Binarizing Degraded Document Image for Text Extraction

by

Radhika Patel

201311004

A thesis submitted in the partial fulfillment of the requirements for the degree of

Master of Technology

in

Information and Communication Technology

to

Dhirubhai Ambani Institute of Information and Communication Technology

Gandhinagar, India

May 2015

Declaration

This is to certify that

a. the thesis comprises my original work towards the degree of Master of Technology in Information and Communication Technology at DA-IICT and has not been submitted elsewhere for a degree.

b. due acknowledgment has been made in the text to all other material used.

Signature of Student

Radhika Patel

Certificate

This is to certify that the thesis work entitled *Binarizing Degraded Document Image for Text Extraction* has been carried out by *Radhika Patel (201311004)* for the degree of Master of Technology in Information and Communication Technology at this Institute under my supervision.

Thesis Supervisor

Prof. Suman K. Mitra

Acknowledgements

First of all I would like to thank my supervisor Prof. Suman K. Mitra from the bottom of my heart for giving me insight in the field of Image Processing, sharing large experience of research and motivating for work when I need it and being there for me to help out on the occasion where I got stuck, even in his very busy schedule.

I am thankful to my examiner Prof Anil Roy, Prof. Asim Banerjee, Prof. Bhaskar Chaudhury and Prof Ranendu Ghosh (all from DA-IICT) who gave me valuable suggestion and feedback to improve my work and encouraged me to strengthen my basic knowledge in the field of my research and inspired me to pursue root level study. Especially I would like to thank to Ms. Purvi koringa (Ph.D student at DA-IICT) who is my 'MATLAB help, grammar ready reference and being skeptical for my idea (and in turn, improving). I would also like to thank Mr. Ashish Phopaliya for Rough set based edge detection code and Ms. Gitam Shikkenawis (Ph.D student at DA-IICT) for suggestions and valuable help.

Besides them, I would like to express my gratitude to the authors of all the books and publications, which helped me to enhance my knowledge and understanding about my work.

I am indebted to my family for their trust in me that made me confident enough to accomplish my work. Their support in a number of ways must also be acknowledged here.

I would like to thank my friends for their continuous support, trust and encouragement. I would also like to mention my deepest thanks to my friends Hiral vasani, Prashant damodiya, Rachit chhaya for giving inspiration for work and suggestion and all my M. Tech Batch mates through their constant support made my stay at the institute worth remembering.

Radhika Patel

Abstract

The recent era of digitization is expected to be digitized many old important documents which are degraded due to various reasons. Binarizing Degraded Document Image for Text Extraction is a conversation of document color image to binary image. Document images have mostly two classes: background and text. It can also be considered as a text retrieval procedure as it extracts text from a degraded document. Degraded document image binarization have many challenges like huge text intensity variation, background contrast variation, bleed through, text size or stroke width variation in a single image, highly overlapped background and foreground intensity ranges etc. Many approaches are available for document image binarization, but none can handle all kind of degradation at the same time. Mostly, a combination of global and/or local thresholding along with various preprocessing as well as postprocessing techniques are used for document image binarization to handle most of the challenges. The approach proposed in this thesis is basically divided into three stages: preprocessing, Text-Area detection, post-processing. Preprocessing employs PCA to convert image from RGB to Gray, followed by gamma correction that enhances the contrast of the image. Contrast-enhanced image is filtered with DoG (Difference of Gaussian) filter to boost local features of a text, followed by equalization. Next stage involves identifying Text-Area. A Rough set based edge detection technique is used to find closed boundary around texts, which results into locating Text-Area along with some non-text area detected as text. Text is detected by applying logical operators on preprocessed image and edge detected image. Postprocessing technique takes care of false positives and false negative based on intensity values of preprocessed and gray image. The algorithm is also expected to be independent of the script. To demonstrate this, the algorithm is tested on Gujarati degraded document images. The Performance is evaluated based on various quantitative measures like Distance Reciprocal Distortion (DRD), Peak Signal-to-Noise Ratio (PSNR), F-Measure, and pseudo F-measure and It is compared with the state-of-the-art (SOTA) method. The proposed approach is close to the SOTA methods based on performance. It is able to binarize without losing text in some of the very challenging images, where state-of-the-art methods lose the text.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Old historical documents need to be digitized to preserve the information stored in them. These documents are scanned as a color image or captured using a camera. Digitized document images can be used for text retrieval, image analysis or optical character recognition. With the boom of hand-held devices like kindle and i-pads, people prefer to have all text data in digital form. Image binarization is also an essential stage in automatic language translation system form documents. Binarization of digitized documents leads to less storage requirement, as the binarized document image needs only one bit to represent one pixel, either 1 (white for background) or 0 (black for text).

Image binarization is a process of converting input color image into the binary image. Image binarization can be done using various approaches, Segmentation and thresholding are intensity-based methods; morphological methods uses the shape of the object to binarize images along with other segmentation methods [1]. Segmentation and thresholding can be explained as a partition of an input image into two class based on its pixel value in an input image. Morphological operators operate based on neighborhood set information to decide foreground from background.

Document image binarization is generally a preprocessing stage for optical character recognition and image analysis. It is also useful for image compression and data mining, data entry tasks. Document binarization is employed wherever text retrieval is necessary for eg, application like information retrieval from a text document, language translation of

the text, digitizing old books etc.

1.1 Motivation and Problem Statement

Binarization of a degraded document have an important role in recognition systems and document analysis systems. It is a primary step for many document analysis systems. Most of the document analysis algorithm and Optical Character Recognition (OCR) technique works on binarized images. The use of bi-level information present in binarized image decreases the computational load and enables the utilization of the simple analysis methods compared to gray-scale or color image information. A fast and accurate document image binarization technique is important for accurate performance of next consecutive document image processing tasks such as OCR. The recent Document Image Binarization COntest (DIBCO2014), held under the framework of International Conference on Frontiers in Handwriting Recognition (ICFHR) 2014 [2], particularly addresses this issue by creating a challenging benchmarking data set and evaluating the recent advances in document image binarization.

1.1.1 Objectives

Though many researcher have proposed various approaches to binarize degraded documents since many years, still there is scope of improving efficiency. This can be explained by the challenges in different types of document such as spots in the image, bleeding-through, text stroke width variation, stroke intensity variation, document background with uneven illumination, huge image contrast variation, smear etc. exist within many document images. These document degradations generate the document thresholding error in image binarization. A big challenge to most state-of-the-art techniques is to develop a automatic, accurate, and fast algorithm for document image binarization of all type of degraded document images. Binarize document such that all texts are cured in output binarized image.

1.2 Philosophy of Proposed Approach

While binarizing a document for OCR, it is important to preserve maximum possible text in output binarized image. The Proposed approach solves this problem. The proposed approach of degraded document binarization can basically be divided in three stages: preprocessing, text detection, postprocessing. This approach tries to retrieve as much as possible text even from degraded area regardless of a type of degradation present in an image.

Figure 1.1: Layout of the problem on which current thesis is based

It is observed that these degraded images have highly overlapped text and background intensity range, thus hard thresholding can not separate the text from the background. Thus proposed approach concentrates on the philosophy of edge detection, edge between text and background generally have high-intensity difference. However, due to degradations present in old document images the edges do not have very high intensity differences, to improve the performance of edge detection (in our approach closed boundary around text) the contrast of image is improved. A series of several image processing technique is used as a preprocessing stage, Which involves RGB to gray conversion using PCA, gamma correction, DoG to improve local features, and equalization. The chain of these processes increases the local contrast in an image, in turn enhancing text, which is useful in edge detection.

The approach detects possible text area in preprocessed image and further decides to classify the pixel as either foreground or background. To identify possible text area, edge detection based on rough set used to exploit the nature of giving closed boundary around an object (here, text). Postprocessing technique is developed to remove the noise (here, false positives and false negatives) to improve overall binarization performance.

Chapter 2

Literature Survey

2.1 Document Image Binarization COntest

Document image binarization contest (DIBCO) is world wide competition, the Competition is aimed to invite new methodologies in document image binarization. A challenging data set of degraded document images with ground truth is provided by competition organizer every year for competitors. Data set include printed as well as a handwritten text document. Various degradations are available in the data set. DIBCO 2009 was organized with ICDAR'09, H-DIBCO 2010 was organized with ICFHR 2010, DIBCO 2011 was organized with ICDAR'11, H-DIBCO 2012 was organized with ICFHR 2012, DIBCO 2013 was organized with ICDAR 2013 and H-DIBCO 2014 was organized with ICFHR 2014 [3]. Evaluation tool [4] is also provided to evaluate the efficiency of algorithms, submitted by a various competitor. Winner of the competition is decided based of performance measures like F-measure, pseudo-F-measure, Peak Signal to Noise Ratio (PSNR) and Distance Reciprocal Distortion (DRD) [5].

2.2 Data Set and Challenges in Data Set

DIBCO 2009, DIBCO 2010, DIBCO 2011, DIBCO 2012, DIBCO 2013 and DIBCO 2014 data set are available to assess performance of newly developed algorithms [6]. Data set

include printed as well as handwritten text documents. Many different type of degradations and challenges are included in data sets like bleed through, spots in image, huge text intensity variation, background contrast variation, water marks, text size variation, stroke width variation etc. as shown in figure 2.1.

(a) (b)

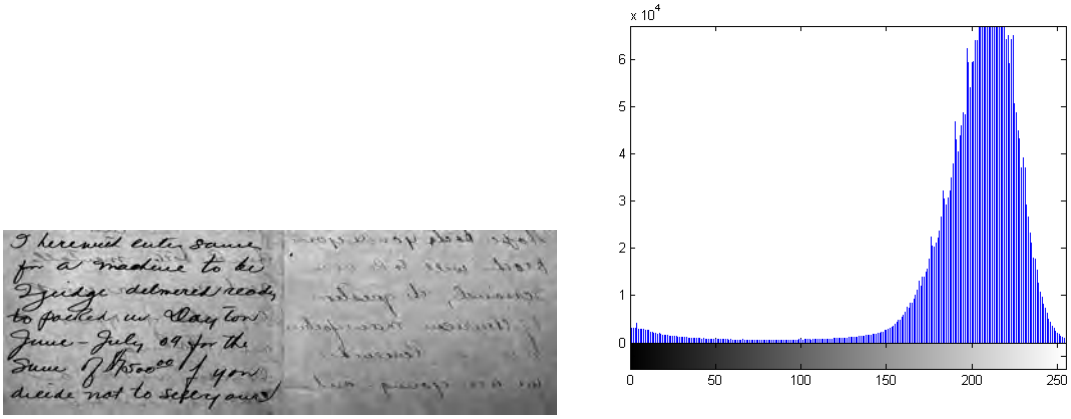
(c) (d)

(e) (f)

Figure 2.1: Degradations and challenges in data set,(a) Bleed through image, (b) Spot in image, (c) Text intensity variation, (d) Stroke width variation, (e) Image contrast variation, (f) Font size variation

- Binarization is basically two-class segmentation problem, the trivial case of binarization is images having clear valley to separate foreground and background in its histogram. However, analysis of images from these data sets revealed that there is no clear valley in the histogram such that it separates text from the background. One of

such histogram is shown in below figure 2.2.



(a) (b)

Figure 2.2: (a)Gray Image, (b) histogram of image (a)

- Figure 2.2 supports previous claim, even if there seems a clear valley in the histogram 2.3(b) of image in figure 2.3(a), the intensity value at valley can not be used as a hard threshold value, because the intensity range of foreground and background highly overlaps each other as can be seen in figure 2.3(c) and 2.3(d) respectively. Performing hard thresholding based on valley will result in segmenting text and background with large error. Text and background pixel position are taken from ground truth image (dibco2013 data set with ground truth images [6]).

- In figure 2.3(b) valley is present at approximately 135, the text intensity range from 0 to 200 figure 2.3(c) and the background intensity range from 25 to 250 figure 2.3(d) . Thus, valley intensity 135 is not useful to segment the text from background.

