all, I would like to express my gratitude to **Dr. Rajiv Kumar, Assistant Professor,** Computer Science and Engineering Department, Thapar University, Patiala for introducing me to image processing and for all his guidance and support. This thesis work was enabled and sustained by his vision and ideas. I have been amazingly fortunate to have an advisor like him who gave me the freedom to explore new ideas on my own and at the same time the guided me to recover when my steps faltered. His patience and support always helped me overcome many crisis situations and successfully complete this dissertation.

I am also thankful to the entire faculty and staff of Computer Science and Engineering Department (CSED) and my friends who devoted their valuable time and constantly supported me in all possible ways towards completion of this work. I thank all those who have contributed directly or indirectly to this work.

Lastly, I would also like to thank **my parents** for their years of unyielding love and encouragement. They have been my backbone and have supported me in all walks of life.

Arushi Singhal
(601303005)

Abstract

The study of historical documents is a topic that presents major challenges for researchers from various fields such as history, political science, psychology, computer science, among others. Historical documents contain significant information about cultural and scientific value. Historical artifacts consist of documents, letters, newspapers, pictures, maps, etc. Many of these are stored in libraries, museums or government archives. However, due to the preservation, few people have access to this material. Also such documents are frequently degraded over time. In order to make easier the access to this rich source of knowledge of the history of a society, digitization of the material comes as a possible solution. Digitized degraded documents require specialized processing to remove different kinds of noise and to improve readability. However, handling these documents is extremely delicate. Getting software to do the work automatically what the user would need to do manually can bring great financial and historical benefits, alongside with better preservation. The problem is further aggravated if the document is written on both sides because with time the ink from the back side of the paper tends to seep through and disturbs the visibility of text on the other side during digitization of paper. This effect is called as "ink-bleed through" or "back-to-front interference". Among the document image processing steps, the segmentation is one of the most important as it will be responsible for identifying what needs to be recognized. The first step of segmentation is the thresholding (or binarization) of the image. Binarization identifies which pixels belong to the foreground image and which belong to the background. A misclassification of the pixels can impair subsequent stages of processing. We present a new approach for this problem by filtering the background first using ideas of *visual perception theory*. When an observer stands back from a document, he/she loses the details of the image (as the acuity of the human vision decreases with the distance). Distant objects project smaller images onto the retina. As we increase the distance from the object, the details are lost and only the main colors remain. This idea is used to binarize the degraded historical documents and remove "back-to-front interference".

Table of Contents

List of Figures

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 small 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 process of converting scanned documents or images into machine readable characters. The input to OCR can be scanned images 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.

Character recognition is also referred as optical character recognition which is a current subject of research and has immense potential in future also where we would want to track and locate every bit of information that is being exchanged 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 into the following categories on the basis of data acquisition:-

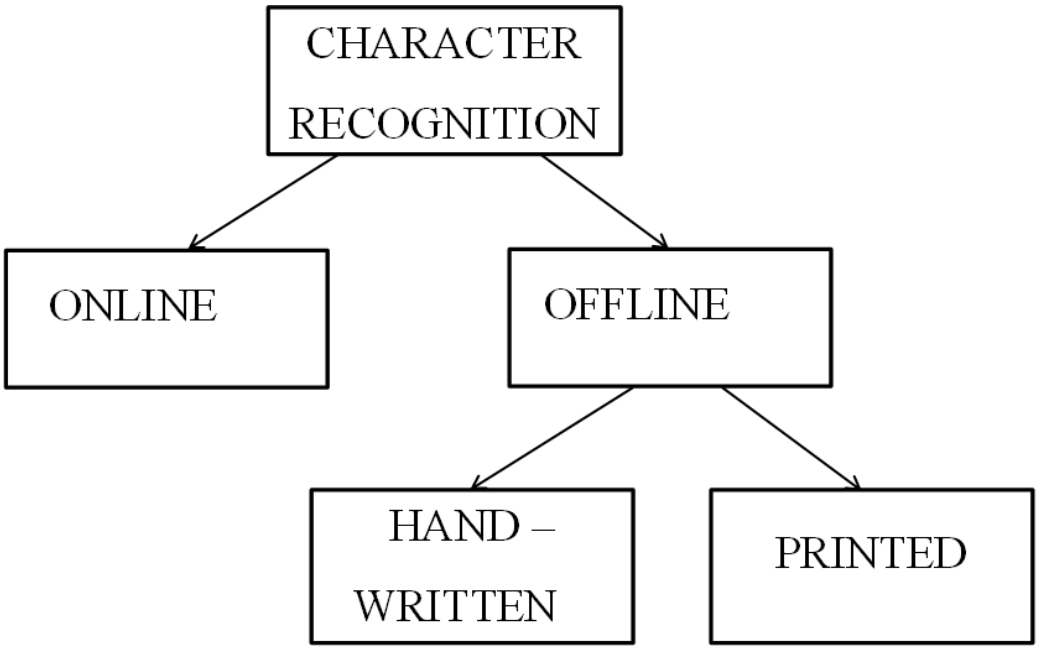


Figure 1: Classification of Character Recognition System

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:-

1. Online Handwriting Recognition
2. 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 function of time and are then stored 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 real time contextual information but offline data does not, a major difference between online and offline recognition. This variation generates a major deviation in processing methods.

1.3. 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:

(i) **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.

(ii) **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.

(iii) **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.

(iv) **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.

(v) **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.

1.4. APPLICATIONS OF OFFLINE HANDWRITING RECOGNITION

Offline handwritten recognition has the following application areas:

1.4.1. 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.

1.4.2. 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.

1.4.3. Other Industries

Many other fields such as education, finance, and government agencies also use OCR. It has made innumerable text available online for researchers and students saving their money and making a large source of knowledge to be shared just with a few clicks and search keywords.

1.4.4. 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.

1.5. 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 postprocessing.

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 below in more detail.

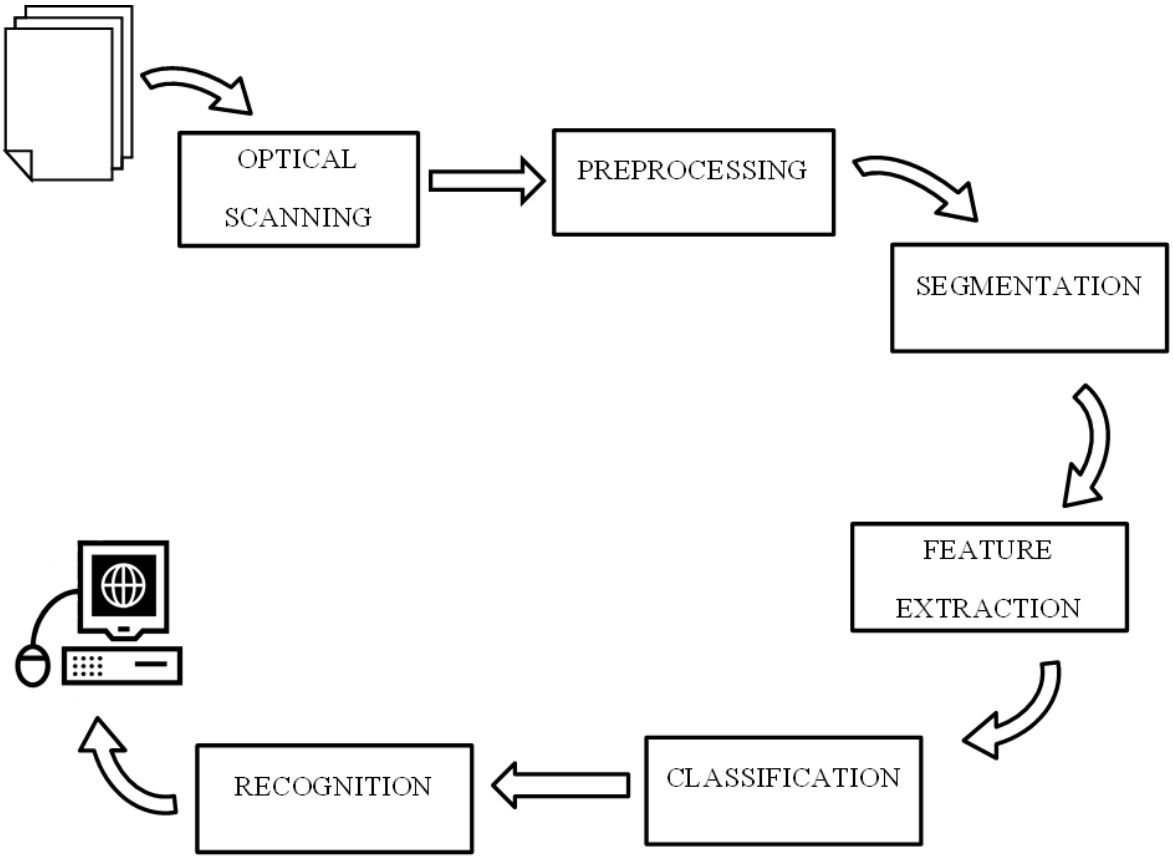


Figure 2: Phases of OCR

1.6. PREPROCESSING

Depending on its type of data acquisition, the raw data is subject to a lot of preliminary 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. Pre-processing methods include noise removal, normalization, compression, smoothing, skew correction and thinning. Removing any unwanted bit-patterns, which do not have significance in the output, is the main objective of noise removal. It also simplifies pattern recognition process without missing any important information. It reduces any inconsistent data, if present. It also enhances the image for the next steps.

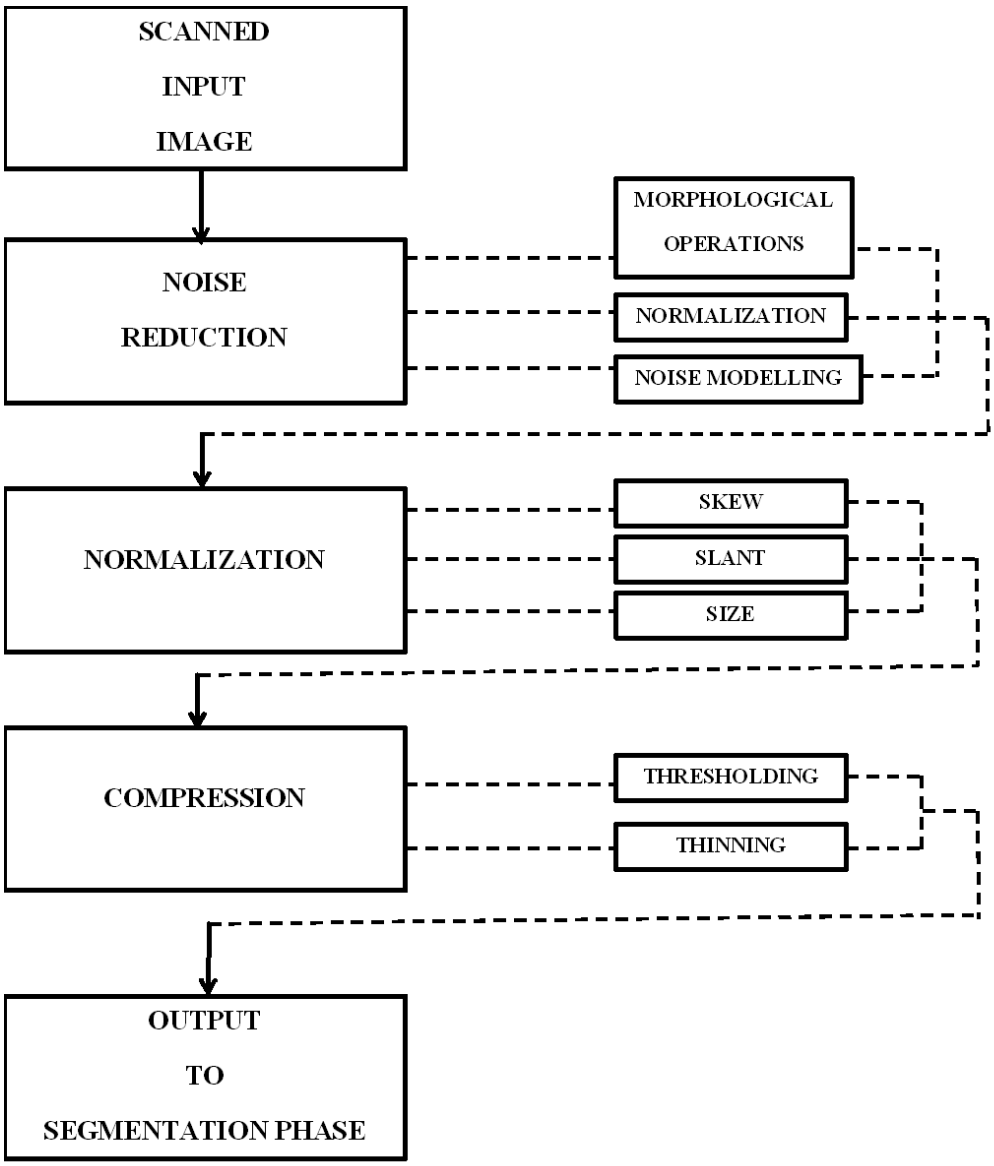


Figure 3: Various stages in Preprocessing

1.6.1. 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.