2.3 Related previous Works

Various thresholding techniques are reported in the literature. Two major approach of binarization is global thresholding and local thresholding. Global thresholding [7, 8, 9] efficiently detect text if degradation in images is not intense or if there is exist a clear valley

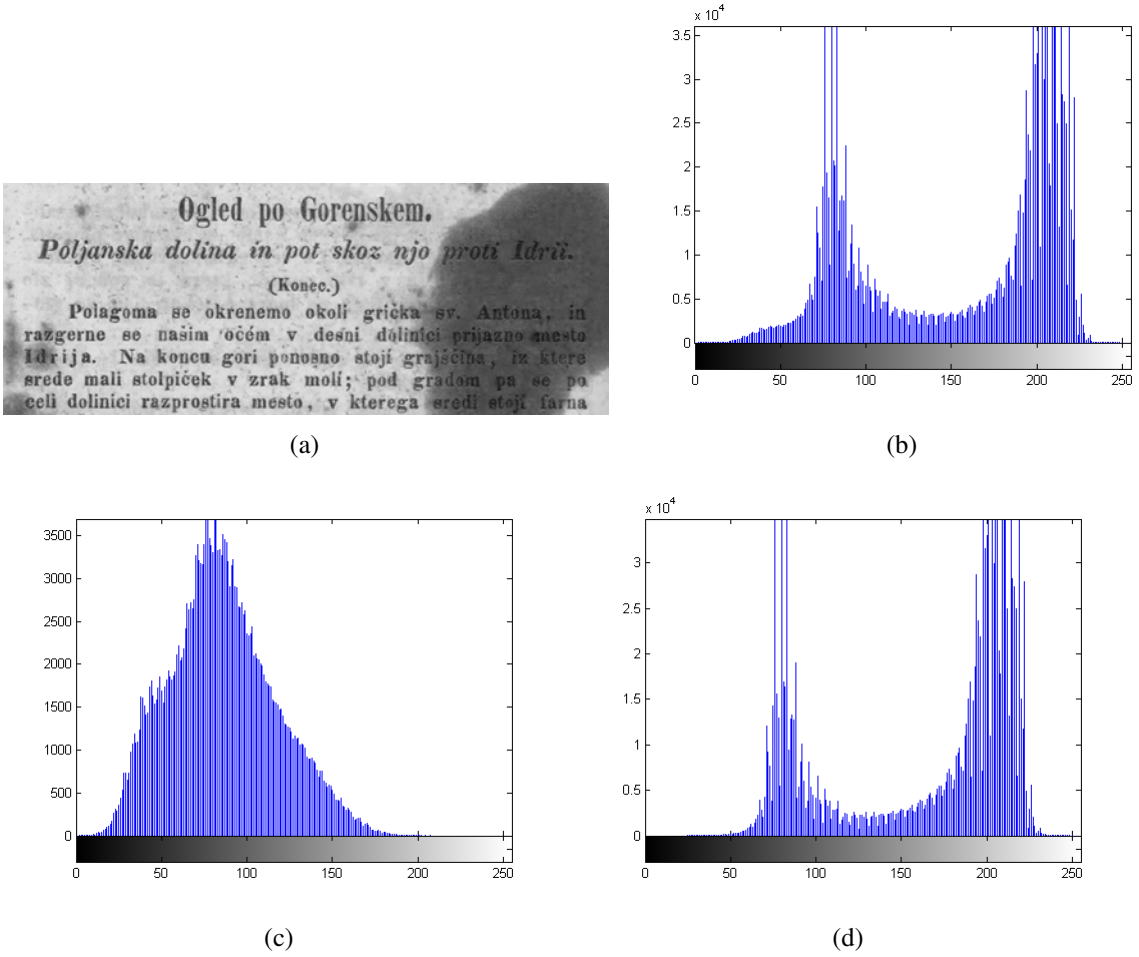


Figure 2.3: Histogram of text and background in grayscale image, (a) Gray Image, (b)Histogram of image (a), (c) Histogram of text intensity, (d) Histogram of background intensity

between text and background in the histogram of an image. Some of the approaches use adaptive thresholding techniques [10, 11, 12, 13, 14], a performance of such techniques is dependent on window size. In some approaches employs edge detection techniques to segment text from background [15].

In adaptive thresholding technique proposed by Sauvola et al. [12] firstly the image is classified into background, pictures and text. Two different approaches are used for thresholding each of this area, namely background, pictures, and text. A soft decision method (SDM) for background and pictures, and a specialized text binarization method (TBM) for textual and line drawing areas is used. The SDM includes noise filtering and

signal tracking capabilities while the TBM is used to separate text components from the background in bad conditions. Finally, the outcome of these algorithms are combined, this approach performs well almost in all kind of images except the images having high local contrast.

Some adaptive thresholding techniques are developed specifically for document binarization using domain knowledge. Gatos et al. [16] developed an algorithm having four stages for binarization; (i) applying wiener filter, (ii) rough estimation of foreground regions and background surface, (iii) thresholding decision based on rough estimation and preprocessed image and (iv) a postprocessing to improve the quality of text regions and preserve stroke connectivity.

Shijian Lu et al, [15] proposed an approach in 2009 DIBCO competition, It first estimates a document background surface through an iterative polynomial smoothing procedure that compensate degradations due to smear and shading, followed by detecting stroke edges, extracting text and postprocessing. This method won DIBCO 2009 with average 91.24 $\%$ F-measure result for DIBCO 2009 dataset. Shijian Lu et al, also proposed new approach in 2010 DIBCO competition in [17, 18], It includes four main steps, (i) image contrast which is evaluated by local maximum and minimum is used to select the high contrast pixels. (ii) the stroke edges which are extracted using Canny's method are combined with those high-contrast pixels to produce a better edge map. (iii) the document image is binarized using a local threshold decided based on the constructed edge map. (iv) postprocessing is applied to improve the final result. This method won DIBCO 2010 with average 91.50 $\%$ F-measure result for DIBCO 2010 dataset.

Shijian Lu et al, also proposed a different approach for binarization in 2012 in [19] considering the problem as the clustering problem, in a first run, various existing binarization techniques are used to segment the image into foreground and background. Based on these binarized results, each candidate pixel is classified as foreground (true positive) or background (true negative) or uncertain (can be either foreground or background). in second stage, certain features are extracted based on the contrast of pixels present neighborhood window and intensity value of the candidate pixel to be classified. The label decision is

taken based on these feature value as well as certainly labeled pixels enclosed within that window. The labeling procedure of uncertain pixels is iterative if the window does not have enough certainly labeled pixels, the decision is postponed till next iteration.

Nicholas R. Howe proposed a new approach in [20] based on energy segmentation using laplacian of image. It finds a darker area and lighter area using laplacian of image. It uses canny edge detection for finding discontinuities and graph cut method based binarization for a continuous area. This method won DIBCO 2012.

Parker et al. [21] proposed robust binarization of degraded document images using heuristics. The method is based on two guiding principles: (i) "Writing" pixels should be darker than nearby "non-writing" pixels, (ii) "Writing" should generate a detectable edge. Firstly, PCA based approach is used to convert RGB image into gray scale image, in gray image (a) locally dark pixel and (b) identify pixel that are near an edge are identified using Sobel edge detection mask, followed by a bilateral filter to reduce noise. The Final stage involves intersecting of (a) and (b), and cleaning to improve binarization result.

Bill Trigs et al. proposed a preprocessing technique that is able to handle shadow effect in face images in [22]. In this paper a chain of basic image processing techniques is employed to remove the shadowing effect due to a texture of the face image, the techniques involved are gamma correction, DoG filtering, and equalization.

Munshi at al. also proposed a preprocessing technique for fingerprint images [23]. The paper uses rough set based binarization technique base on rough set theory proposed by Pawlak [24] to get closed boundaries around different objects present in an image and the method performs better than Canny edge detection [25]. Image is decomposed into several blocks such that each block has at least 90 its pixels belonging to the same range of intensity value.

Chapter 3

Proposed Approach

The Proposed approach of document binarization can basically be divided in three stages: preprocessing, text detection, postprocessing. This approach tries to retrieve as much as possible text even from degraded area regardless of a type of degradation present in the image.

Scanned degraded document images are of RGB form. To reduce three-fold computation on RGB image, it is converted in the gray scale image. RGB to gray conversation done using PCA based method as discussed in section 3.1 in detail. Preprocess increase local contrast in the image, in turn, enhancing text, which is useful in edge detection. Preprocessing involves several image processing techniques as explained in section 3.2.

Preprocessed image have high contrast and suitable for edge detection, rough set based edge detector is used to take advantage of its nature of giving closed boundaries around separate classes. These closed regions are filled and considered as ROI, i.e. Text-Area. The region filling procedure also fills closed region of text, which are unwanted false positives in our case. The procedure is described in section 3.3. Text detection is performed on ROI based on intensity values in preprocessed image as explained in section 3.4.

The output of the text detection stage gives text (Foreground) as well as noise regions, say 'Black blob' (false positive due to closed shape within alphabets) and 'White blob' (false negative), and non-text border just around the text which are mostly 1 pixel wide. Border removal, black and white blob removal, and small object removal are applied to

solve this problem. The postprocessing method is explained in section 3.5.

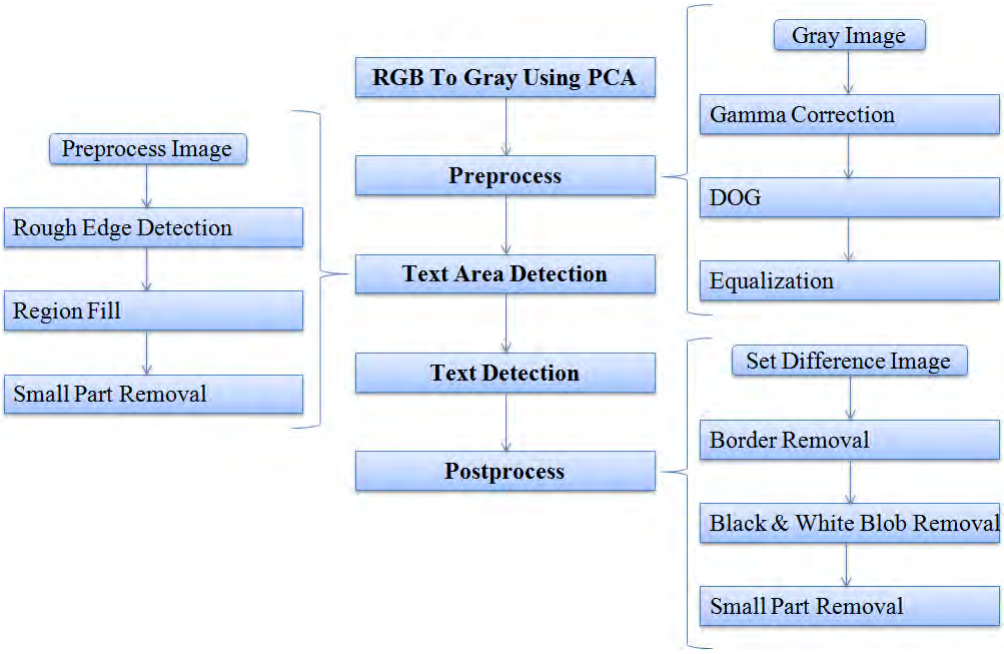


Figure 3.1: Flow chart of proposed approach

3.1 RGB to Gray Conversion

Document images are scanned or camera photo images so they are in RGB format or any color model format. Work with the three-dimensional data set is computationally complex, and time-consuming. First input document image is transformed in gray scale level using Principal Component Analysis (PCA) [26] based approach [21]. PCA is generally used for dimensionality reduction based on the eigenvalue of the feature set matrix. Here normalized eigenvalue is used to decide weight for each of the R, G and B color plane.