1.6.1.1. 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 remove these inadequacies prior to Character Recognition.

There are three major groups of noise reduction techniques:-

(i) ***Filtering:*** To remove noise and diminish unnecessary points usually introduced by uneven writing platform and/or poor rate of sampling of data 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 designed for contrast adjustment purposes, smoothing, sharpening, thresholding and removing slightly textured or colored background.

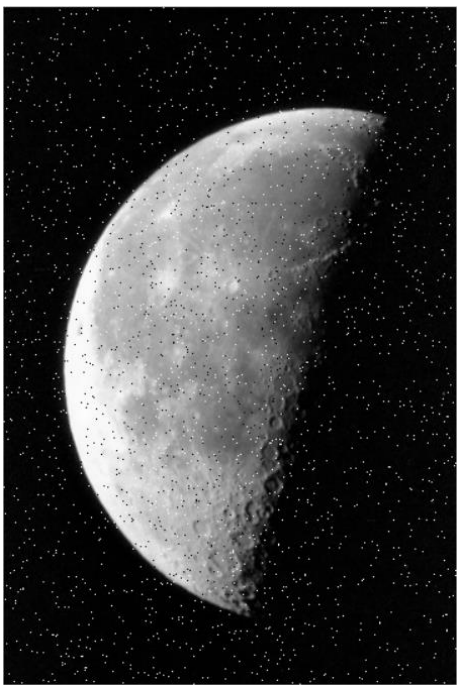


Figure 4: Image with Salt and Pepper Noise

Figure 5: Image after noise removal

(ii) ***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 operations can be used for removal of noise in the document.

(iii) ***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.

1.6.1.2. 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:-

(i) ***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. 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.

Figure 6: Skewed image [26]

(ii) ***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.

Figure 7: Image before and after slant normalization [30]

(iii) ***Size Normalization:*** Adjusting the size of characters to a standard size is called size normalization. The character is divided into a number of zones where each zone is scaled separately and hence normalization is performed both horizontally and vertically.

(iv) ***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 inorder to represent the document.

1.6.1.3. 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:

(i) ***Thresholding:*** Thresholding is used to reduce the storage requirements and increase the processing speed of the images where grayscale images as represented as binary images 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 entire document image based on an estimation of the background intensity level from the intensity histogram of the image 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.

(ii) ***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 *nonpixel* 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 Nonpixel wise thinning.

1.6.2. THRESHOLDING

One of the techniques used for image compression is thresholding. It is performed to reduce storage and increase the performace by converting grayscale images into binary images 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 differentiate the pixels we want to analyze from the remaining, a comparison of every pixel intensity value with respect to a *threshold value* is performed. Once we have separated the required pixels, we can assign them with a fixed value to identify them which means a value of 0 (representing black), 255 (representing white) or any other value according to our needs can be assigned.

1.6.2.1. GLOBAL THRESHOLDING

Global Thesholding selects an optimal or several optimal threshold values automatically in order to separate the objects of our interest in an image from the background based on their gray-level histogram distribution. It assigns a single unique threshold value to the whole image based on an estimation of the background intensity level from the intensity histogram of that image. 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.

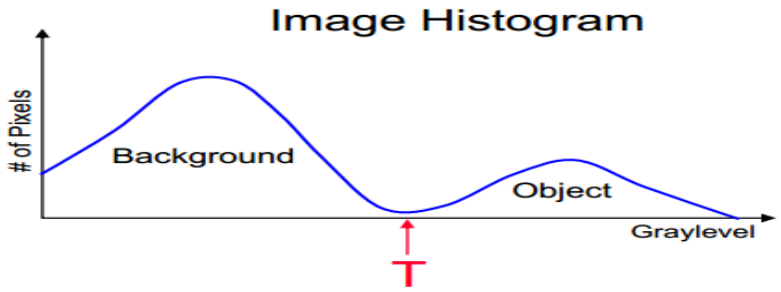


Figure 8: Image Histogram for Global Thresholding [34]

- $g(x,y) = \begin{cases} 1, if\ f(x,y) > T \\ 0, if\ f(x,y) \leq T \end{cases}$

GLOBAL THRESHOLDING ALGORITHM

The basic global threshold, T, is calculated as follows:

1. Select an initial estimate for T (the average gray-level in the image)
2. Segment the image using T to produce two groups of pixels: G_1 consisting of pixels with gray levels >T and G_2 consisting pixels with gray levels ≤ T

3. Compute the average intensity of pixels in G_1 to give μ_1 and G_2 to give μ_2
4. Compute a new threshold value: $T = \frac{\mu_1 + \mu_2}{2}$
5. Repeat steps 2 – 4 until the difference in T in successive iterations is less than a predefined limit T i.e. If $|T - T_{new}| >$, back to step 2, otherwise stop.

This algorithm can be applied to find the threshold value when the histogram is bi-modal.

The thresholding technique is demonstrated below in the figure 10 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.

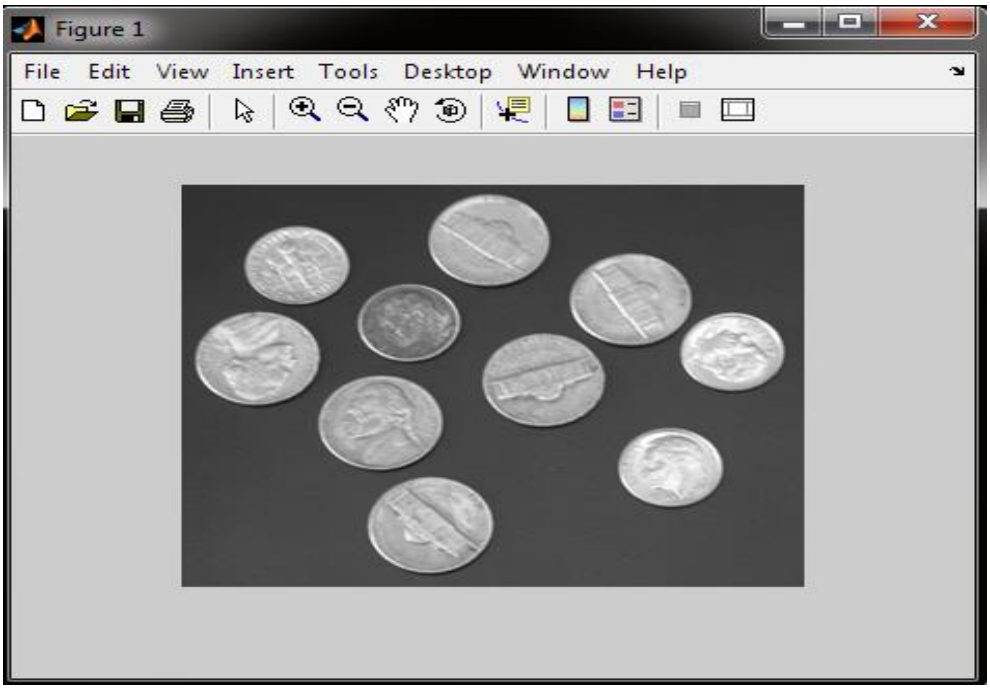


Figure 9: Grayscale Image for Global Thresholding

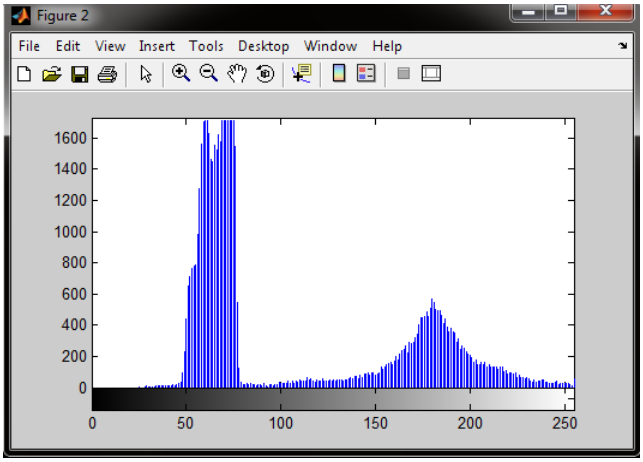


Figure 10: Histogram for image in Figure 8

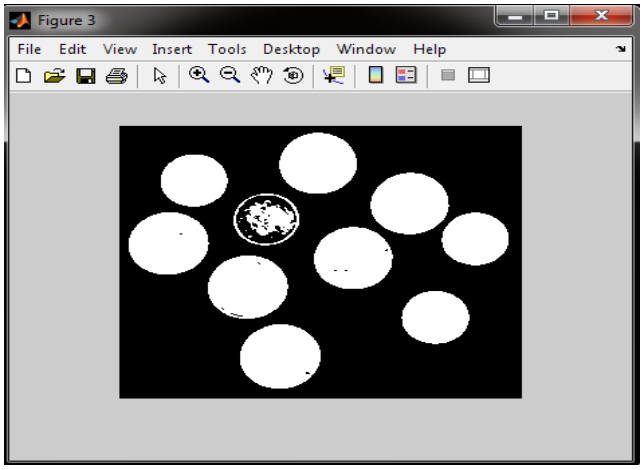


Figure 11: Image after Global Thresholding

1.6.2.2. 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 image into several sub images and then threshold each of these sub image 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 analysed within the local windows across the entire image inorder to determine the local thresholds. The local area information of the pixel determines 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 inoder to obtain a good estimate of the average values but not too large to pick over average non-uniform intensity of background.

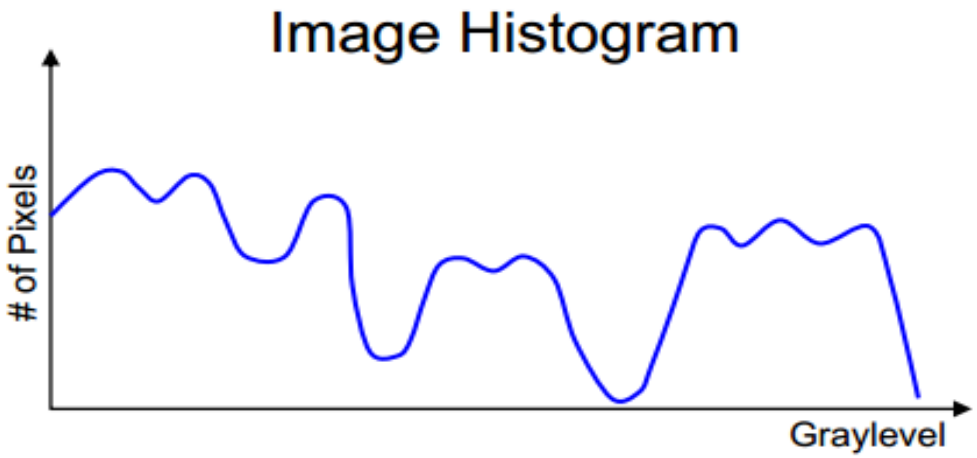


Figure 12: Image Histogram for Local Thresholding [34]

- $g(x,y) = \begin{cases} a, iff(x,y) > T2 \\ b, if\ T1 < f(x,y) \leq T2 \\ c, if\ f(x,y) \leq T1 \end{cases}$

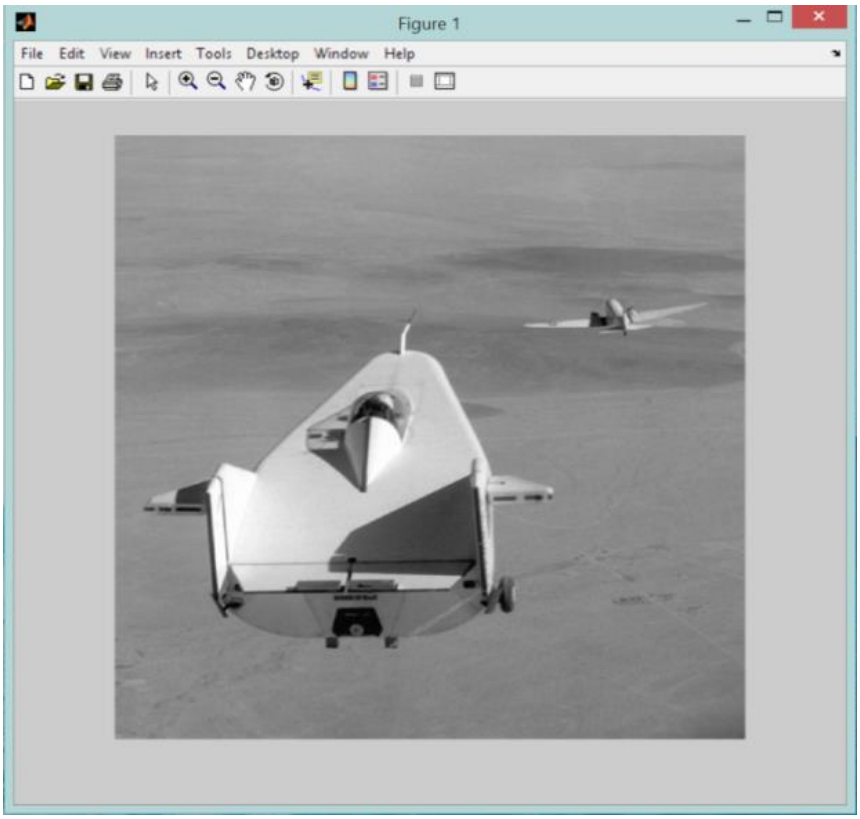


Figure 13: Grayscale image for local thresholding

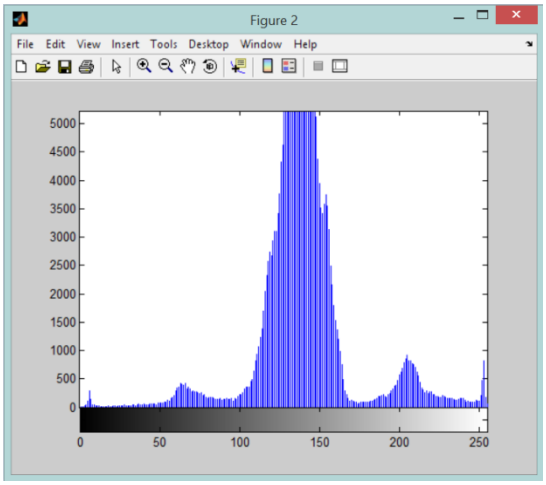


Figure 14: Histogram for image in Figure 12

1.7. 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.

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 inorder 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.

1.7.1. 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. 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:-

Based on two aspects of images:

- **Detecting Discontinuities:** Partitioning an image on the basis of sudden changes in intensity i.e. edge detection.

- **Detecting Similarities:** Partitioning an image into different regions on the basis of similarity according to a predefined criterion i.e. thresholding, region growing, region splitting and merging.

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 15: Classification of Image Segmentation Methods

1.7.2. 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 grouped into two main categories:

(i) Gray Histogram Technique

In this technique, 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.

(ii) 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 inorder to form closed boundary of regions. Sobel, laplacian of gaussian (log), lapclace 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. If the level of accuracy is too high, then noise can bring up fake edges which make unreasonable image outline. If the level of noise immunity achieved is too high, then some parts of image outline may remain undetected and object position can be wrong. Therefore, this method can generate equally good results on complex and noisy images along with simple and noiseless images.

(iii) 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.

1.7.3. 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.

(i) 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 are based on image space segmentation of regions. The basic idea behind this method is based on the characteristics of an image.

Thresholding is applied to select a threshold value T 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 entire image then we can say that any pixel intensity value which is greater than the threshold value T belongs to the object and the pixel intensity values which are smaller than the threshold value belong to the background.

(ii) 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 form a region.

It is performed as follows:

- Find a seed pixel as an initial point for each of needed segmentation.
- Join together the similar or like properties of pixel (Based on a predetermined growing or similar formula to determine) with the seed pixel around the seed pixel domain into the domain of seed pixel.
- These new pixels act as a new seed pixel to continue the above process until no more pixels that satisfy the condition can be included.

b) **Region Splitting and Merging** In this method, the image is subdivided into a set of arbitrary disjoints regions and then merge and/or split the region according to the given condition for segmentation. Region based segmentation is largely influenced by splitting. This splitting method can be represented in the form of quad trees in which every node has exactly four branches.

It is performed as follows:

- Split the region into the four disjoint branches.
- When no further splitting is possible, merge any region.
- Stop when no further merging is possible.

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

1.7.4. 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.

(i) **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.

a) Fuzzy c means clustering This method identifies the natural groups of data inorder to form a large set of data which produces a brief representation of the systems behaviour. Fuzzy c-means is one of the data clustering methods in which datasets are grouped into ‘n’ clusters where every data point from the dataset belongs to every cluster to a certain degree.

b) K-Means Algorithm This algorithm groups data vectors into a predefined number of clusters. The centroids of the predefined clusters are initialized randomly initially. The dimensions of the centroids are same as the dimensions of data 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 changes are noticed for all cluster means or for a fixed number of iterations.

(ii) **Neural Network-based segmentation**

The image 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

- Highly parallel ability and fast computing capability make it apt for real-time application.
- Unrestricted nonlinear degree and high interaction among processing units makes this method capable to establish modelling for any process.
- Reasonable robustness makes it insensitive 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.

1.7.5. Model Based Segmentation

The above segmentation methods use only local information about the objects or pixels. Since humans have the capability to recognize objects 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

shape of objects is required. This information can then be compared with the local information. This is the basic concept behind model based segmentation.

1.8. 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, a more compact and distinctive representation is required for avoiding complexity and increasing accuracy of the algorithms. Inorder to accomplish this, we need to extract a set of features from objects/letters to form a feature vector for each class. This feature vector helps the recognition system to distinguish an object from other classes while remaining immune to the characteristic differences within 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 between different classes by looking at these features.

1.9. CLASSIFICATION

The CR process assigns a character image to a class by using a classification algorithm based on the features extracted and the relationships among the features. Since members of a character class are equivalent or similar in as much as they share defining attributes, the measurement of similarity, either explicitly or implicitly, is central to any classifier. In this stage, we train the neural net using the feature vectors obtained during feature extraction method against the required targets.

Feature extraction is concerned with recovering the defining attributes hidden by imperfect measurements. To represent a character class, either a prototype or a set of samples must be known. The feature selection process attempts to recover the pattern attributes characteristic of each class. The classification stage identifies each input character image by considering the detected features.

1.10. POSTPROCESSING

The incorporation of context and shape information in all the stages of character recognition systems is necessary for meaningful improvements in recognition rates. This is done in the post-processing stage with a feedback to the early stages of character recognition. The simplest way of incorporating the context information is

the utilization of a dictionary for correcting the minor mistakes of the character recognition systems. 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 basic idea is to spell check the character recognition output and provide some alternatives for the outputs of the recognizer that do not take place in the dictionary.

1.11. HISTORICAL DOCUMENTS

Historical documents are of great importance to us due to their cultural and scientific value. Therefore, the study of historical documents is a huge challenge among researchers from various fields such as history, psychology, political science, computer science, and many others. Historical artefacts consist of documents, maps, pictures, newspapers, letters, etc. Most of these are stored in museums, libraries, and/or government archives. However, only few people have access to this material due to the preservation. In order to avail easier access to this rich source of information and knowledge 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. Inorder 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 readability 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.

1.11.1. PROBLEMS IN HISTORICAL DOCUMENTS

Paper is very fragile and susceptible to aging still it is the primary medium to store information. There are oodles of precious historical documents preserved in libraries all over the world. This invaluable archive is exposed to progressive decay even with all the careful 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 documents which occur due to various different reasons [4].

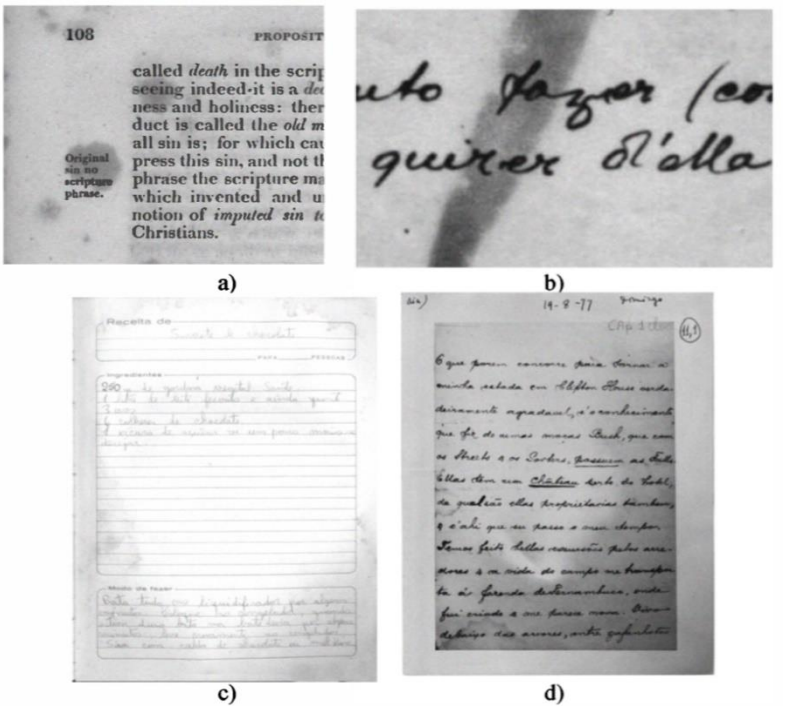


Figure 16: Sample historical documents with (a) smudges (b) marks of adhesive tape (c) faded ink and different levels of degradation and (d) non-uniform illumination in a two-color paper.

Some of the problems found in historical artefacts are:

- paper assume dark or light brown tones
- damages in the paper due to deterioration
- smudges, blemishes or dirt from 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 [2].

1.11.2. TYPES OF DEGRADATION IN HISTORICAL DOCUMENTS

On the basis of origin of degradation we can classify document image degradation into various types. For example we classify degradation which occurred due to time and degradation which occurred due to digitization into two different classes. All the effects which occur in documents due to environment like bad climatic conditions at the place in which the documents are stored can be classified in degradation with time class.

Even after digitization, the defects in the document that is scanned remain a problem to be solved. But some additional defect might be added during scanning. So a suitable typology for degradation of documents must be defined in order to deal with the above problems. The typology proposed here is based on the type of treatment required for restoration of historical documents. It can be divided into three main categories: background degradation (paper), foreground degradation (images of text, features, characters, drawings) and global degradation.

1.11.2.1. Noisy Background Degradation

The presence of artefacts in the background causes difficulty in reading the document. This may occur due to several reasons: digitization and humidity can cause spots; bleed-through or back to front interference can be caused if the ink on the back side of the paper seeps through and shows up on the front side of the document, unwanted features like strokes of pen, underlined text etc. For example, we need to remove the line strokes caused by underlining the text for efficient segmentation and recognition of text. This type of degradation is simulated by new layers with different gray level values which are superimposed on the image of document. The restoration of these documents includes segmentation and

classification phases to extract the desired components and convert the image into readable text.

1.11.2.2. Noisy Foreground Degradation

This type of degradation causes touching or broken letters which lead to OCR errors. The main reason behind this type of degradation is interference of foreground degradation (like spots) with the text of the document. So, every pixel of the image which lies inside the spot has information about the actual data and noise both. The ink of the paper is also affected with time. The chemicals of which the paper is composed can lead to fading of the ink and causes gaps in the document text. These gaps result in losing significant information. Gaps create regions in the document image which become difficult for recovery. On one hand, gaps cause complete loss of information whereas semi-transparent blotches cause partial loss (some part of the information can still be preserved). So filling the gaps in the text document image is more complicated than treating blotches. There are several other reasons also for ink degradation like compression of the image and low resolution of the document image. While compression can lead to space saving and speeds up the image processing but it also causes ruptures to be introduced in character layouts and causes many modifications in their contours. So, one of the main characteristics of this type of degradation is loss of information. The restoration of documents suffering from these degradations require repairing shapes of features to extrapolate information which was lost to attain readability of the text in these documents. This requires prior information about the document and is quite difficult to deal with.

1.11.2.3. Global Degradations

This type of degradation affects the complete document. Geometrical degradation is a common example for global degradation which is caused by scanning of thick documents. The paper surface of a thick book gets curved while scanning. The presence of this curvature leads to warped words appearing around the book spine area. Time can also cause degradation and fading of the text colours. This type of degradation can be characterized by the application of transformations on the image of original document. These transformations are either applied on the local information of the pixel like skew, curves or on the color of the pixel. It is important

to notice that the restoration is oriented towards modelling image degradation and applying inverse procedures which succeed in finding the desired information.

1.12. MOTIVATION BEHIND THIS WORK

We know that the information contained in historical documents is of extreme value for us and therefore the need of preserving these arises. But since these documents are either handwritten of printed on paper and paper ages with time and becomes susceptible to deterioration, we need to digitize these documents to make them available for later generations to study and cherish the past civilizations. However, digitization of these documents is not an easy task. We must handle the documents very carefully and delicately while digitization. But digitization might introduce various noises during the process. So, the removal of this noise is necessary in order to correctly recognize the information available in the documents. Hence, we need efficient techniques to be developed to remove all the inaccuracies and noises from the documents and make their content available to the researchers, students and teachers worldwide.

1.13. OBJECTIVE OF THESIS

The objective behind this thesis was to study various types of problems in historical documents as they are a significant source of knowledge and great historical value to us. After studying certain problems, it was observed that “bleed-through” effect was one of the common problems in historical documents caused due to noisy background i.e. degradation of paper so there arises the need to remove the interference from the back of the historical document to the front of the document also known as “show-through” which occurs when the ink seeps through and affects the readability of the text in the document. Therefore, to develop and implement an algorithm that can remove “bleed-through” or “back-to-front interference”. For this, a prior requirement was to obtain a database of historical document images and classify them on the basis of level of interference in them into three categories: strong interference, medium interference or light interference. Classification requires a dataset about the features of the document so we need to extract the features of these historical document images for correct classification. Lastly comparison of the results with the previous algorithms and its detailed analysis is shown.