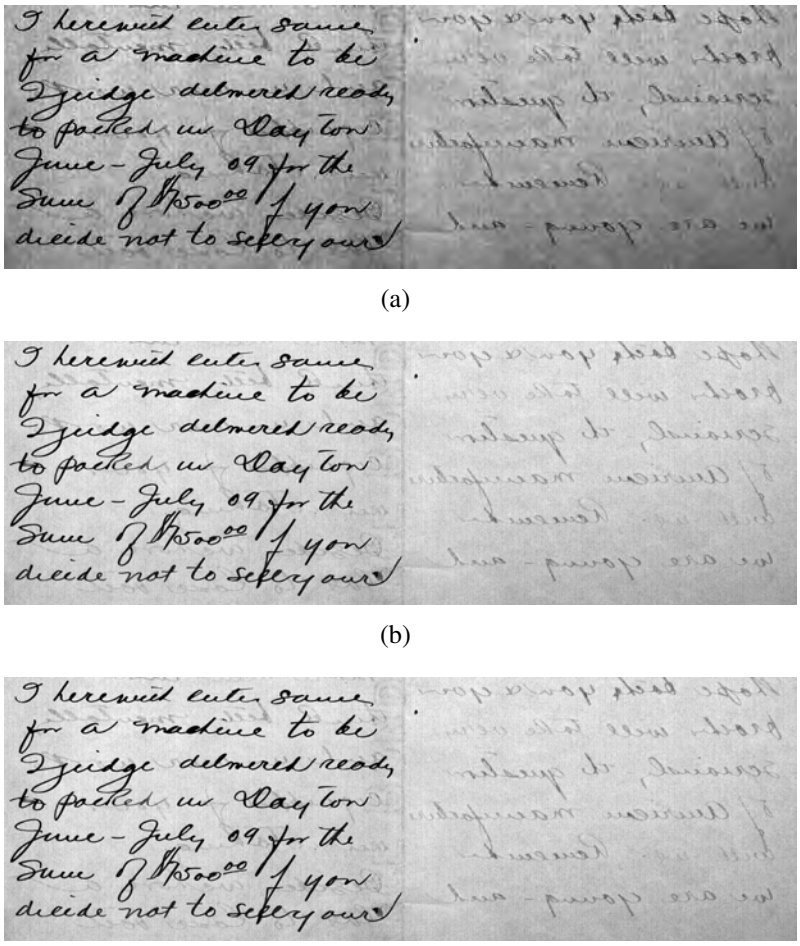
Standard formula for RGB to Gray conversion is $0.2989 * R + 0.5870 * G + 0.1140 * B$ [27]. Where R, G, and B are intensity values of the red color plane, green color plane, and blue color plane respectively. Here, assignment of weight is dynamic based on R, G and B component of individual image unlike the static weight assignment of standard conversion formula.

PCA basically seeks for the basis such that the transformed data have maximum covariance in the direction of first principal component and so on. The components are found by solving maximization problem, resulting in eigenvalue problem $X^T Xv = \lambda v$, where, X is the covariance matrix of data, λ= eigenvalue and v= basis vector. This philosophy of PCA is used to find the weight based on the variance of R, G, and B plane. Firstly, the R, G and B plane of size $m \times n$ are vectorized and placed as a column to create a matrix X of $mn \times 3$. Find eigenvalues of matrix $X^T X$. Give normalized eigenvalue as a weight to R, G and B component. Based on PCA, it can be assumed that highest eigenvalue corresponds to the color plane having the highest variance and likewise. The variance of an individual plane is calculated and the accordingly normalized eigenvalue is used to weigh the corresponding color plane in gray conversion. It is observed that in most of the document images, blue color plane have highest contrast and thus have the larger contribution in the gray image. Results based on Standard formula and proposed method are as shown in figure 3.2 here, difference in contrast but it is not perceivable with naked eyes.

Root Mean Square(RMS) contrast [28] of image is high in PCA based gray scale image compare to standard formula based gray scale converted image. Formula for RMS contrast is, $\sqrt{\frac{1}{MN}\sum_{i=0}^{N-1}\sum_{j=0}^{M-1}(I_{ij}-\bar{I})^2}$. where I_{ij} is intensity value of pixel position (i, j) and $\bar{I}$ is mean intensity value. Image size is $M \times N$.

3.2 Preprocessing

Document images are degraded, removing the degradation or suppressing the effect of degradation will improve the results of binarization. The approach proposed here uses preprocessing technique described in [22], which consists a chain of basic image processing techniques to improve local contrast and suppressing the noise from background texture in order to efficiently detect ROI in next stage. The proposed approach uses following chain of techniques.



(a)

(b)

(c)

Figure 3.2: Results of RGB to gray, (a) Input RGB image, (b) Result of RGB to Gray using standard formula based method, (c) Result of RGB to Gray using PCA based method

3.2.1 Gamma Correction

Gamma Correction is contrast enhancement technique [1]. It is nonlinear gray-level transformation that replaces gray-level I with I^{γ}(for $\gamma > 0$) or $\log\ I$ (for $\gamma = 0$) where $\gamma\,\epsilon\,[0, 1]$ is a user defined a parameter. It enhances the local dynamic range of the image in dark while compressing the range of bright regions and highlights. The intensity transformation curve is shown in figure 3.3.

Gamma value (γ) in the range $[0, 0.5]$ is a good adjustment. For given text documents, experiments were carried out with different γ values results are shown in figure 3.4. $\gamma = 0.2$ is set experimentally considering the end result of whole preprocessing chain in light of text

enhancement and background variation suppression.

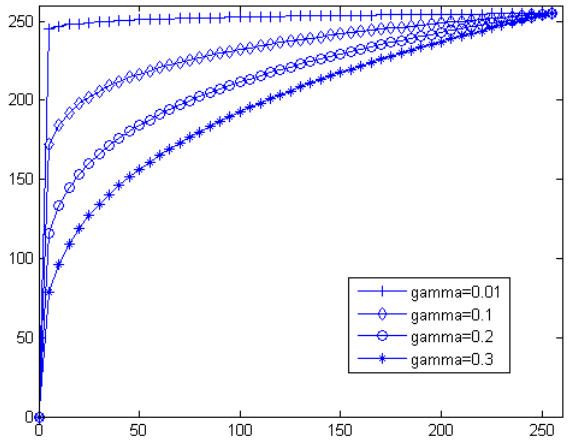


Figure 3.3: Gamma correction Graph

3.2.2 Difference of Gaussian (DoG)

Gamma correction does not remove the influence of overall intensity variations. Document images have slow/smooth background variations contributing to low-frequency component, text edges contributing middle range frequency components and noise contributing high-frequency component. DoG filtering [22] is a convenient way to achieve the bandpass behavior. As DoG name suggests, it is basically a difference of $2-D$ Gaussian filter having different variances (Outer mask is normally $2-3$ times broader than the inner mask). The inner Gaussian is typically quite narrow (usually variance $\sigma_1 \leq 1$ pixel essentially works as high pass filter), while the other σ_2 is $2-4$ pixel wide or more, depending on the spatial frequency at which low-frequency information becomes misleading rather than informative. Experimenting on different text images, values for σ_1 and σ_2 are set to 1 and 2 respectively. The resultant images for different value of σ_1 and σ_2 on images processed with different gamma values contain vary small variation in gray values, which are not perceivable with bare eyes (thus output images are not printed here).

3.2.3 Equalization

Processed image still typically contains extreme values produced by highlights, small dark regions, garbage at the image borders, etc. Following two-stage approximation [22] is used to rescale the gray values present in preprocessed image.

First stage,

$I(x,y) \leftarrow \frac{I(x,y)}{(mean(I(x,y))^{\alpha})^{\frac{1}{\alpha}}}$

The second stage,

$I(x,y) \leftarrow \frac{I(x,y)}{(mean(min(\tau, I(x,y))^{\alpha}))^{\frac{1}{\alpha}}}$

Here, $I(x,y)$ is image intensity at (x,y) location, α is a strongly compressive exponent that reduces the influence of large values, τ is a threshold used to truncate large values after the first phase of normalization and the mean is average intensity value of the image. By default, authors [22] set $\alpha = 0.1$ and $\tau = 10$ experimentally. The resulting image is well-scaled, but it may still contain some extreme values, to reduce its effect a nonlinear mapping is done using $I(x,y) \leftarrow \tau * tanh(I(x,y)/\tau)$, it scales whole image in the range of $[-\tau, \tau]$ The result of an end to end preprocessing are shown in figure 3.5.

3.3 Text Area Detection

Preprocessed image contains locally high contrast that visibly separates text region from nearer background with large intensity difference as shown in figure 3.6, parallel it suppresses local texture of background as well as remove several type of degradations like smear, Text intensity variations and background contrast variation, while suppressing other degradation like spots in image and bleed through as shown in figure 3.7. Preprocessed image is now more suitable for edge detection, proposed approach uses edge detection to identify possible text regions as explained below.

3.3.1 Edge Detection Using Rough-Set Theory

In most of the preprocessed documents, histogram shows one peak (due to mid-range gray values belonging to background), intensities lesser than the peak values generally belong to text, and intensities more than the peak value generally belong to background; but in practice the intensities lying nearer to peak intensity may contain pixels from text as well as background. Thus, applying edge detection method that works based on hard thresholding of image gradient magnitude like Sobel [29], canny [25] etc. will not work, to exploit local high contrast around text in preprocessed image rough set based edge detection [23, 30] is used [24].

Step 1: approximate thresholds are obtained from the histogram by sliding window based approach. It is adapted to find local minimum in the histogram and use multiple minimum as thresholds. The window size could be varied to get desired number of thresholds from an image (here, the window size is fixed to 5 empirically). The approximate thresholds are optimized separately using rough entropy measure. The optimization of rough entropy reveals minimizing roughness of the object at a threshold T. Rough entropy function used is defined in [31],

$RE_T = -\frac{1}{2}[Ro_T log_e(\frac{Ro_T}{e}) + Rb_T log_e(\frac{Rb_T}{e})]$

where, $Ro_T = 1 - \frac{|\bar{o_T}|}{|\underline{\mathbf{O}}_T|}$is roughness of object and $Rb_T = 1 - \frac{|\bar{b_T}|}{|\underline{\mathbf{b}}_T|}$ is roughness of background ($|\underline{X}|$ is lower approximation and $|\bar{X}|$ is upper approximation).

Step 2: The image under consideration is binarized with each threshold. For K number of thresholds, the image is binarized with K different thresholds, resulting in K binary images having different/same symbols 1 and 0 at each pixel location. Rough-set can be defined in terms of lower (certain member of a set) and upper (possible member of set) approximation. Object boundaries in the image are often uncertain due to the gradient of intensity present in the image around edge pixels. The binarized images will have constant 0 labeled region as well as constant 1 labeled regions in the area of almost constant intensity; along with region having label 0 as well as 1around the edges.

Step 3: To identify these possible edge regions the image is partitioned into different

non-overlapping blocks, if the block contains 90% or more than that pixels having same label consider the block as non-edge block (assign label 0), other wise consider it as edge block (assign label 1). Edges are considered one pixel thick thus 2×2 blocks are used while deciding edge block and non-edge blocks. This procedure is repeated for all K binary images.

Step 4: Now considering K binary images all at once, look out for a pixel location, if a pixel have same symbol 0 across all K images assign the 0 (certainly non-edge pixel) and if any of K symbol is 1, consider the pixel as possible edge. This procedure will give multiple edges for a single text, however, Rough-set edge detection have two-fold advantage; (i) it gives closed boundary (ii) it will detect text area without loosing any single pixel of text stroke. Results of Rough-set edge detection performed on preprocessed image are shown in figure 3.8

3.3.2 Region Filling

Region filling (Hole filling) [1] is process of filling hole in image. Here, hole is possible text area which is surrounded by connected text edges. Region filling is morphological algorithm based on set dilation and intersection. $X_k = (X_{k-1} dilation\, B)$ Here, A is our rough edge image as shown in figure 3.8. It starts with $X_0$of 0(same size as A), here given one point in each hole which is set to 1. B is structuring element $\begin{bmatrix} 0 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 0 \end{bmatrix}$. Algorithm terminates at iteration step k if $X_k = X_{k-1}$. Results of region filling on rough edge detected images are shown in figure 3.9. The text area of the document is clearly detected without loss of any text. Some extra pixels are also detected as foreground (contributing to false positives) around the text.