1.14. ORGANIZATION OF THESIS

The thesis is organized in a sequential way starting with introduction to optical character recognition and its phases followed by description of historical documents and various types of problems associated with them due to degradation. Chapter 2, this chapter includes the description of the literatures that reveals various developments and the character recognition process and different techniques used so far to remove the "back–to-front interference" in the historical documents. Chapter 3, this chapter details about the problem definition. Chapter 4, in this chapter the authors have implemented a new hybrid algorithm that is efficient in removing "back-to-front interference" or "bleed-through" problem in historical documents. Chapter 5, thesis is concluded with a summary of this work and discussion on future scope of research.

CHAPTER 2

LITERATURE SURVEY

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.

The true-color to gray-scale conversion is done by:

$$Gray_{level} = 0.299r + 0.587g + 0.114b$$

where $Gray_{val}$ is the new pixel value

r, g and b are red, green and blue values of the original pixel.

The entropy-based algorithms take image histogram as an input and normalize each of its entries by total number of pixels in the input image, producing a distribution of probabilities of gray levels. Therefore,

$$p_{i=\frac{n_i}{N}} \tag{1}$$

$$P_i = \sum_{i=0}^{t} p_i \tag{2}$$

where n_i is the number of pixels with gray level i varying from 0 to 255,

N is the total number of pixels in the image,

p_i, i=0,1,…,t, is the probability distribution of the histogram of the image taking into account the relative pixel frequency, and

P_i is the sum of all probabilities from i=0 to i=t in the histogram.

N.Otsu [1975] developed an effective thresholding technique called Otsu's thresholding used widely in real thresholding tasks. He gave a nonparametric and unsupervised approach for automatic thresholding used in image segmentation. In this method, an optimum value of threshold is selected by “minimizing the sum of within-class variances of foreground and background pixels” using discriminant criteria. It is done inorder to maximize the separability of the resultant classes in gray levels. This

technique is very simple. It utilizes only the zeroth and the first-order cumulative moments of the gray-level histogram. It can be easily extended to multi-threshold problems. Considering the above said, we can say that this method can be suggested as a simple and standard one for automatic threshold selection that can be applied to many practical problems [25].

The mean and variance of the object and background in relation to the threshold t are defined as follows.

$$m_b(t) = \sum_{i=0}^{t} i.p_i \qquad \sigma_b^2(t) = \sum_{i=0}^{t} [i - m_b(t)]^2 \; p_i$$

$$m_w(t) = \sum_{i=t+1}^{255} i.p_i \qquad \sigma_w^2(t) = \sum_{i=t+1}^{255} [i - m_w(t)]^2 \; p_i$$

The optimal value for this limit is the argument that maximizes the following expression

$$\eta(t) = \frac{P_t(1 - P_t)[m_b(t) - m_w(t)]^2}{P_t\sigma_b^2(t) + (1 - P_t)\sigma_w^2(t)}$$

Pun et al. [1981] proposed an image segmentation method that selects threshold automatically. He used the concept of entropy to calculate threshold which was associated with the asymmetry of the gray level histogram. In his algorithm, the gray levels are considered to be produced by a source with an alphabet consisting 256 statistically independent symbols. But it was not suitable for back-to-front interference removal. It allows un-supervised choice of one or more threshold values [27].

He considers the ratio between the a posteriori entropy

$$H'(t) = -P_t \log[P_t] - [1 - P_t] log[1 - P_t] \tag{3}$$

and the source entropy

$$H(t) = H_b(t) + H_w(t) \tag{4}$$

where H_b and H_w are:

$$H_b(t) = -\sum_{i=0}^{t} p(i)\log(p(i)) \tag{5}$$

$$H_w(t) = -\sum_{i=t-1}^{255} p(i)\log(p(i)) \quad (6)$$

and $p(i) = p_i$ is given by equation (1).

$$\frac{H'(t)}{H} \geq Fe(\alpha)$$

$$= \alpha \frac{logP(t)}{\log[\max(p_0, \dots, p_t)]}$$

$$+ (1-\alpha)\frac{\log[1-P(t)]}{\log[\max(p_{t+1,\dots,}p_{255})]},$$

$$H_b(t) = \alpha H \quad (7)$$

where,

The threshold is obtained for the value of t that satisfies equation (7), where α is the argument that maximizes $Fe(\alpha)$.

Johannsen et al. [1982] gave an algorithm which minimized "the sum of entropies of the foreground and the background" and uses a log of the foreground entropy. The objective of this algorithm is to minimize the function $S(t)$ defined as follows:

$$S(t) = S_b(t) + S_w(t) =$$

$$= \log(P_t) + {}^{1}/_{P_t}\,[E(p_t) + E(P_{t-1})] + \log(1-P_{t-1}) + {}^{1}/_{(1-P_{t-1})}\,[E(p_t)$$

$$+ E(1-P_t)]$$

where $E(p) = -p.\log(p)$ and p_i and P_t are provided by equations (1) and (2) respectively.

The value of t that minimizes $S(t)$ is its optimal value [16].

Kapur et al. [1985] suggested an algorithm which considers the foreground and background as two distinct sources such that whenever the addition of the two entropies reaches a maximum, its argument t reaches the optimal value [18].

The distribution of object A, the distribution of the points that correspond to the written part (ink or paint) that will be mapped onto black pixels is given by:

$$A: p(i) = \frac{p_i}{P_t}, 0 \leq i \leq t$$

Conversely, the distribution of object B, the distribution of the points that correspond to the document background (paper) that will be mapped onto white pixels is:

$$B: p(i) = \frac{p_i}{1 - P_t}, t + 1 \leq i \leq 255$$

The values of the entropies H_w and H_b are calculated through equations (5) and (6), with $p(i)$ according to A and B as above.

Yen et al. [1995] recommended a new criterion for multilevel thresholding. The algorithm was based on two factors, discrepancy between the thresholded and original images and the number of bits required to represent the thresholded image. An entropic correlation is defined as

$$TC(t) = C_b(t) + C_w(t) =$$

$$= -log\left\{\sum_{i=0}^{t}\left[\frac{p_i}{P_t}\right]^2\right\} - log\left\{\sum_{i=t+1}^{255}\left[\frac{p_i}{1 - P_t}\right]^2\right\}$$

and the threshold is the argument that maximizes that expression. The functions $C_b(t)$ and $C_w(t)$ are known as Renyi entropy with ρ=2 [32].

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 [6].

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 [14].

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 [24].

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 [3].

Mello [2000] gave an algorithm that finds the most frequent gray level of the image and takes it as initial threshold to evaluate the values H_b, H_w and H by equations (5), (6) and (4) [5]. The entropies here are calculated with base N. The entropy H determines the value of weights m_b and m_w:

- If $H \leq 0.25$ then $m_w = 2$ and $m_b = 3$.
- If $0.25 < H < 0.30$, hten $m_w = 1$ and $m_b = 2.6$.
- If $H \geq 0.30$, then $m_w = 1$ and $m_b = 1$.

and the threshold is directly calculated by

$$t^* = 256(m_b H_b + m_w H_w)$$

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 [1].

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 [21].

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 [19].

M. Feldbach and K. D. Tonnies [2003] proposed a new approach for segmentation of dates in historical documents of church registers from 18^{th} and 19^{th} 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 [9].

C.A.B. 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 [23].

S. L. Feng and R. 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 significantly outperforms the other models and prior work using hidden Markov models on this heavily unbalanced dataset [10].

F. Drira [2006] suggested a typology for different types of degradation of old document images. 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 techniques are applied recursively to separate original text from interfering and overlapping areas of text [8].

Silva et al. [2006] proposed a segmentation method to generate high quality monochromatic images of historical documents based on entropy of histogram on an image. This algorithm can be used to remove 'back-to-front interference' in historical documents [15]. The basic idea behind his algorithm is considering the histogram distribution as the 256-symbol source (a priori source) distribution, as in Pun's algorithm, and that all symbols are statistically independent. This algorithm can also be applied to typed document files to improve the responses of OCR's tools. After the image synthesis process, it produces a series of 256 images with different interference levels two quality factors are calculated:

- The first step takes a reference image $s^{(manual)}$ which is obtained from S by manually searching a threshold that yields a good quality binary image. Then we compare the $S^{(manual)}$ image with each of the 256 $G_{fade}^{(k)}$ images and calculates the number of mismatching pixels:

$$q_{fade}^{(absolute\ reference)} = \sum_{n=1}^{N}\sum_{m=1}^{M}\left|S^{(manual)}(m,n) - g_{fade}^{(k)}(m,n)\right|$$

  This is the absolute reference mismatching factor, because the reference image is the same for all algorithms.

- The second quality factor, called self-referent mismatching factor, takes as reference image $S^{(k)}$, which is obtained by the application of the algorithm k onto image S, for its binarization. The number mismatching pixels between the $S^{(k)}$ and $G_{fade}^{(k)}$ is calculated as:

$$q_{fade}^{(self-referent)} = \sum_{n=1}^{N}\sum_{m=1}^{M}\left|S^{k}(m,n) - g_{fade}^{(k)}(m,n)\right|$$

  Where S is the document that plays the role of foreground information (signal image).

  I is the image that plays the role of back-to-front noise (interfering image).

  G_{fade} is the synthesized image by overlapping the S image with a faded mirrored version of the I image I_{fade}.

M. Makridis et al. [2007] proposed a new methodology for word segmentation in historical and degraded machine-printed documents. The major problems 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 [20].
It is based on:

(i) a dynamic run length smoothing algorithm that helps grouping together homogeneous text regions,
(ii) noise and punctuation marks removal,
(iii) obstacle detection in order to facilitate the segmentation process and
(iv) draft text line estimation procedure that guides the final word segmentation result.

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

G. Vamvakas et al. [2008] gave a complete OCR methodology for recognizing historical documents (printed and handwritten) without any prior knowledge of font. This methodology consists of three steps: The first two steps refer to creating a database for training using a set of documents, while the third one refers to recognition of new document images. First, a pre-processing step that includes image binarization and enhancement takes place. At the second step, a top-down segmentation approach is used in order to detect text lines, words and characters. A clustering scheme is then adopted in order to group characters of similar shape. Then, a database is created in order to be used for recognition. Finally, for every new document image the above segmentation approach takes place while the recognition is based on the character database that has been produced at the previous step. It is a semi-automatic procedure since the user is able to interact at any time in order to correct possible errors of clustering and assign an ASCII label [31].

I.B. Yosef et al. [2009] gave a new approach for text line segmentation based on adaptive local projection profiles for degraded documents with text lines written in large skew. This approach is fast and it applies the local algorithm in an incremental manner which adapts to the skew of each text line as it progresses. It gives highly accurate results for degraded documents with lines written in different skew angles and curvatures [33].

Carlos A.B. Mello [2010] has suggested a new approach to segment images of historical documents. 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 foreground information (the ink) by simulating the information we perceive when we go far from the document image. As we stand back, the text tends to disappear. Only the main colors from the background remain. This method is quite efficient in documents with various types of degradation although it is not suitable for small noises. For stained paper, his method achieved better results than 24 classical thresholding algorithms (including histogram-based, entropy-based and adaptive algorithms [22].

Rao et al. [2011] described that a gray-scale image of the noisy document is generally represented by $I(x, y)$.

$$I(x, y) = S, S \in [0,1]$$

Where x and y are the horizontal and vertical coordinates of the image $I(x, y)$ and S can take any value between 0 and 1 where $S = 1$ and $S = 0$ stands for black. There are two parts in the proposed Modified IGT. In the first part the level shifting of the pixels of an image is evaluated, while the second part of the algorithm, determines the relative importance of pixels with respect to object information [28]. After each iteration, some amount of pixels will be moved from fuzzy region to background. The iteration process will continue as long as the following criterion is satisfied is expressed by the equation given below

$$|T_i - T_{i-1}| = t$$

Where T_i is the threshold used in the i^{th} iteration

T_{i-1} is the threshold before the i^{th} iteration

t is used as a sensitivity parameter of threshold

Clausner et al. [2012] presented a hybrid text line segmentation method that uses a novel data structure and a rule base to combine the strengths of top-down and bottom-up approaches while minimising their weaknesses. The method uses a combination of rule based grouping of connected components (bottom-up) and projection profile analysis (top-down). The method works with bitonal images. Both input and output are represented using the PAGE format. Text regions, described by their outlines, are

populated with detected text lines. They followed the steps, Connected component analysis, Rule-based grouping of connected components to text line candidates, Splitting of large components in under segmented lines using local projection profile and repeat, Merging small line candidates to their nearest neighbor, Creating final text lines based on the candidates [7].

Garz et al. [2012] suggested a novel binarization free line segmentation method that is robust to noise and copes with overlapping and touching text lines. First, interest points representing parts of characters are extracted from gray-scale images. Next, word clusters are identified in high density regions and touching components such as ascenders and descenders are separated using seam carving. Finally, text lines are generated by concatenating neighboring word clusters, where neighborhood is defined by the prevailing orientation of the words in the document [12].