3.3.3 Small Object Removal

As shown in figure 3.10 some very small bunches of pixels are detected as foreground. In some of the images noise present in background (bleed through or very high textural background) results in wrongly detected text area (false positive) as shown in figure 3.10(b), where as in some images such bunch of pixels actually corresponds to text (true positive) for eg. semi colon symbol ($'$,$'$), period symbol ($'.'$), dot of letters like $'i'$ and $'j'$. In some of the images as shown in figure 3.10(a) where text stroke is very thin, identified text area have the trail of the very small disconnected component detected as foreground (true positive).

Some of these small connected components [1] needs to be removed, deciding the size of small connected components to be removed is a trade-off between loosing text (true positives) and keeping noise (false positives). Targeting to retain as much as possible text in the output image, it is observed that the components having the area less than 50 pixels are contributed by noise only. Thus, a size of a small component to be removed is set experimentally (The images from DIBCO data set are of very high resolution. The decision of the size of the connected component to be removed is highly dependent of image resolution. For images scanned at a different resolution the value will differ and need to be set experimentally). Results are shown in figure 3.11

3.4 Text Detection

Take all foreground pixels (pixels those are identified as text in previous stage), say, T from output image of stage 3.3.3 (small object removal). For a pixel under consideration, if intensity value in preprocessed image is higher than 173, label the pixel as background, resulting in image Ip. The threshold value is set to 173 experimentally, based on properties of preprocessed image as shown in figure 3.12. In most of the images, applying equalization after DoG filtering in preprocessing emphasizes text with high-intensity region enclosing it. It is observed that thresholding preprocessed image with some high value will separate background region surrounding the text. Ip contains excessive area segmented as fore-

ground in T, and is subtracted from T to leave us with only text (along with some noise namely, white blob, black blob and additional border like structures as explained in section 3.5) Results of text detection are shown in figure 3.13

input: preprocessed image, Text-area detected image (T)

output: Text detected image

algorithm:

step1: binarize preprocessed image with threshold 173, say, Ip

step2:

for a pixel p_i from T if pi(T)=1

if pi(Ip)=1

pi(T)=BG

else

do nothing in T

end

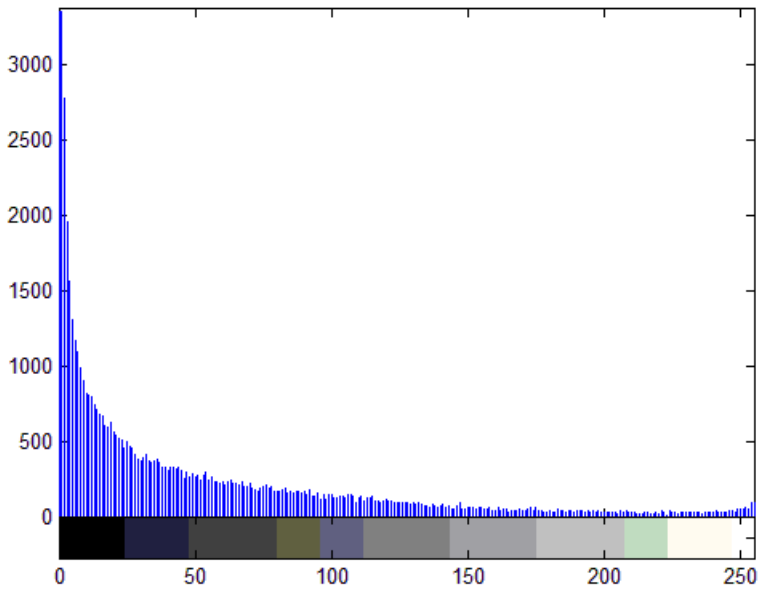


Figure 3.12: Histogram of Text intensity in preprocess image

3.5 Post Process

The output of the above stage gives text (foreground) as well as noise regions say Black blob (false foreground due to closed shape within alphabets) and white blob (false background), and non-text outer border around text which are mostly 1 pixel wide. As shown in fig 3.14

3.5.1 Border Removal

As can be seen from figure 3.14 the extra border like structure are one pixel wide and disconnected in nature, and have majority background pixels in its neighborhood, while every rightly classified text pixel has more than 50% text pixel (Foreground pixel) in its 8-neighborhood. Now consider each pixel that is classified as foreground, which has more than 60% background pixel in 8-neighborhood are considered background pixel. In this process, modification are carried out for next pixel decision. Results are shown in figure 3.16.

3.5.2 Black Blob Removal

Output of above step 3.5.1 is binarized text image which has most of the text region segmented as foreground, but some letters such as $'e',' o',' D',' R'$ have closed regions which get filled in region filling step and it appears as foreground in the above procedure which are actually not a foreground. To take care of such cases following procedure is performed.

- consider the pixels classified as Foreground in above stage, take the intensity values of this foreground region from preprocessed image. Now this values majority represent the text region in original input image, thus, mean (μ) and variance (σ) of these pixel values can be used to set an image dependent threshold intensityI_{th}.

- Considering μas threshold may result in some False Foreground pixel detection because foreground and background intensity largely overlaps. To reduce the amount of false foreground we can employ $\mu + \sigma$ as threshold I_{th}. This threshold covers almost

all the text pixel in foreground, simultaneously reduces false foreground occurrence in the above mentioned enclosed region of letters 'o', 'D' etc. The enclosed blobs are basically background (majority) having higher intensity value than $I_{th}$and thus will be classified as background.

The result of black blob removal is shown in fig3.17

3.5.3 White Blob Removal

There are some degraded images with some of the text area having very high intensities, such images performs poorly after enhancement, because such intensities get enhanced in preprocessing, and thus get classified as background as white blobs within foreground text areas.

To remove the white blobs following steps are performed

- Image contrast of gray scale image (output of stage 3.1) is enhanced such that 1% of total pixels having lowest and highest intensity values are saturated to 0 and 1 respectively.

- Binarizing the above image such that no text region is lost (i.e. False Foreground can be tolerated, but False Background is not allowed). Observing all the available images from database 128 intensity is set experimentally such that no text area is lost. This image is considered as reference image for the white blob removal decision making as shown in figure 3.19(b)

- Find all white connected component from the output of previous step, except the largest connected component which is clearly a background in the document image.

- If any white connected component has more then 30 % pixel in the reference image are classified as background then it remain background in output image else it is set to foreground in output.

Results of white blob removal is shown in fig3.19

After removing while blob, black blob and the unwanted border like structures image is almost correctly segmented into text and background with some small components present due to noise in the input image. to cleanup the output image small object removal is performed as explained in section 3.3.3. The size of the component is decided empirically such that it does not remove text pixels (true positive) like the period symbol, semi colon symbol and dots of letters like 'i' and 'j'.

3.6 Evaluation Measures

To evaluate the results quantitatively different measures are used like F-Measure, Peak Signal-to-Noise Ratio (PSNR), Distance Reciprocal Distortion (DRD), pseudo F-Measure, as mention in [32, 5]. The online tool is available on DIBCO website to the evaluate performance of algorithms [4].

F-Measure :

A Higher value of F-Measure represents a good result.

$Recall = \frac{TP}{TP+FN}$,

$Precision = \frac{TP}{TP+FP}$

where TP=True Positive, FN =False Negative, FP =False Positive

$$FM = \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}}$$

.

PSNR:

PSNR represents the closeness of one image to another, in this case, binarized image to its ground truth. Higher the PSNR better the performance of the algorithm. Following equation shows the formula to calculate PSNR.

$$\text{PSNR(dB)} = 10 \log_{10} \frac{P^2 MN}{\sum_{x=0}^{M-1} \sum_{y=0}^{N-1} (g(x,y) - f(x,y))^2}$$

where, P is maximum value possible value a signal can take (here, highest possible intensity i.e. 1), g is binarized image, f is ground truth image, $M \times N$ is the size of image g and f.

DRD [33]:

It is a measure of distortion in binary images as perceived by human vision. It uses weight matrix with each of its weights determined by the reciprocal of a distance measured from the center pixel. Here weight matrix is W_m of size $m \times m$, $m = 2n$+1 $n = 1, 2, 3, 4, 5...$ center element is (i_c, j_c) where $i_c = j_c = (m + 1)/2$.
$\mathbf{W_m}(i, j) = \frac{1}{\sqrt{(i-i_c)^2+(j-j_c)^2}}$,if $i = i_c$and $j = j_c\mathbf{W_m}(i, j) = 0$ matrix normalization can be done using,

$$\mathbf{W}_{\mathrm{Nm}}(i, j) = \frac{\mathbf{W}_m(i, j)}{\sum_{i=1}^{m}\sum_{j=1}^{m}\mathbf{W}_m(i, j)}$$

If R, number of pixels flipped in output image $g(x, y)$, Each pixel will have a distance reciprocal distortion DRD_k $k = 1, 2, 3, ...R$. kth flipped pixel at $(x, y)_k$ in the output image $g(x, y)$, the resulted distortion is calculated from an $m \times m$ block B_k in $f(x, y)$ (Ground truth image) that is centered at $(x, y)_k$. The distortion DRD_k measured for this flipped pixel $g(x, y)_k$ is given by

$$\mathrm{DRD}_k = \sum_{i,j}[\mathbf{D}_k(i, j) \times \mathbf{W}_{\mathrm{Nm}}(i, j)]$$

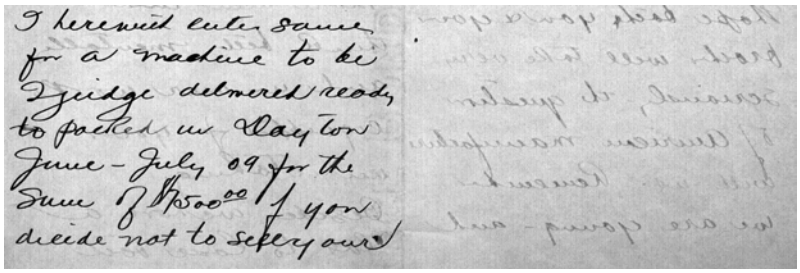
$$\mathbf{D}_k(i, j) = |\mathbf{B}_k(i, j) - g[(x, y)_k]|$$

$$\mathrm{DRD} = \frac{\sum_{k=1}^{S}\mathrm{DRD}_k}{\mathrm{NUBN}}$$

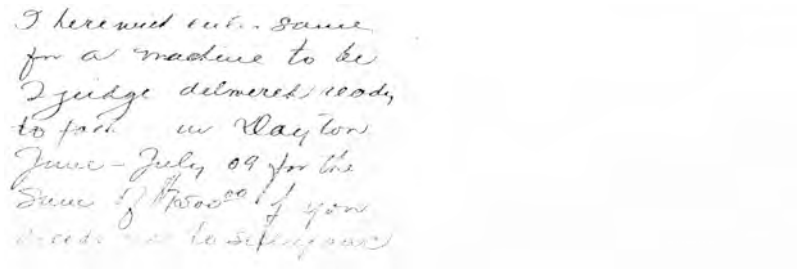
NUBN = Number of nonuniform blocks (8×8) in ground truth image.

pseudo F-Measure:

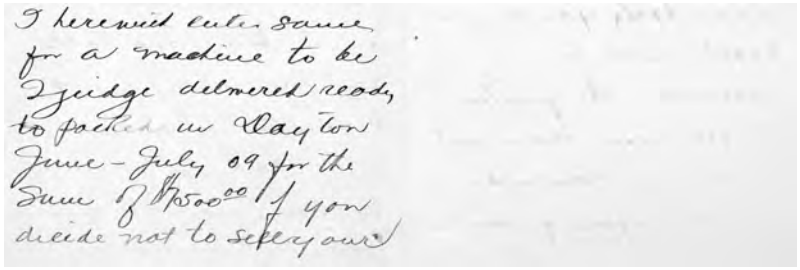
This measure is proposed in [32]. It uses pseudo recall and pseudo-precision instead of recall and precision in F-measure. pseudo recall and pseudo-precision use weight matrix according to the contour of ground truth. In pseudo-Recall the weights of the ground truth foreground are normalized according to the local stroke width and In pseudo-Precision the weights are constrained within an area that expands to the ground truth background, taking into account the stroke width of the nearest ground truth component. Inside this area, the weights are greater than one (generally delimited between (1,2) while outside this area they are equal to one [32].