Fornes et al. [2013] hypothesized that show-trough are low contrast components, while foreground components are high contrast ones. A Multi-resolution Contrast (MC) decomposition is presented in order to estimate the contrast of features at different spatial scales. This decomposition is also able to enhance the image removing shadowed areas by weighting spatial scales. The main hypothesis is that show-through are low contrast components, while foreground components are high contrast ones. Furthermore, they hypothesized that background variation are spatially wide components, hence they can be considered low spatial frequency features. Thus, the method has three main steps. First, decompose the image into a Multi-resolution Contrast representation, which allows the contrast of components at different spatial scales. Second, enhance the image by reducing the contrast of low spatial frequency components. Finally, perform show through cancellation by thresholding low contrast components [11].

B. Gatos et al. [2014] gave a text zone detection and text line segmentation method for historical documents in order to achieve accurate text recognition performance. They faced several challenges such as horizontal and vertical rule lines overlapping with the text, two column documents and characters of different text lines touching vertically. Robust and efficient page segmentation is necessary for historical handwritten document images. For text zone detection, they analysed vertical rule

lines, connected components as well as vertical white runs while for text line segmentation. They enhanced an existing approach based on Hough transform in order to better treat cases of vertical connected characters. Their method proved gave promising results after an evaluation of a set of historical handwritten documents [13].

P. Kale et al. [2015] gave a hybrid binarization approach to improve the quality for the old documents using a combination of global and local thresholding techniques. Initially, global thresholding is applied to the whole image. The image areas that still have background noise are detected and the technique is again re-applied to each area separately. This achieves a better adaptability for the algorithm where different kinds of noise re-exist in different areas of same image. Global thresholding avoids the computational and time cost of applying a local thresholding in the entire image which is main advantage of using this technique. Hence it is effective in removing background noise and improving the quality of degraded images [17].

From this vast study of literature survey it is found that many different types of algorithm were given by many researchers to remove back-to-front interference in the historical stained documents. 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 an efficient algorithm was 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 artefacts are a significant source of knowledge about our history and civilization. There are many historical artefacts 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 Justification

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. According to Mello and Lins, "No single algorithm can provide equally good results to remove bleed through for all types of documents." The gaps can be summed up as:

- If prior information was available about the document nature then it would be easier to decide whether or not an algorithm will work for that particular type of the document.
- Therefore, there is a need of classification of historical documents with bleeding noise based on the strength of noise in them.
- Hence, depending on the strength or level of interference, a document must be first classified into light back-to-front interference, medium back-to-front interference and strong back-to-front interference.
- Thereafter, a suitable algorithm can be applied to the document to remove the interference.

Thus, the above gaps define objective of this thesis:

- To attain a database of historical documents, extract their features and classify them based upon the strength of bleeding noise.
- To develop an algorithm to remove interference from back side of the paper to front side of the paper.
- To test the algorithm on the database acquired and analyse if the results are desirable.
- To compare the algorithm proposed and compare the results with the existing algorithms.

The next chapter describes the proposed solution divided into three sections where section 4.1 describing preprocessing techniques applied to filter the image, section 4.2 detailing the segmentation to remove back-to-front interference, section 4.3 provides details on classification, feature extraction of the documents into three types based on their interference level and recognition.

CHAPTER 4

PROBLEM SOLUTION

"Back-to-front interference" is one the most common problems that occurs in historical documents with time. It affects the readability of the document and the recognition of text by OCR. Therefore, the need arises to develop an algorithm that removes any interference caused by the ink which seeps through the paper and attain readability and recognition of text by OCR. The proposed solution to remove back-to-front interference has been described in three sections below.

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 (O_{min}).
2. Let maximum of the original grayscale image be (O_{max}).
3. Find the range of intensity for the grayscale image.
4. Assign desired minimum intensity (D_{min}) as 0.
5. Assign desired maximum intensity (D_{max}) as 1.
6. Find the desired range of intensity.
7. Subtract O_{min} 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 objects 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. Segmentation

There are various ways to segment an image into lines, words and characters. The mechanism used in the proposed solution is based on visual perception theory. This principle simulates the concept of visualizing an image by standing back. When a person views an image by standing back, only the main colors from the background are visible while the text will not be visible. A similar concept was implemented using MATLAB for historical documents.

Algorithm for Segmentation

1. Subtract the image obtained after preprocessing from the original image and obtain I_{diff}.
2. Convert all black pixels into white.
3. For every pixel, if $I_{diff} \neq white$, negate I_{diff}.

4.3. Classification and Recognition

Classification of historical documents is required on the basis of strength of noise because researchers have pointed that no algorithm is good enough to remove bleed through in all kind of documents. A test set of documents with bleed through was formed by 150 real-world images which were obtained from historical files (Trinity College Library Dublin). Images were then hand labelled into three categories:

Strong, medium and light. SVM is used for the purpose of training while GLCM has been used for feature extraction and classification. The working of the classifier to handle spotting bleed through is as follows:

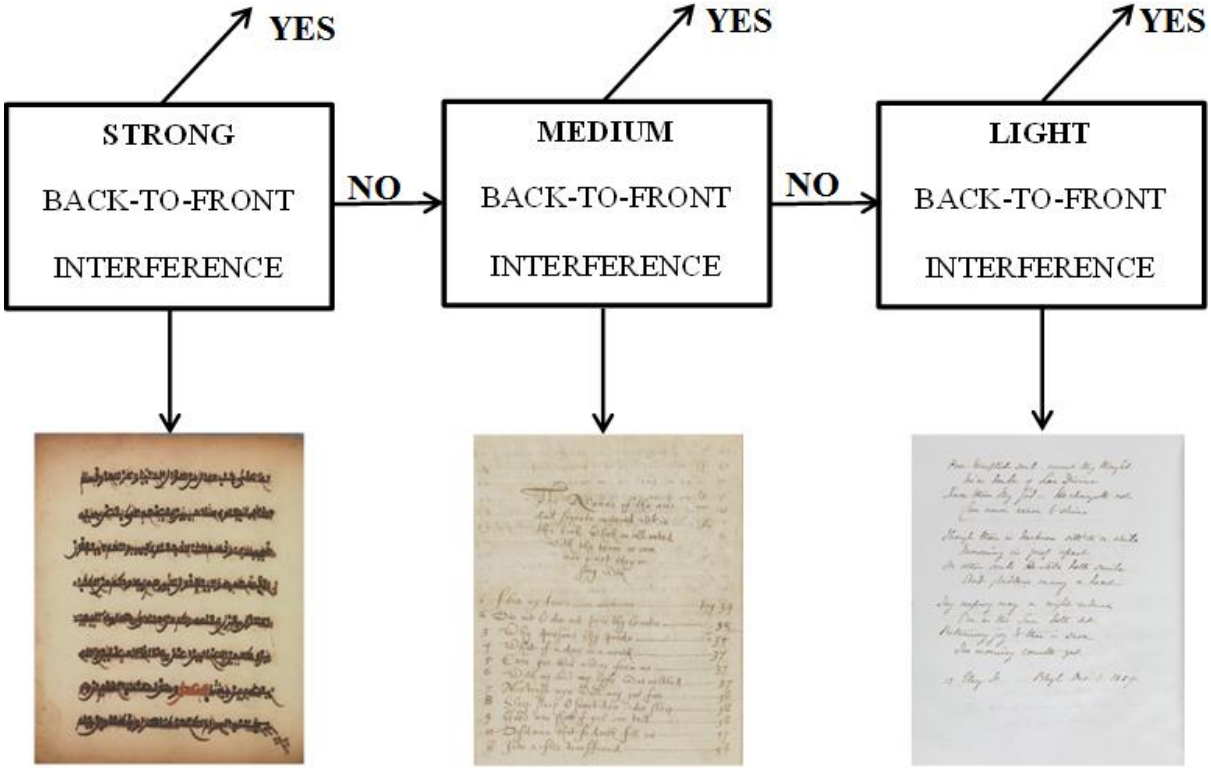


Figure 17: Classification of documents based on strength of bleeding noise

The proposed algorithm can be summarized as:

Algorithm 1: Training

1. Acquire a training set of 'n' images.
2. Extract the features (say F1) of this training set based upon Contrast, Correlation, Autocorrelation, Cluster Prominence, Dissimilarity, Energy, Entropy, Homogeneity, Sum of square, Sum average, Sum variance, Sum entropy, Difference in variance, etc. using GLCM algorithm and store them.
3. Group the images into three classes: Strong, Medium and Light using SVM training.

The above algorithm can be used for training a set of images to fall into their respective classes based upon the strength of bleeding noise observed in the historical documents. This algorithm is called by Check_Interference to test the images.

Algorithm 2: Check_Interference(Sample image: T1, Extracted Features: F1)

/* **T1** is a sample image file and **F1** is the matrix of extracted features. */

1. Store T1 in a temporary variable S1.
2. Extract the features (say F2) of S1.
3. Compare F2 with F1.
4. A class which has the maximum features matching with the test image is assigned to it.
5. Return class.

This algorithm takes a test image T1 as an input and it features are extracted. The extracted features are then compared with the stored features of the training set. The class with which the maximum features of the test image match is then assigned to it. This defines a basis to call the bleed through removal program.

Algorithm 3: BleedThroughRemoval (Sample image: T1)

1. Let T1(x,y) be the original grayscale image.
2. Normalize the intensity value to lie between 0 and 1 where 0 represents black and 1 represents white.
3. Apply morphological closing operation twice with two disks as structural elements with suitable radius.
4. Use filtering operation on the previous image to remove the noise (I_filter).
5. Calculate I_diff = abs(I_filter – I).
6. Convert all black pixels to white.
7. For every pixel(x,y), I_diff ≠ white, negate I_diff.
8. Return image T2.

Here, T2 is the image returned after removal of back-to-front interference.

This algorithm is used to remove the bleeding noise in the historical documents. It gives best result for images with light and medium bleeding noise. But it might give inappropriate results for images with strong interference. Therefore, a main program is design which checks the type of interference in the historical document. If the documents has light to medium type of bleed through effect, then we can call the above document directly but if it is of strong bleed through interference type then the user is prompted to continue or not. If the user still wishes to view the results obtained from the algorithm, he/she can click the 'yes' button else he/she can exit successfully by clicking on the 'no' button.

Main Program

1. Get the test image T1.
2. result = Call Check_Interference (T1).
3. If result = 1 or 2

```
        Then
            display "image has light or medium interference"
            a = Call BleedThroughRemoval (T1)
    Else
        If result = 3
            Then
                display "image has strong interference" and
                get the choice of user to continue or not in ch.
                    If (ch = yes)
                        Call BleedThroughRemoval (T1).
                    Else
                        If (ch = no)
                        Exit.
                    End if.
        End if.
    End if.
```

The above is the main program which is called by the user. It first checks the type of interfering bleed through noise in the document and then applies the back-to-front interference removal algorithm based on the result obtained.

All the above algorithms have been implemented in MATLAB and their results and analysis is discussed in the next chapter.

CHAPTER 5

RESULTS AND DISCUSSIONS

5.1. Results of Proposed Algorithm

The proposed algorithm in chapter 4 is implemented in MATLAB R2007b version and an interactive GUI is built using the same. The algorithm was applied to a test set of 150 historical documents with bleeding noise. The dataset was acquired from Trinity College Library Dublin.

First, the features of the images in the training set are extracted using GLCM and stored into a matrix after labelling each one of them into any of the three classes: strong, medium and light. This training is done by SVM based on the strength of bleeding noise in them. Then, the features of image belonging to test set are compared with the stored feature matrix of the images in training set and labelled as one of the three classes.

Based on this classification, the user is prompted if he/she wants to continue or not if the interference type is strong. Thereafter, the proposed bleed-through removal algorithm is applied to remove the interference from back side of the paper to the front side of the paper. This algorithm produces visually better results for light and medium interference document images when compared to the previous algorithms discussed in chapter 2.

The following section shows the snapshots of the running program in sequential order.

Results

First Window after Running the Main Program

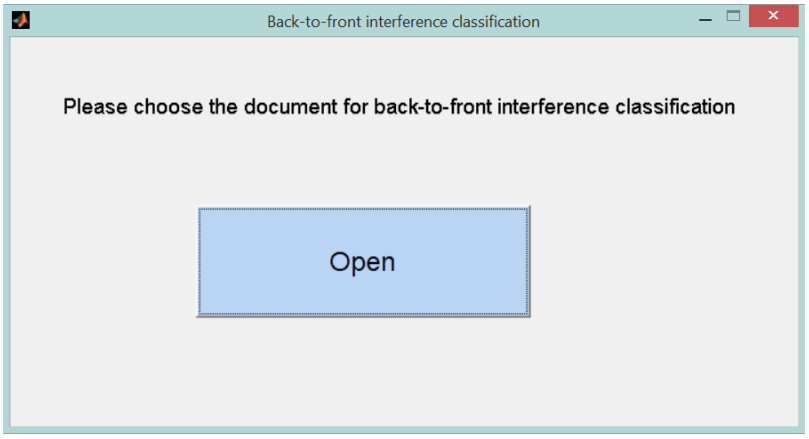


Figure 18: Selection window to choose a degraded document

The above figure shows the screen which appears just as the user runs the program. An input dialogue box appears on clicking the button “Open”.

Input dialogue Window to choose a historical Document

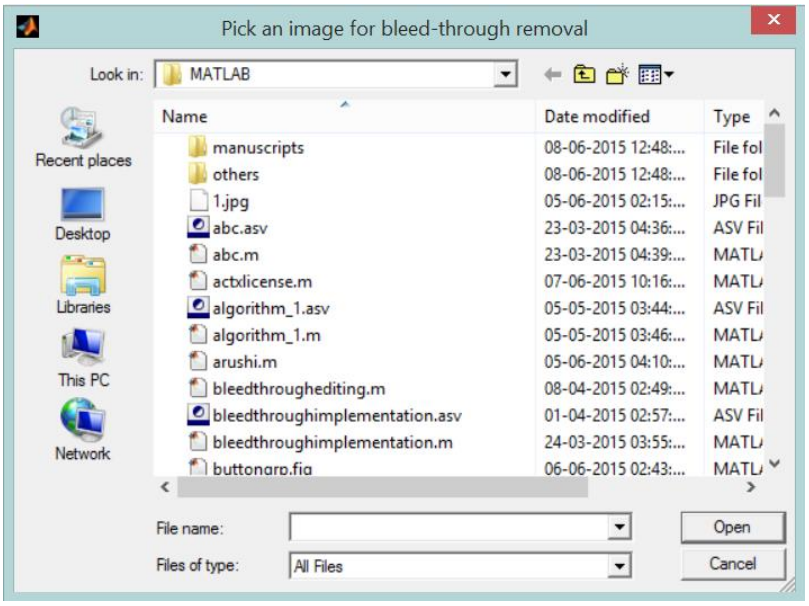


Figure 19: Popup window in response to click on Open button

The above window appears in response to the click of Button Open in figure 18. With this window the user can browse the location of the historical document.

Image belonging to class 1: Light interference

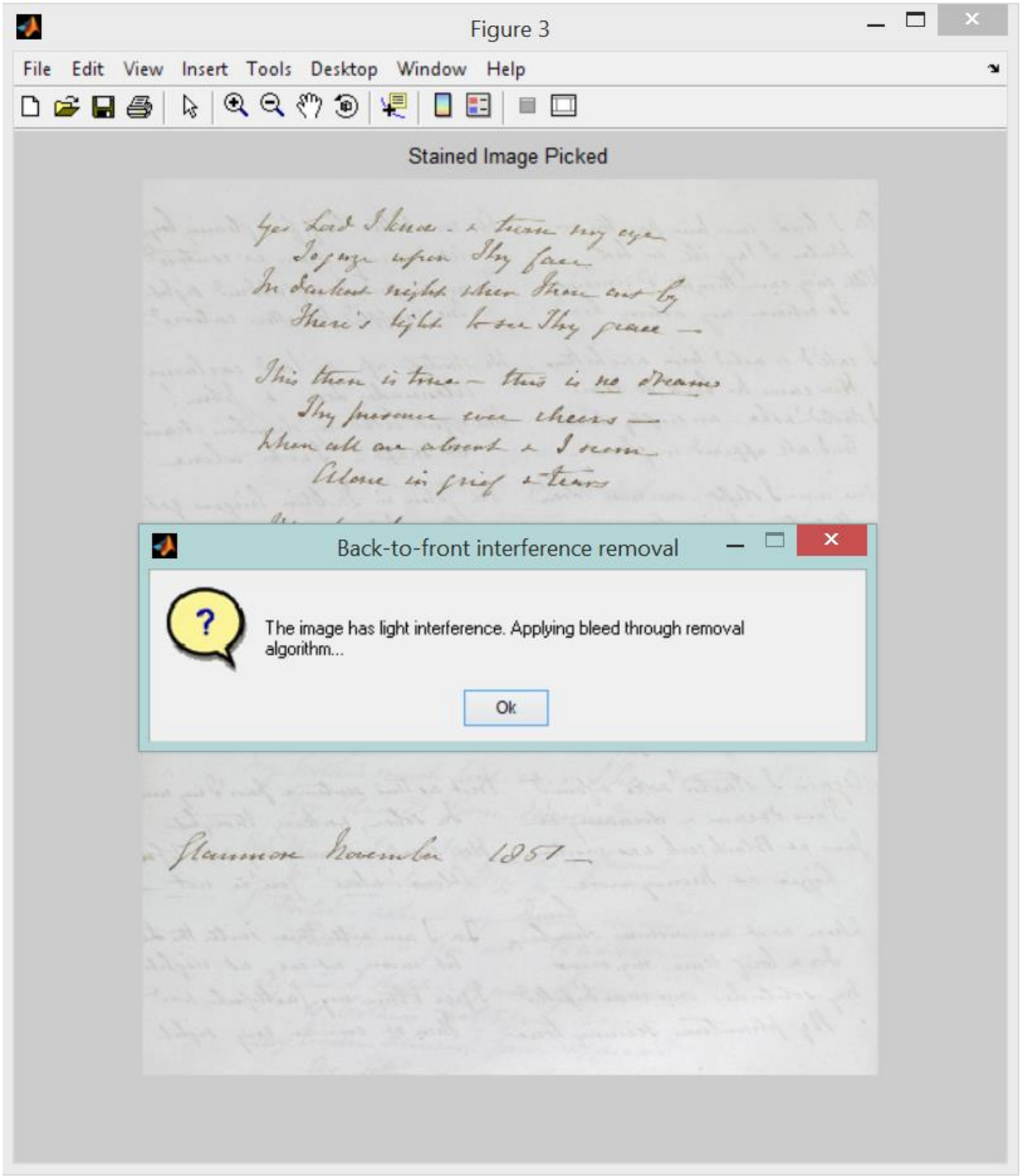


Figure 20: The image chosen by user with its class: light interference

The image chosen by user is displayed along with a message about the class to which it belongs. Since, the image belongs to light interference class, the algorithm is applied directly.

Class 1 image after bleed through removal

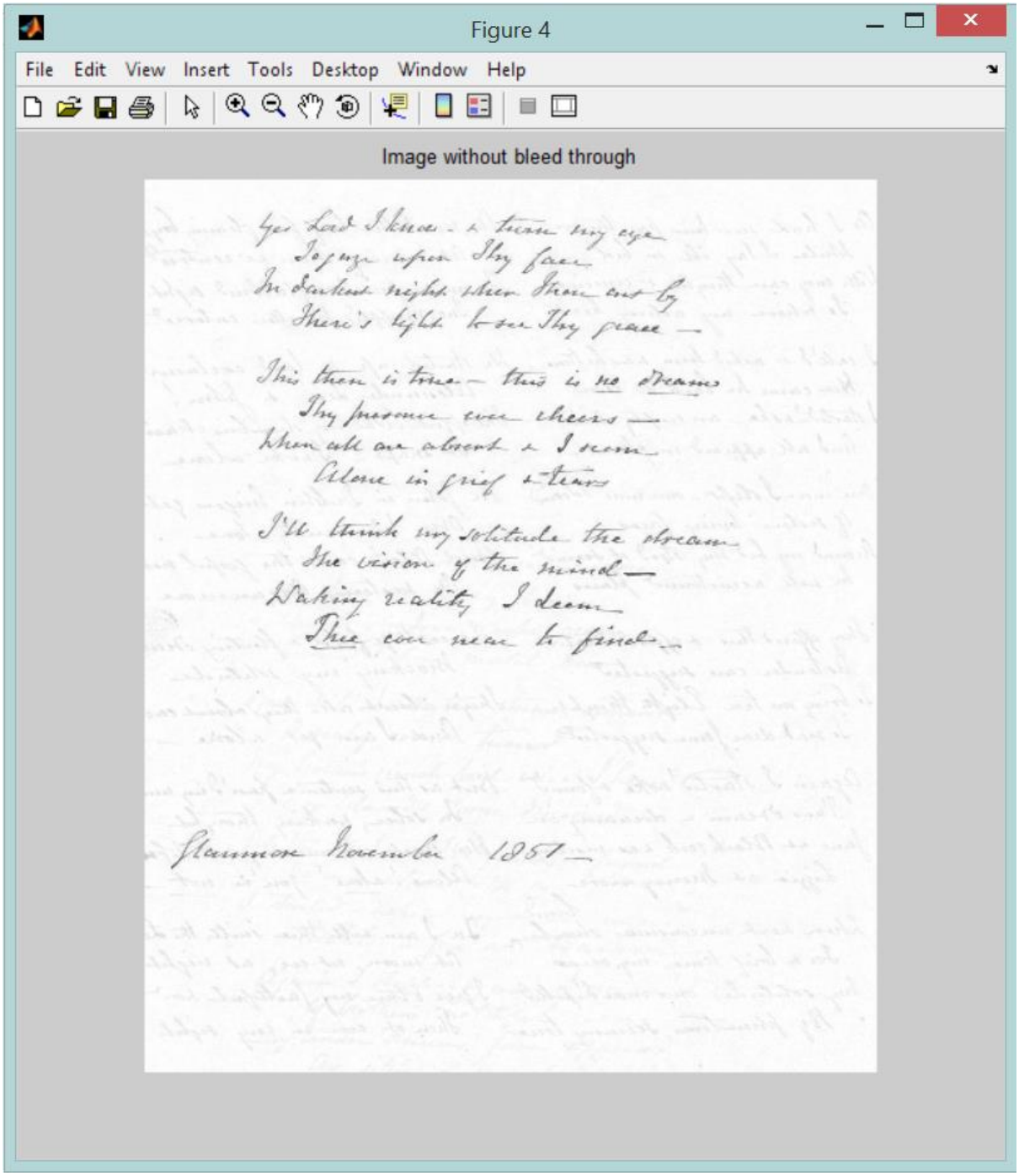


Figure 21: Image after light bleeding noise removal

The results are then displayed after removing the bleed through interference as shown in Figure 19.

Image belonging to class 2: Medium interference

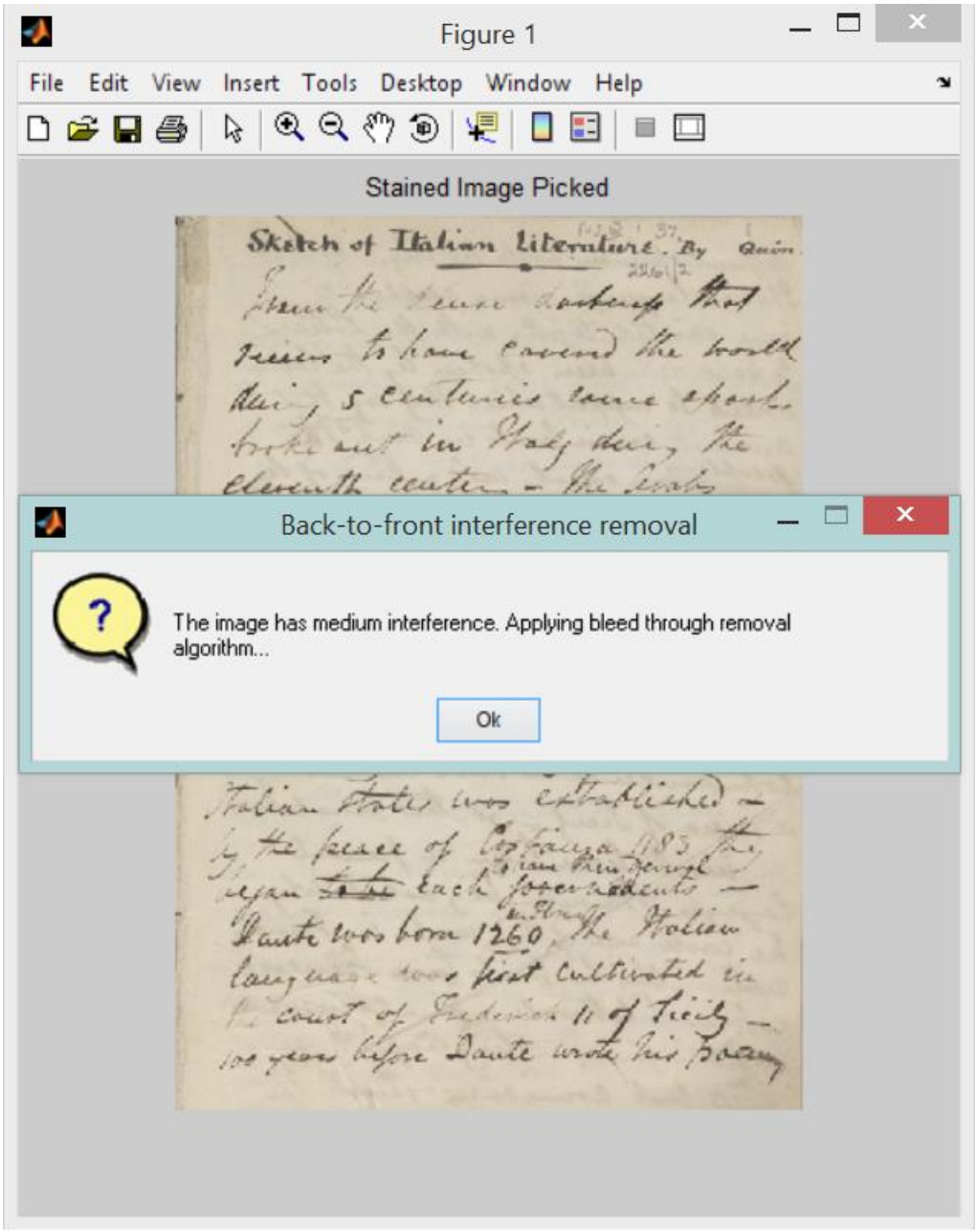


Figure 22: Image chosen by user with its class: medium interference

Another document chosen by the user which has medium bleeding noise. It is also removed by the proposed algorithm as shown in figure below.