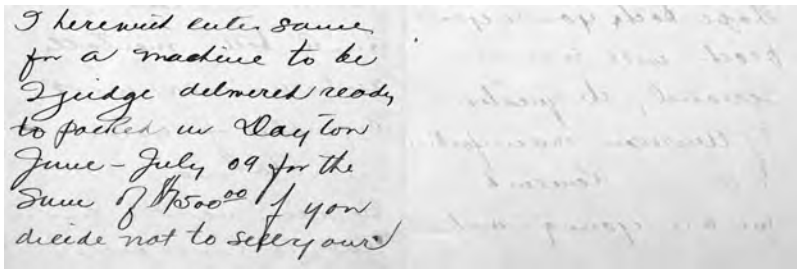
(a)



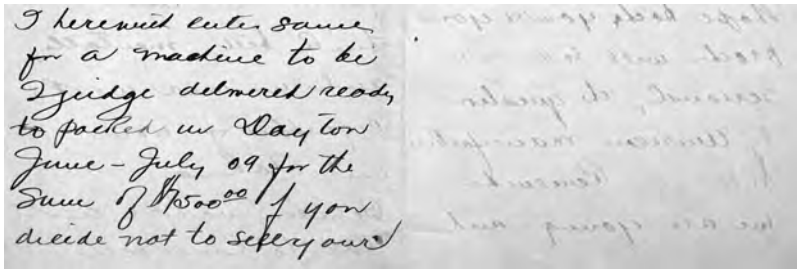
(b)



(c)

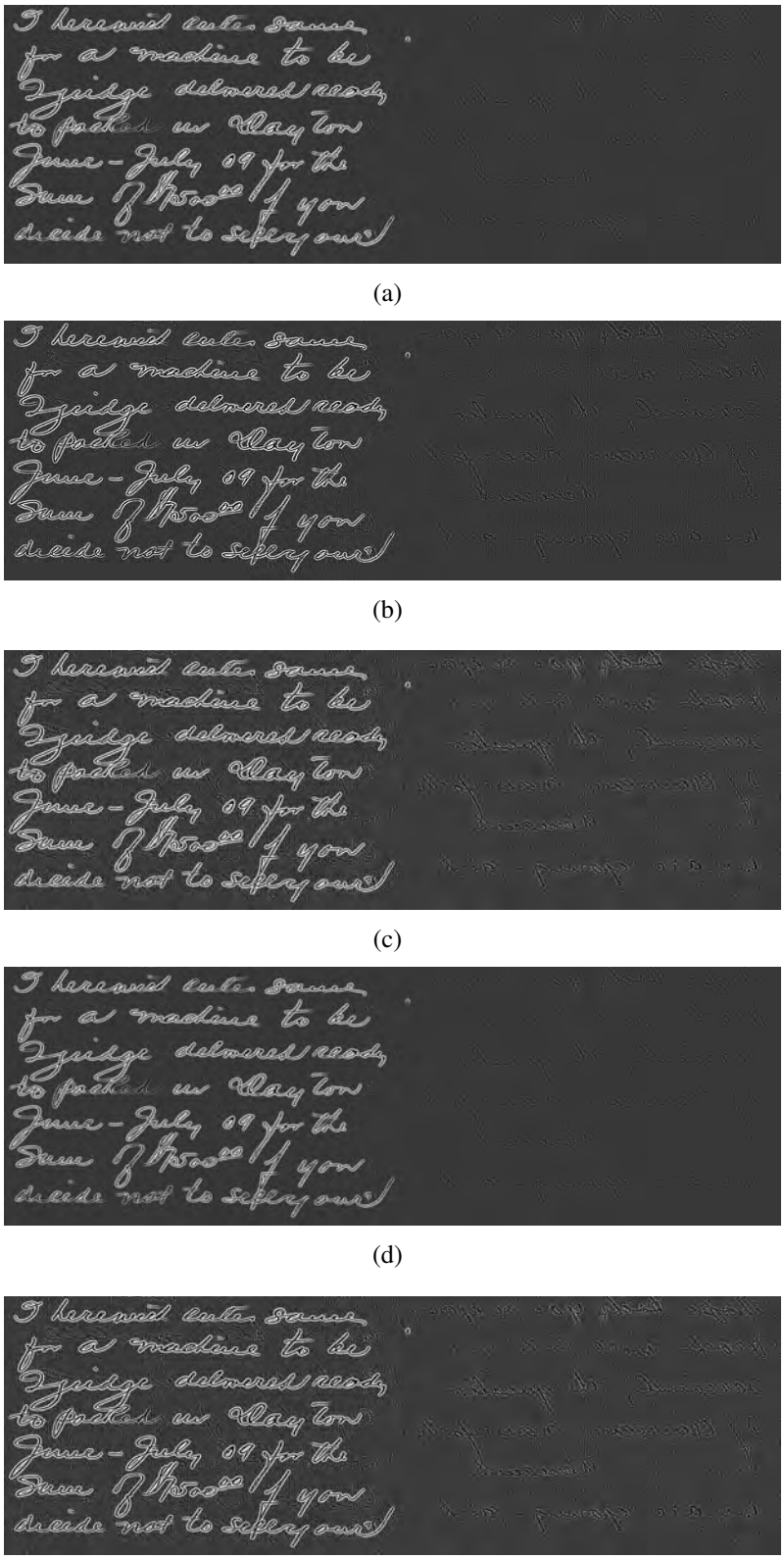


(d)



(e)

Figure 3.4: Results of Gamma correction with different γ values, (a) Gray image using PCA, (b) $\gamma = 0.01$, (c) $\gamma = 0.1$, (d) $\gamma = 0.2$, (e) $\gamma = 0.3$



(a)

(b)

(c)

(d)

(e)

Figure 3.5: Results of Equalization process with different τ value on different gamma value and Dog parameters output images, (a) $\gamma = 0.1, \sigma_1 = 1, \sigma_{,2} = 2, \tau = 10$, (b) $\gamma = 0.2, \sigma_1 = 1, \sigma_{,2} = 2, \tau = 10$, (c) $\gamma = 0.2, \sigma_1 = 1, \sigma_{,2} = 2, \tau = 5$, (d) $\gamma = 0.2, \sigma_1 = 0.5, \sigma_{,2} = 2, \tau = 10$, (e) $\gamma = 0.3, \sigma_1 = 1, \sigma_{,2} = 2, \tau = 5$

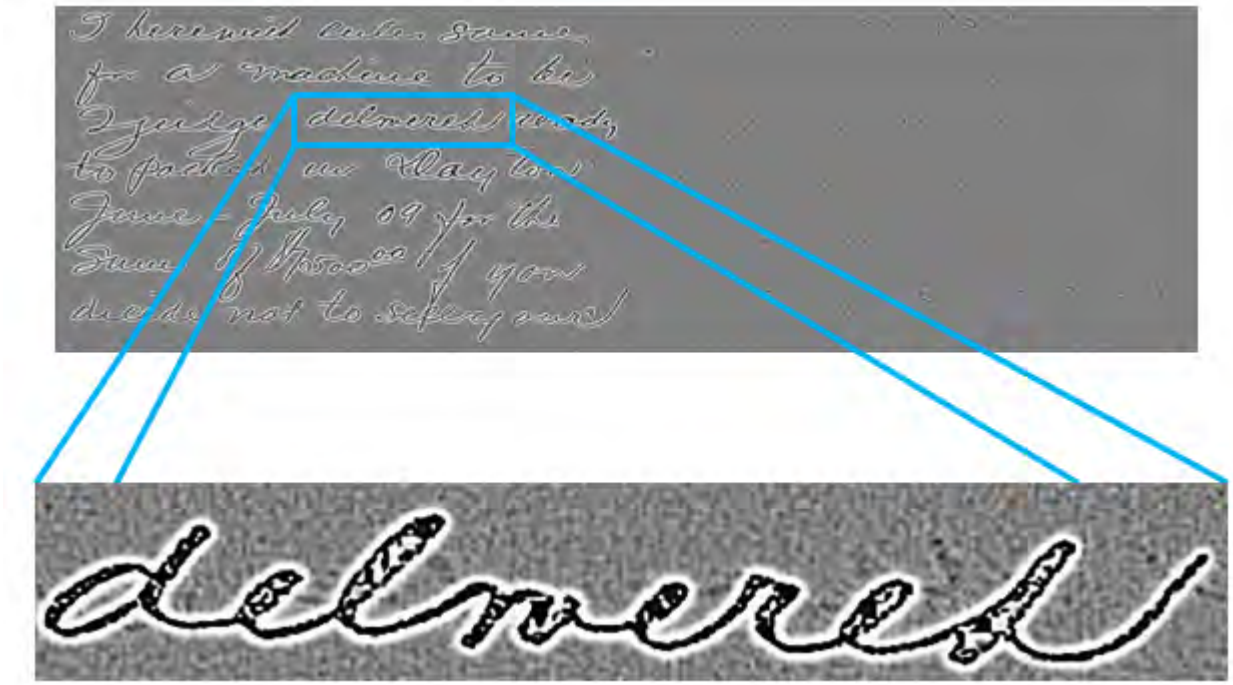
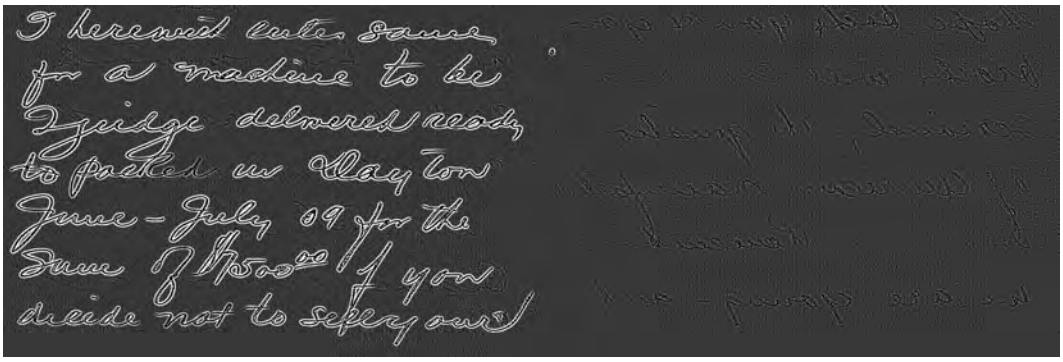
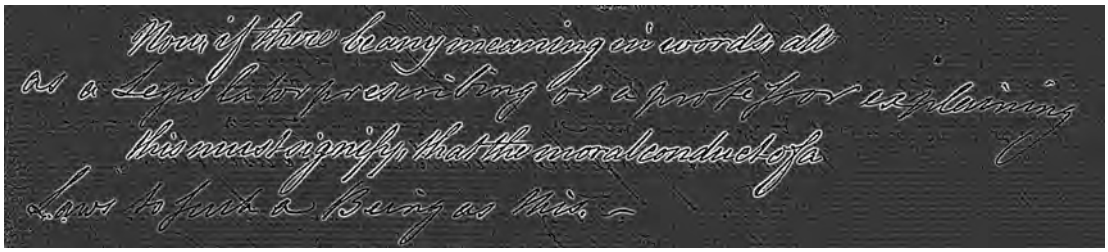


Figure 3.6: Result of preprocess

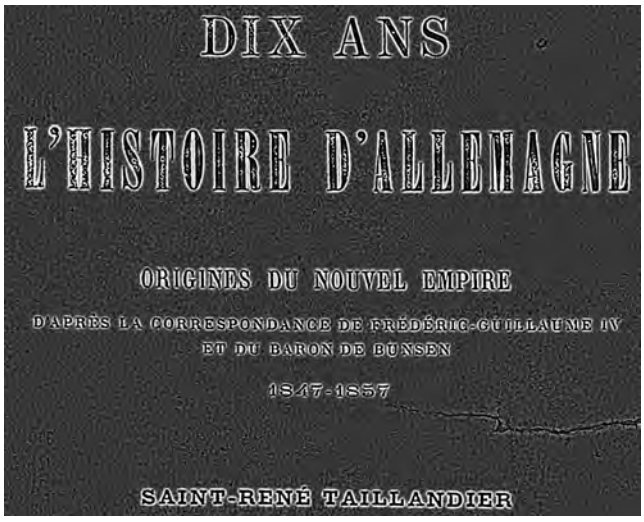
(a)



(b)



(c)



(d)

Figure 3.7: Results of preprocess on different degraded images, (a) Suppress spot degradation in image, (b) Suppress bleed through degradation, (c) Remove text intensity variation (d) Remove background contrast variation

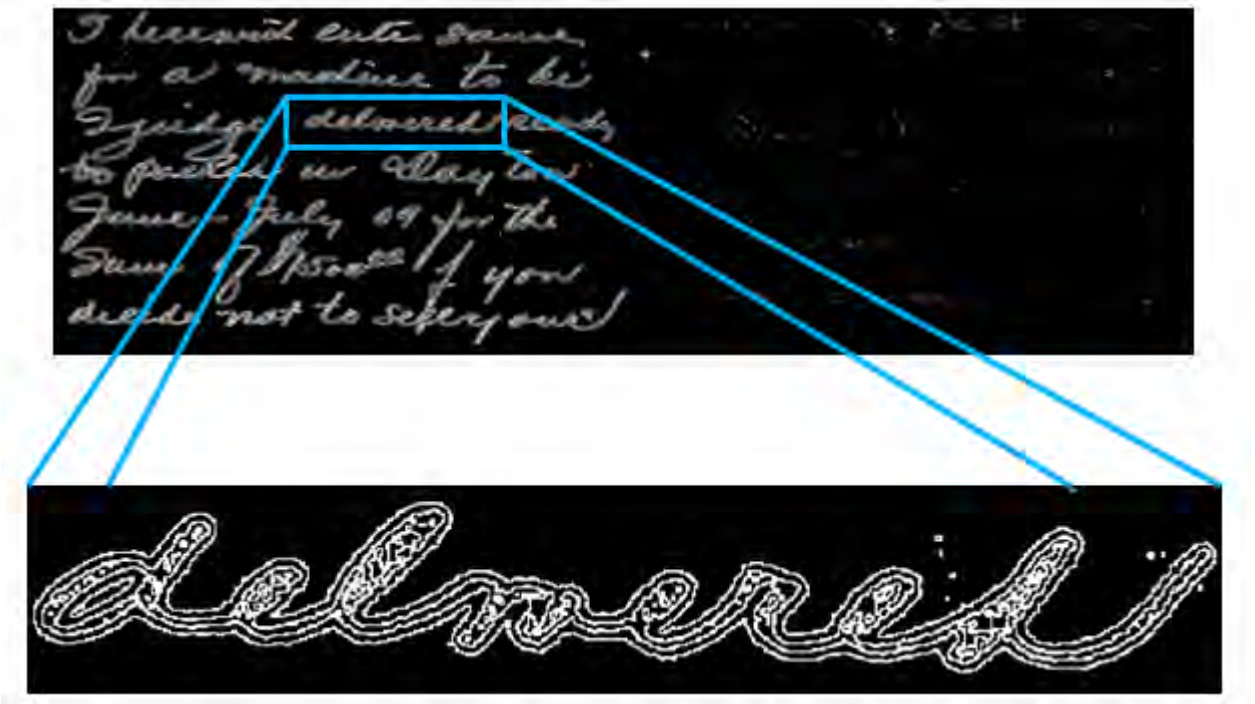


Figure 3.8: Result of Rough_edge detection

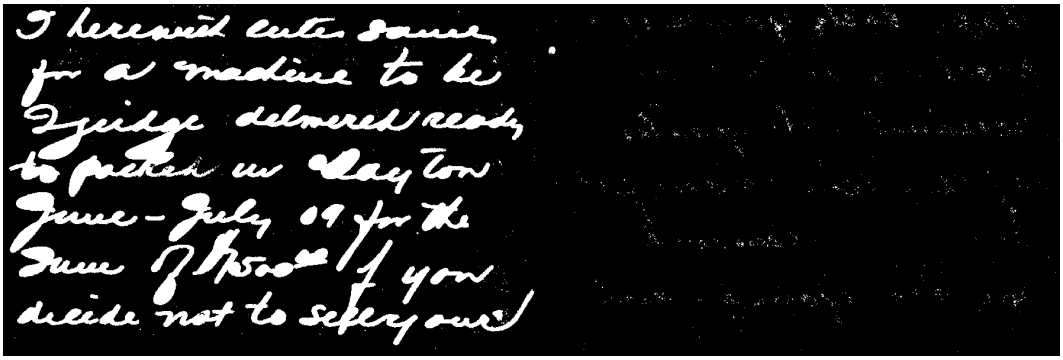
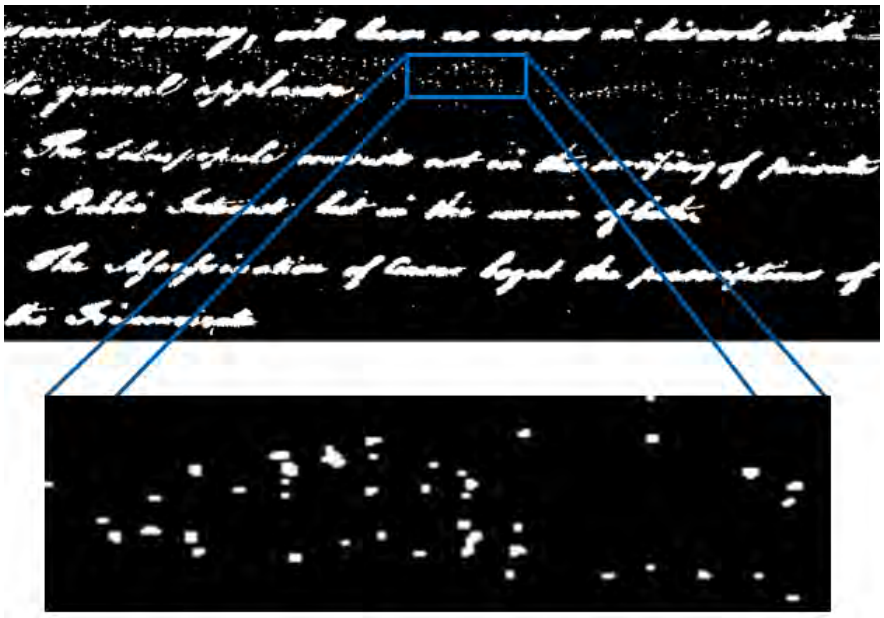


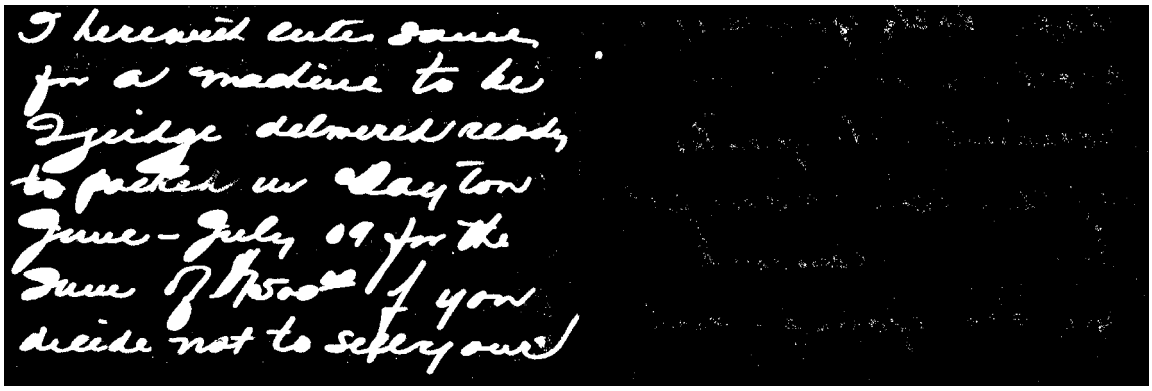
Figure 3.9: Result of Region Filling process

(a)

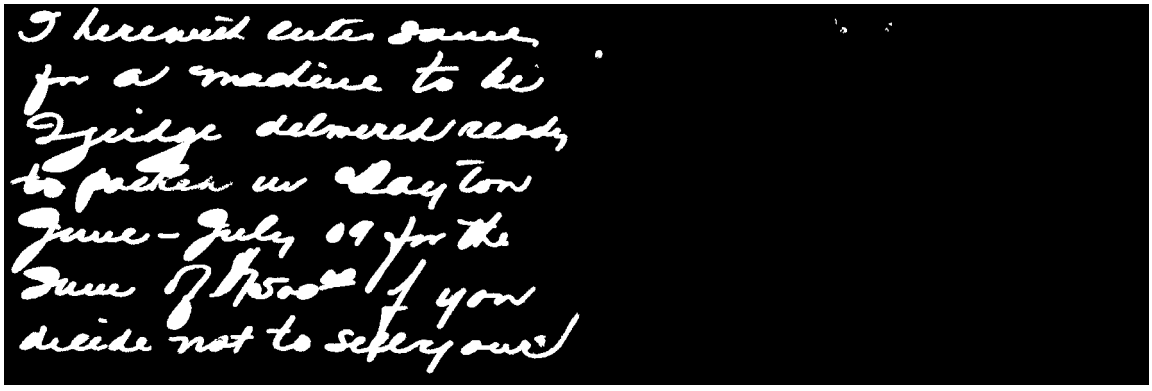


(b)

Figure 3.10: Demonstration of small object as text and noise in image, (a) Result of Region fill image with true positive, (b) Result of Region fill image with false positive

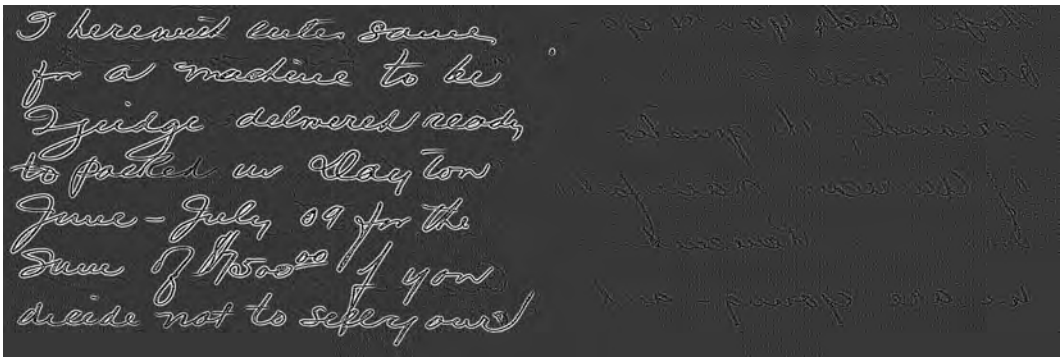


(a)