Class 2 image after bleed through removal

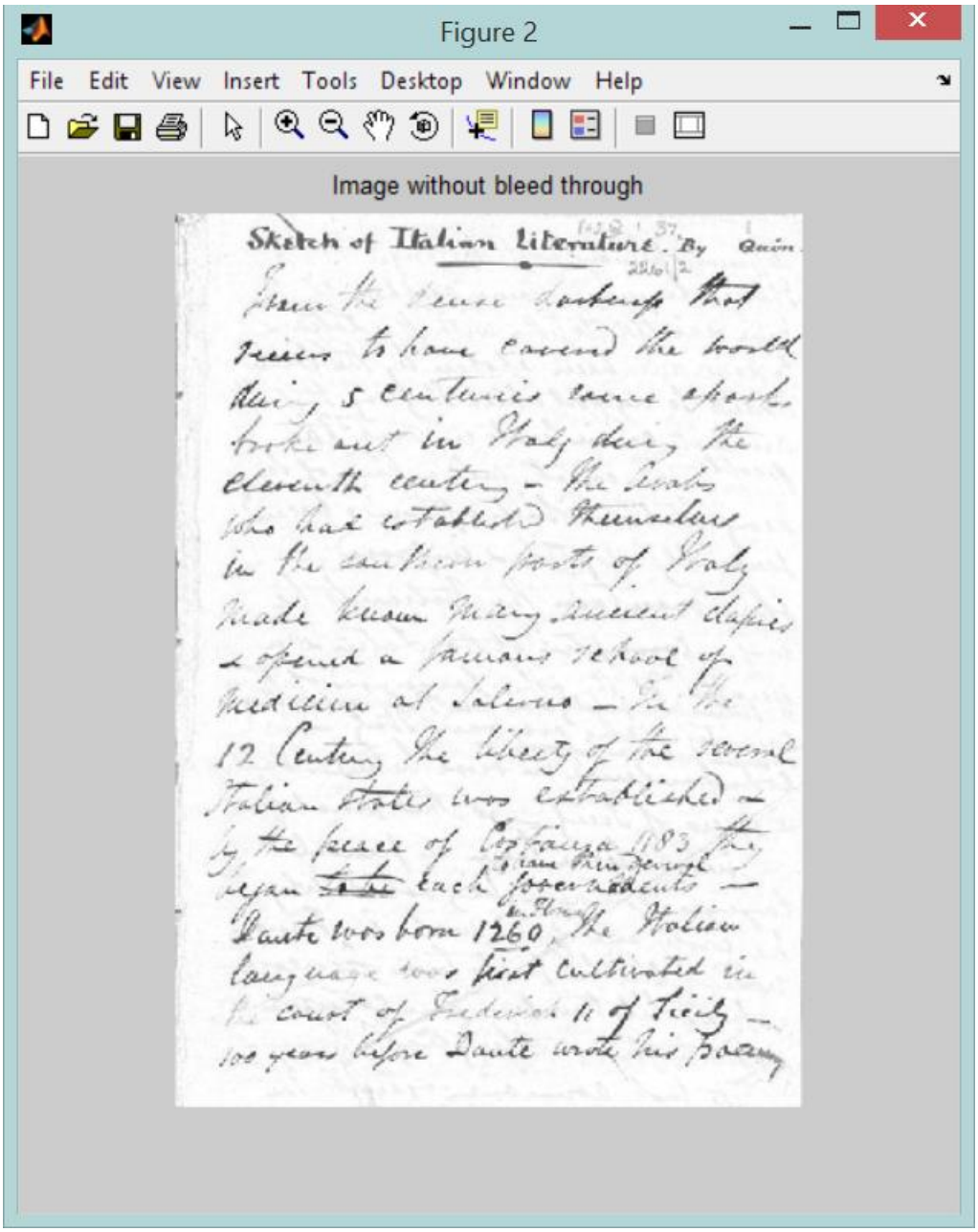


Figure 23: Image after medium bleeding noise removal

The algorithm works perfect for light and medium bleed through noise but doesn't work well for strong interference.

Image belonging to class 3: Strong interference

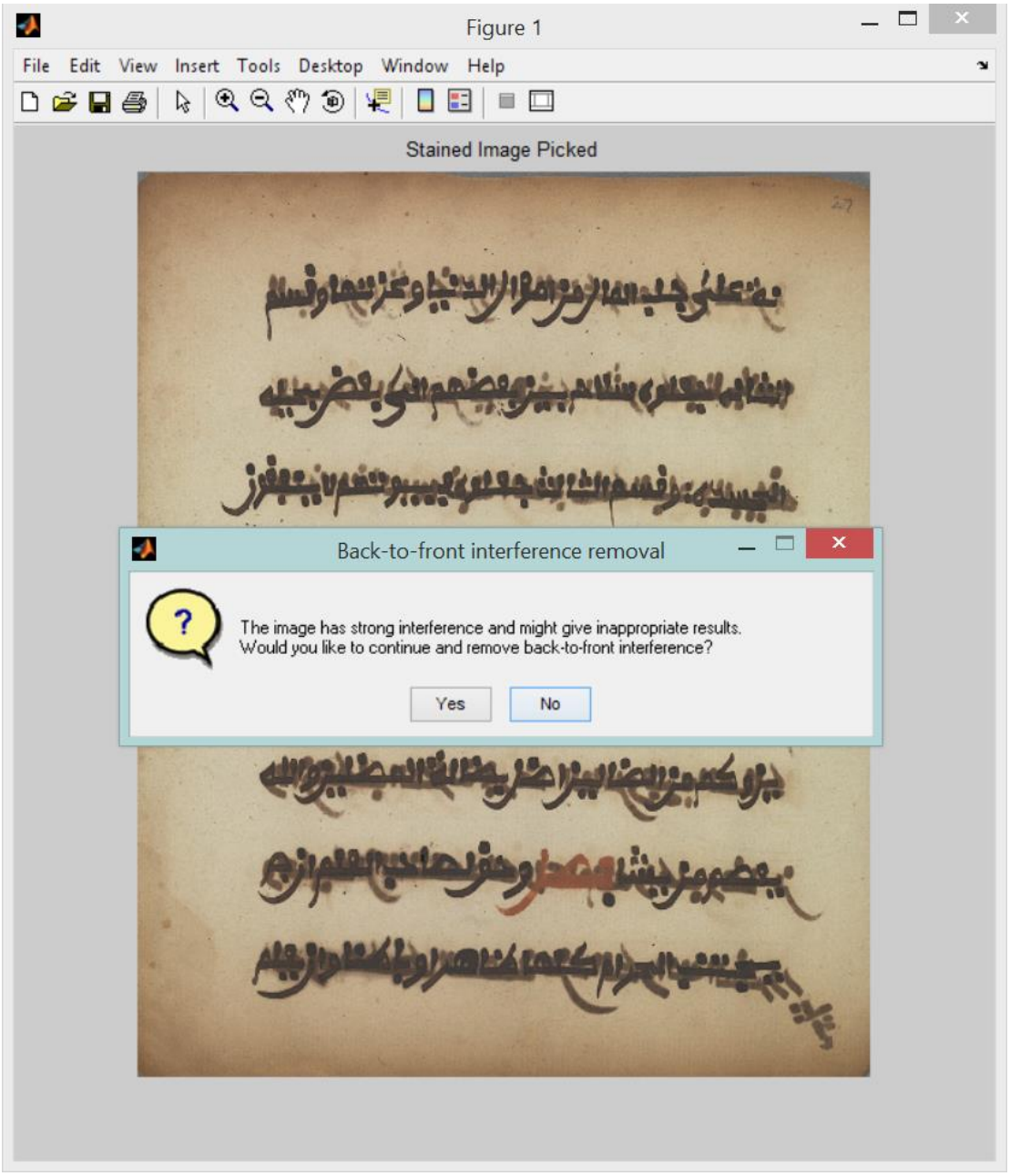


Figure 24: Image chosen by user with its class: Strong interference

If a document has strong back-to-front interference, it does not give desired results always. Therefore, a message box appears asking the user if he/she wants to continue applying the bleed-through removal algorithm.

Class 3 image after bleed through removal

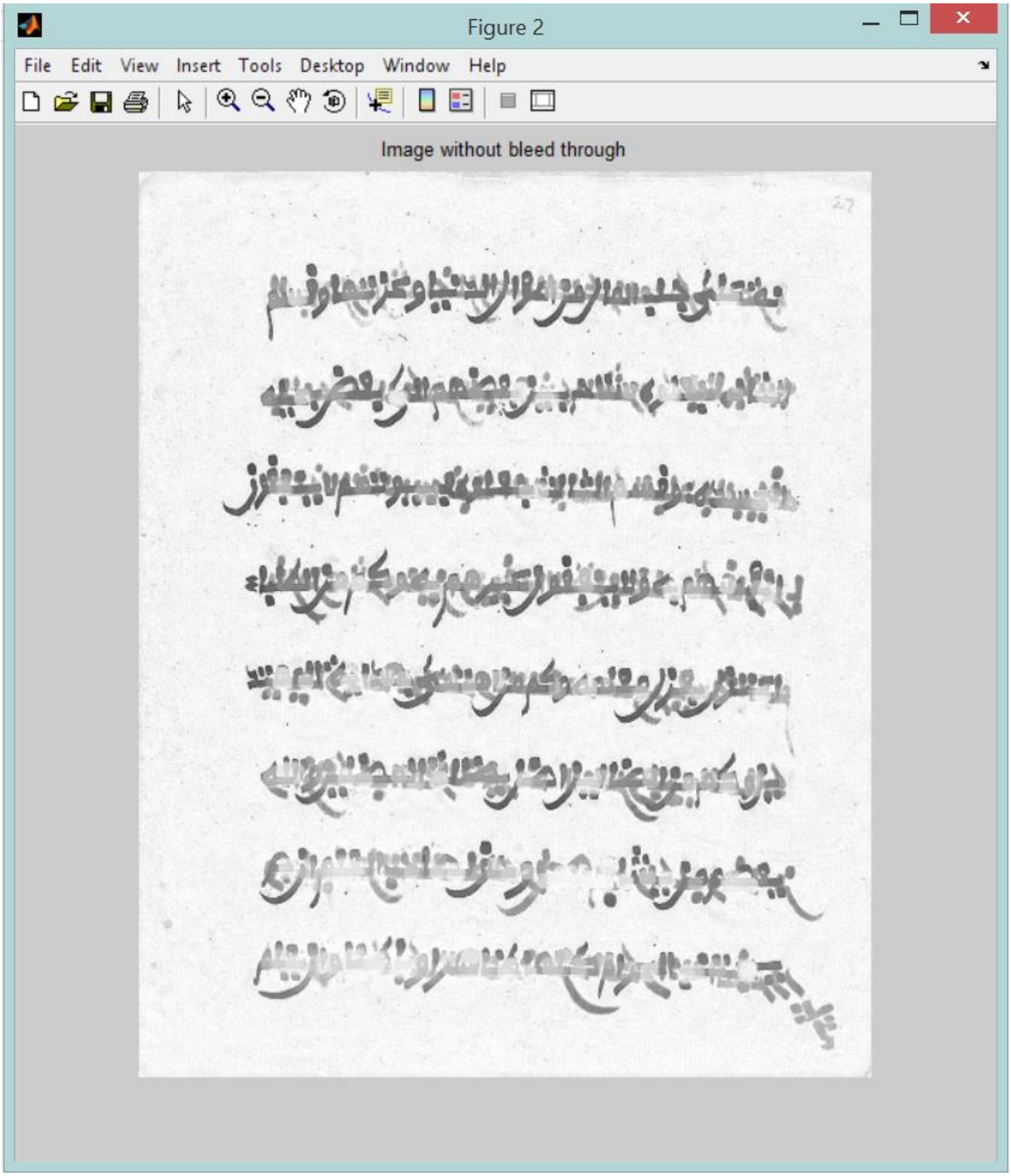


Figure 25: Image after removal of strong bleeding noise

If the user clicks on "yes" option in figure 23, results like the figure above might be displayed. It can be noticed that there are traces of ink from the back of the paper and the noise is not removed completely.