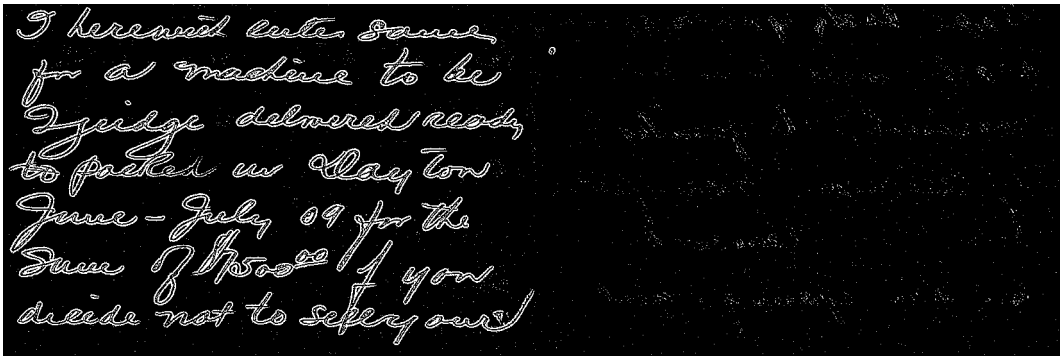


(b)

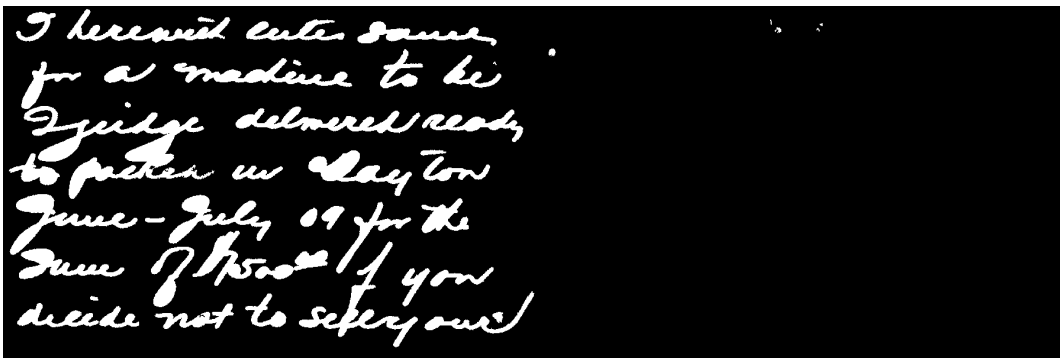
Figure 3.11: Result of Small object removal process, (a) Result of Region fill image, (b) Result of small object removal on region fill image



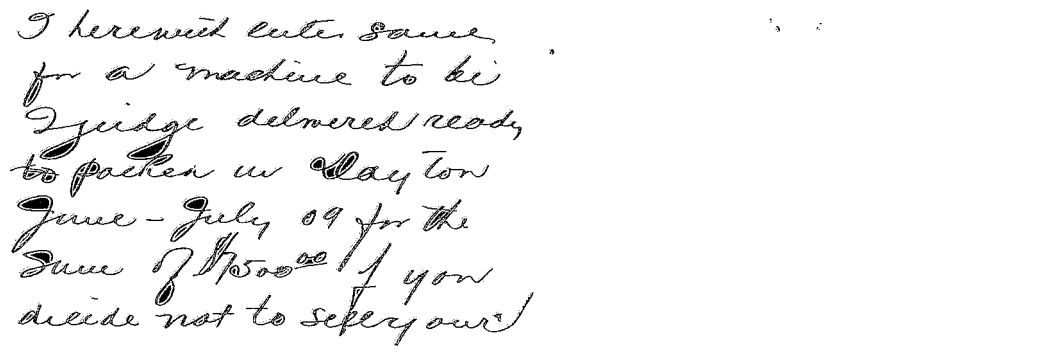
(a)



(b)



(c)



(d)

Figure 3.13: Result of Text detection process, (a) Preprocess image, (b) Result of thresholding with 173 threshold value, (c) Result of small part removal, (d) Result of Text detection

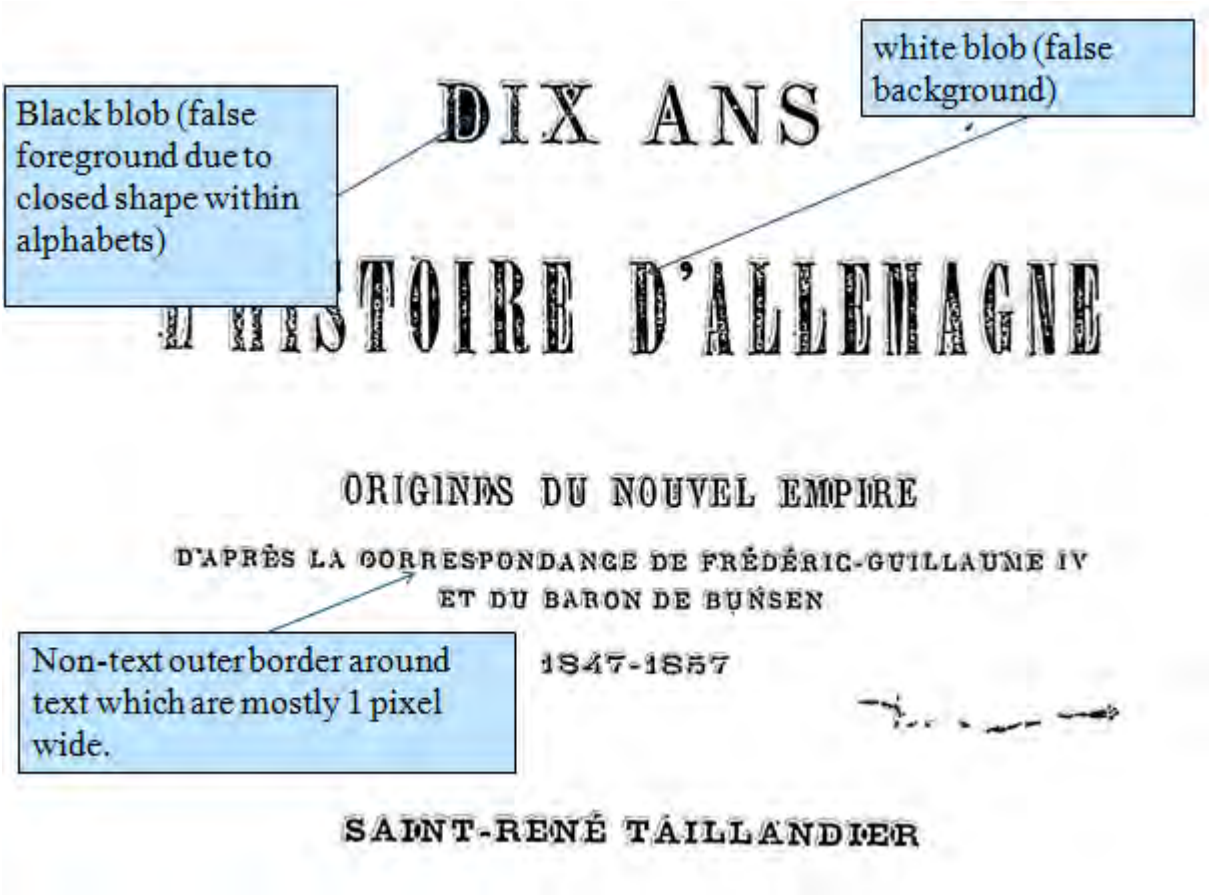
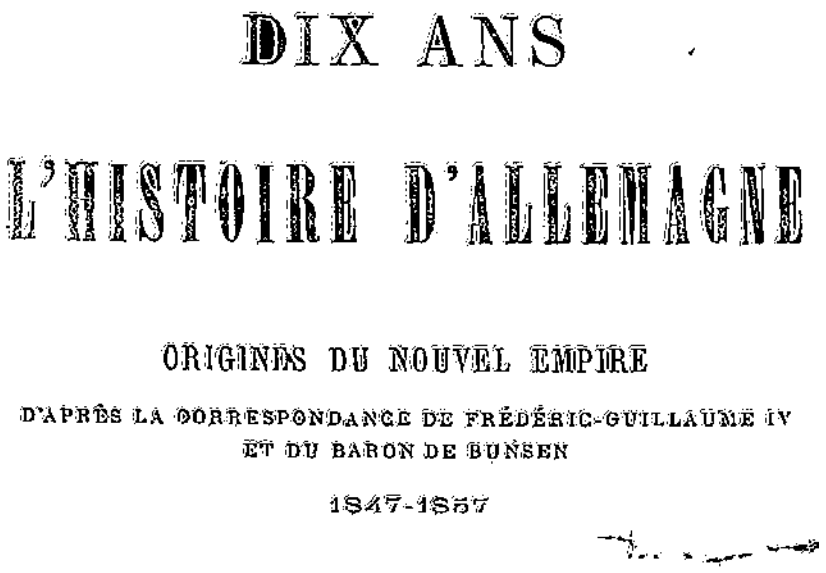
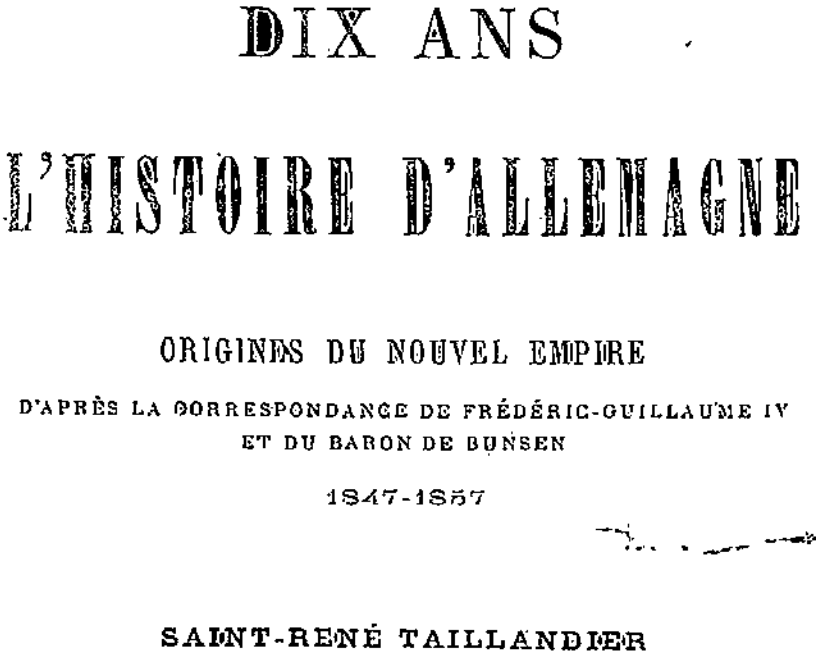