Prompt box when the user exits the program

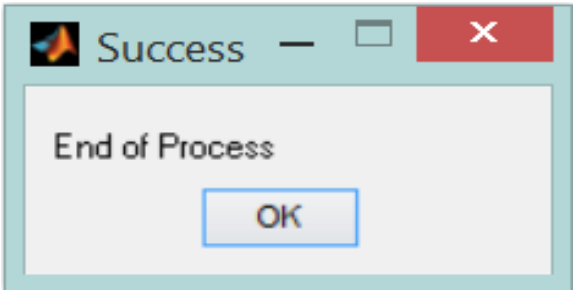


Figure 26: Message box at the end of process

A dialog box saying the process has completed appears if the user clicks on "no" option in figure 23.

Results on other images

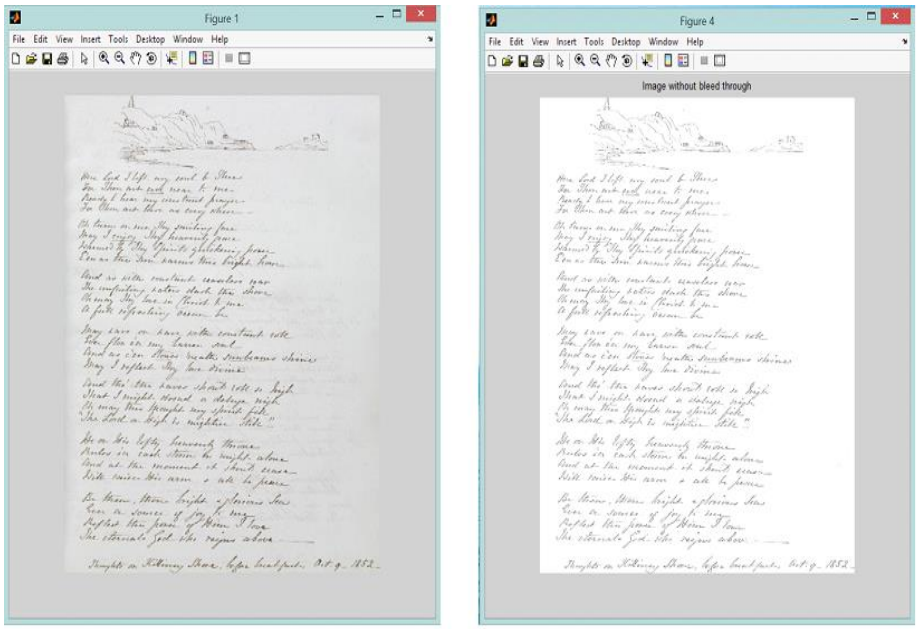
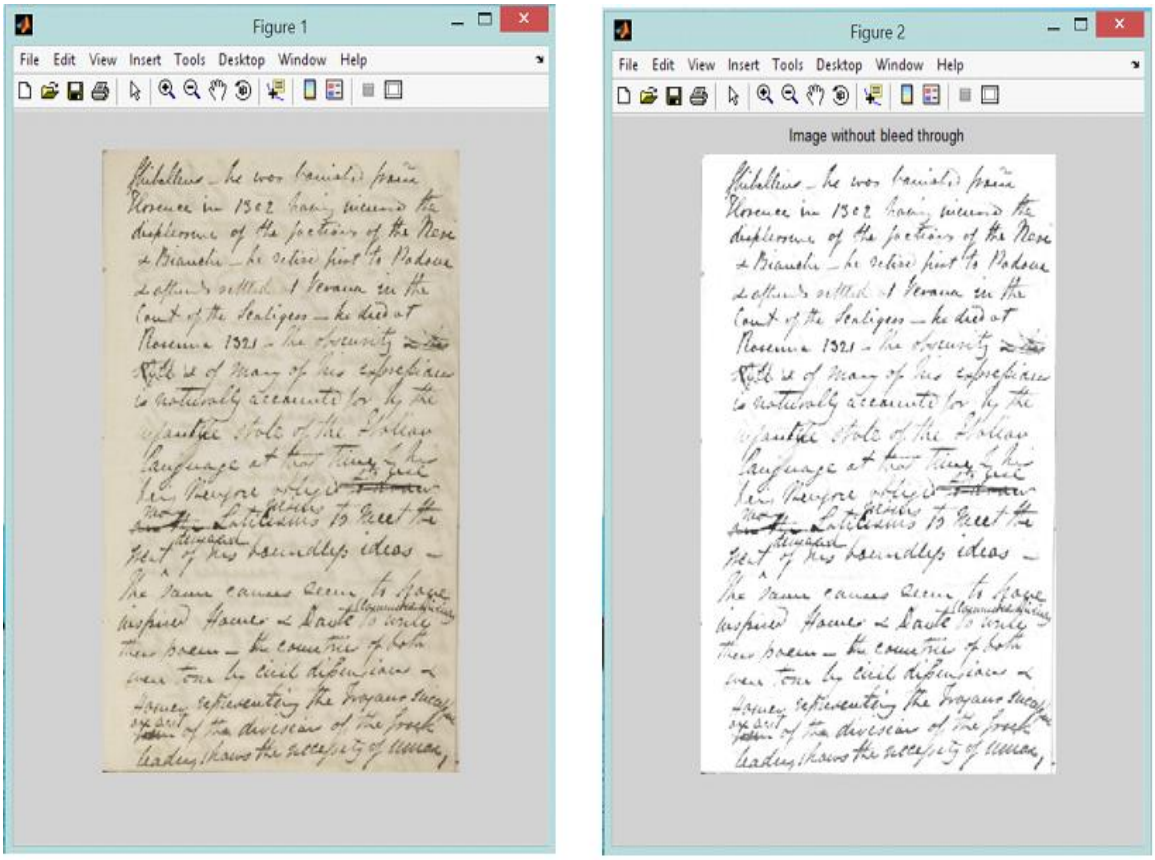


Figure 27: (a) Original image-1with bleed through noise (b) Result from Proposed Algorithm



(a) (b)

Figure 28: (a) Original image with bleed through noise (b) Result from Proposed Algorithm

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. A sample database of 200 images was used to test the algorithm.

Since, the proposed work uses a classification of the historical documents on the basis of the strength of bleeding noise the user has prior information regarding the type of the document. Based upon this information, the user can decide whether or not the bleed through algorithm should be applied to it. A hybrid combination of distance perception and intensity normalization is used to remove the bleeding noise in the historical documents. It is evident by visual comparison that the proposed algorithm gives better results than the existing algorithms for images with light and medium bleed through.

6.2. Future Scope

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

- It can be worked upon to remove noise in all types of historical documents.
- It is observed that the algorithm gives best results for light and medium bleeding noise strength but gives inappropriate results for strong noise.
- It can be improved to calculate the radius of the structural element i.e. disk automatically.
- It can be extended to recognize the characters in the historical documents.

References

[1] Arica N. and Fatos T., "An Overview of Character Recognition Focused on Off-Line Handwriting", IEEE Transactions on Systems, Man and Cybernetics-Part C Applications and Reviews, 31(2), 2001.

[2] Baird H.S., "State of the Art of Document Image Degradation Modelling," IAPR 2000 Workshop on Document Analysis Systems, Brazil, December 2000.

[3] Cai J. and Liu Z., "Integration of Structural and Statistical Information for Unconstrained Handwritten Numeral Recognition", IEEE Transactions on Pattern Analysis and Machine intelligence, 21 (3), 263-270, 1999.

[4] Cannon, M., Hockberg, J. and Kelly, P. (1999) Quality assessment and restoration of typewritten document images. International Journal of Document Analysis and Recognition, 2(2-3), 80–89, 1999

[5] Carlos A.B. and Lins R.D., "Image Segmentation of Historical Documents", 2000.

[6] Casey R. and Lecolinet A., "Survey of Methods and Strategies in Character Segmentation", IEEE Transactions on Pattern Analysis and Machine Intelligence, 18(7), 690-706, 1996.

[7] Clausner C. , Antonacopoulos A. and Pletschacher S., "A Robust Hybrid Approach for Text Line Segmentation in Historical Documents", International Conference on Pattern Recognition Tsukuba, Japan, pp. 335-338, 2012.

[8] DRIRA F.,"Towards Restoring Historic Documents Degraded Over Time", Proceedings of the Second International Conference on Document Image Analysis for Libraries, 2006

[9] Feldbach M. and Tonnies K.D., "Word Segmentation of Handwritten Dates in Historical Documents by combining semantic A-Priori-Knowledge with Local Features", Proceedings of the Seventh International Conference on Document Analysis and Recognition, 2003.

[10] Feng S.L. and Manmatha R., "Classification Models for Historical Manuscript Recognition", Proceedings of the 2005 Eight International Conference on Document Analysis and Recognition, 2005.

[11] Fornes A., Otazu X. and Llados J., “Show-through Cancellation and Image Enhancement by Multiresolution Contrast Processing”, International Conference on Document Analysis and Recognition, 2(1) , 323-327, 2013.

[12] Garz A., Fischer A., Sablatnig R. and Bunke H., “Binarization-Free Text Line Segmentation for Historical Documents Based on Interest Point Clustering”, IAPR International Workshop on Document Analysis Systems, 3(2) , 301-306, 2012.

[13] Gatos B., “Segmentation of Historical Handwritten Documents into Text Zones and Text Lines”, 14th International Conference on Frontiers in Handwriting Recognition, 2014

[14] Ha T. M. and Bunke H., “Off-Line, Handwritten Numeral Recognition by Perturbation Method”, IEEE Transactions on Pattern Analysis and Machine Intelligence, 19 (5), 535-539, 1997.

[15] Joao M. and Lins R.D.,”Binarizing and Filtering Historical Documents with Back-to-Front Interference, 2006.

[16] Johannsen G. and Bille J., “A threshold selection method using information measures”, ICPR’82: Proc., 6th International Conference on Pattern Recognition, 140–143, 1982

[17] Kale P., Gandhe S.T., Phade G.M. and Dhulekar P.A., “Enhancement of old images and documents by Digital Image Processing Techniques”, International Conference on Communication, Information &Computing Technology, 2015

[18] Kapur J.N., Sahoo P.K. and Wong A.K.C., “A New Method for Gray-Level Picture Thresholding using the Entropy of the Histogram”, Computer Vision, Graphics and Image Processing, 29, 273-285, 1985.

[19] Kasturi R., Gorman L. and Govindaraju, V. “Document image analysis: A primer. Sadhana”, No. 27, 3-22, 2002.

[20] Makridis M. , Nikolaou N. and Gatos B., “An efficient word segmentation Technique for Historical and Degraded Machine Printed Documents”, Ninth International Conference on Document Analysis and Recognition, 2007

[21] Mello C.A.B and Lins R.D., “Generation of images of historical

documents by composition", ACM Document Engineering McLean, VA, USA,2002

[22] Mello C.A.B., "Segmentation of Images of Stained Papers Based on Distance Perception", IEEE, 2010

[23] Mello C.A.B., "Synthesis of Images of Historical Documents for Web Visualization", Proceedings of the 10th International Multimedia Modelling Conference, 2004.

[24] Mo S. and Mathews V. J., "Adaptive, quadratic preprocessing of document images for binarization", IEEE Transaction Image Processing, (7), 992–999, 1998.

[25] Otsu N., "A Threshold Selection Method from Gray-Level Histograms", IEEE Transactions on Systems, Man and Cybernetics, 9(1),1979

[26] Panwar S. and Nain N., "A Novel Approach of Skew Normalization Text Lines and Words", International Conference on Signal Image Technology and Internet Based Systems, 8(1), 2012.

[27] Pun and Thierry, "Entropic thresholding", Computer Graphics and Image Processing, 16(3), 210-239, 1981.

[28] Rao N.V., Rao A.V.S., Balaji S. and Reddy L. P., "Cleaning of Ancient Document Images Using Modified Iterative Global Threshold",International Journal of Computer Science Issues, 8(6), 128-133, 2011.

[29] Silva G.P., Lins R.D., Silva J.M., "HistDoc - A Toolbox for Processing Images of Historical Documents",7th International Conference, ICIAR, Vol 6112, 2010, pp 409-419, 2010.

[30] Slavik P. and Govindaraju V., "Equivalence of Different Methods for Slant and Skew Corrections in Word Recognition Applications", IEEE Transactions on Pattern Analysis and Machine Intelligence, 23(3), 2001.

[31] Vamvakas G., Gatos B., and Peratonis S.J., "A Complete Optical Character Recognition Methodology for Historical Documents", The Eighth IAPR Workshop on Document Analysis Systems,2008.

[32] Yen J.-C., Chang F.J. and Chang S.,"A New Criterion for Automatic Multilevel Thresholding", IEEE Transactions on Image Processing, 4(3), 1995.

[33] Yosef I.,Kedem K. and Dinstein I., “Line segmentation for degraded handwritten historical documents”, 10th International Conference on Document Analysis and Recognition,2009

[34] Hagit (15-Jan-2006 14:13). [Online]. Available: http://cs.haifa.ac.il/hagit/courses/ip/Lectures/Ip12_Segmentation.pdf [Accessed: Novemeber 28, 2014].

Video Presentation

The video presentation of the thesis presented herein is available at the following link.
Available at: https://youtu.be/0BifzwfZ4oQ