Figure 3.14: Result of Text detection process with problems

Figure 3.15: Border removal Input and Output image

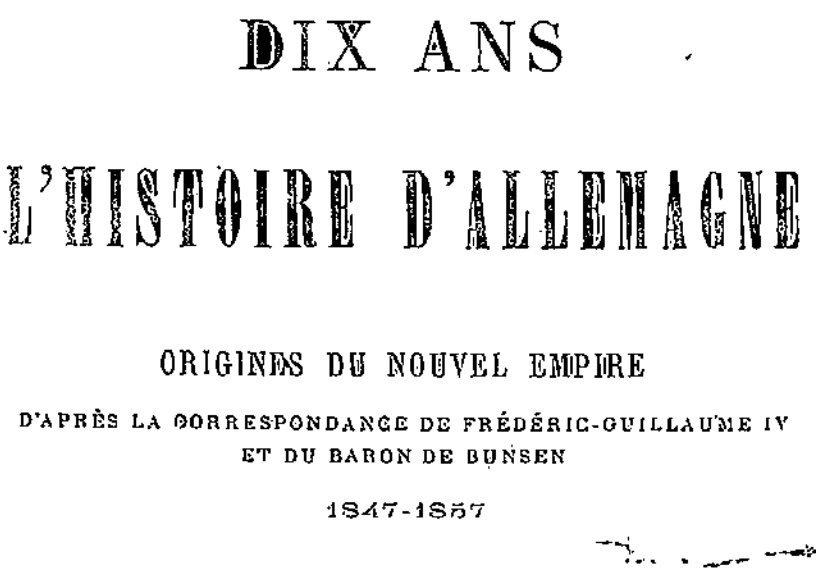


(a)

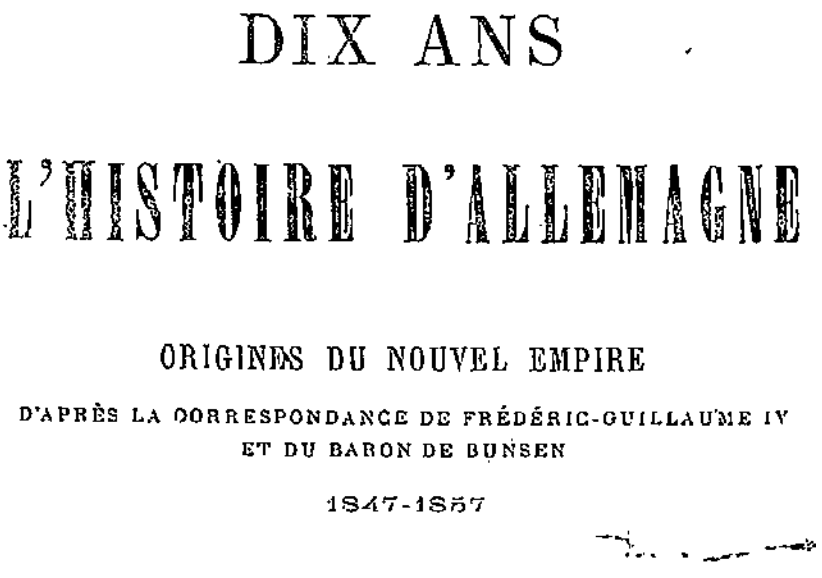


(b)

Figure 3.16: Result of Border removal process, (a) Result of text detection, (b) Result of Border removal

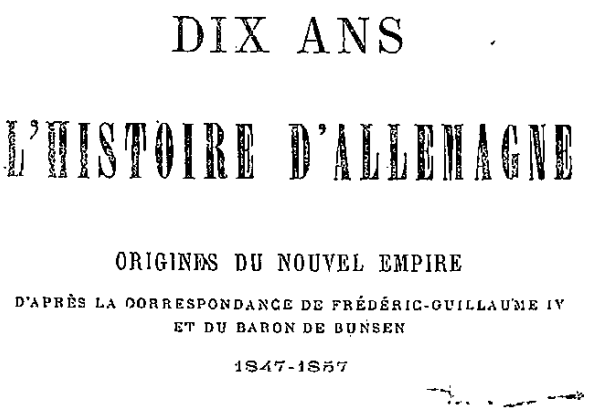


(a)



(b)

Figure 3.17: Result of Black blob removal process, (a) Result of Border removal, (b) Result of black blob removal


DIX ANS

L'HISTOIRE D'ALLEMAGNE

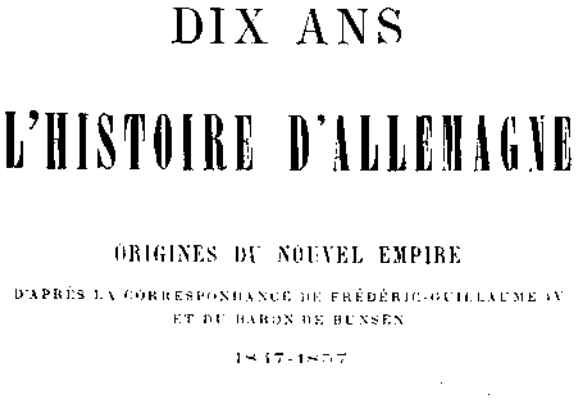
ORIGINES DU NOUVEL EMPIRE

D'APRÈS LA CORRESPONDANCE DE FRÉDÉRIC-GUILLAUME IV
ET DU BARON DE BUNSEN

1847-1857

SAINT-RENÉ TAILLANDIER

(a)


DIX ANS

L'HISTOIRE D'ALLEMAGNE

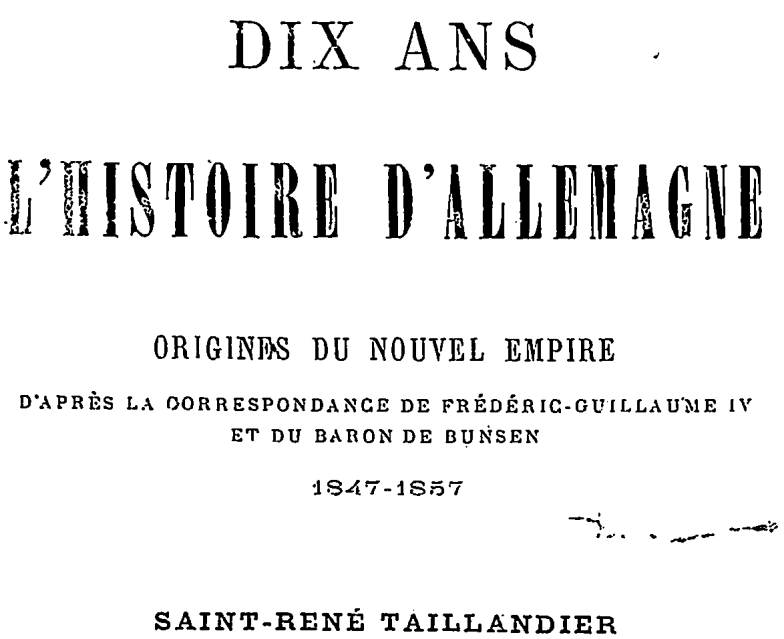
ORIGINES DU NOUVEL EMPIRE

D'APRÈS LA CORRESPONDANCE DE FRÉDÉRIC-GUILLAUME IV
ET DU BARON DE BUNSEN

1847-1857

SAINT-RENÉ TAILLANDIER

(b)


DIX ANS

L'HISTOIRE D'ALLEMAGNE

ORIGINES DU NOUVEL EMPIRE

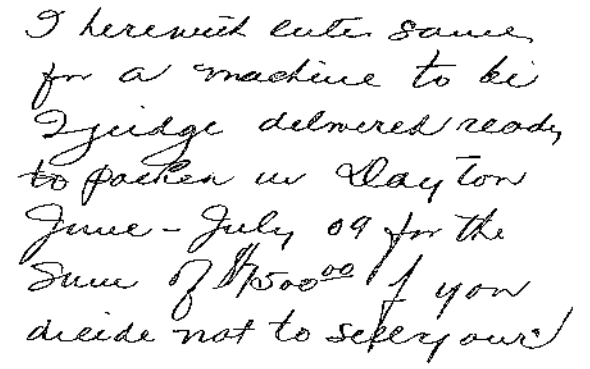
D'APRÈS LA CORRESPONDANCE DE FRÉDÉRIC-GUILLAUME IV
ET DU BARON DE BUNSEN

1847-1857

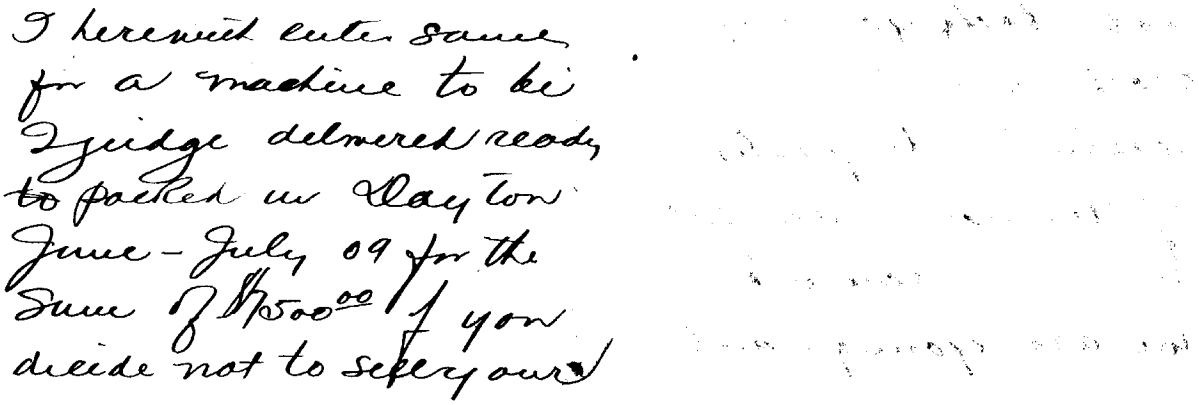
SAINT-RENÉ TAILLANDIER

(c)

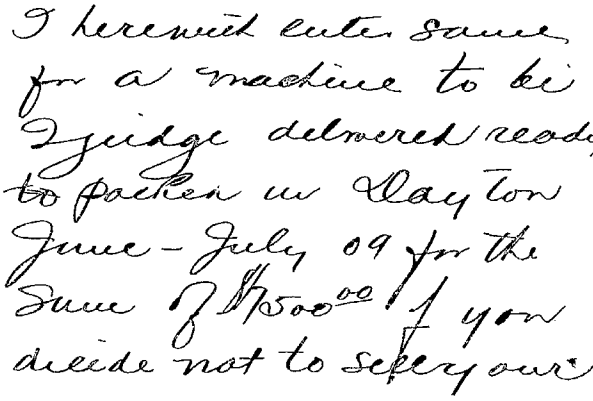
Figure 3.18: (a) Result of black blob removal, (b) Reference Binarized image, (c) Result of white blob removal



(a) Result of black blob removal



(b) Reference Binarized image



(c) Result of white blob removal

Figure 3.19: Result of White blob removal process, (a) Result of black blob removal, (b) Reference Binarized image, (c) Result of white blob removal

Chapter 4

Results

Proposed approach is tested on all the DIBCO data set [6]. Results are shown in figure 4.2. Performance of proposed approach on DIBCO data set is measured using DIBCO evaluation tool which is available at [4], which required ground truth image and resultant image. Quantitative evaluation measurements of proposed approach are reported in 4.1. The best result of the proposed approach is shown in figure 4.2 (h).

Moreover, proposed algorithm does not depend on text size and stroke width so it can be used on image containing text of almost all script, To test the suitability of proposed approach, Gujarati degraded document image data set is created. The algorithm performs equally well on Gujarati document images, that justifies the claim. Results on Gujarati data set are shown in figure 4.3. Ground truth of data set is not generated so performance evaluation is measured qualitatively by visual appearance. Result in figure 4.3(f) shows that the proposed approach can efficiently binarize English numerical and Gujarati numerical also.

4.1 Results of Proposed Approach

This section includes quantitative performance measures of proposed approach using evaluation measure explain in section 4.1. Following tables include results on DIBCO data set of year 2009 in table 4.1, 2010 in table 4.2, 2011 in table 4.3, 2012 in table 4.4, 2013 in

table 4.6 and 2014 in table 4.5.

Table 4.1: Result measurement of proposed method on data set DIBCO2009

Image	FM	FPS	PSNR	DRD	RECALL	PRECISION	RPS	PPS
H01	88.91	87.92	17.96	3.89	95.86	82.89	99.35	78.86
H02	73.80	81.62	19.39	11.45	74.99	72.65	93.97	72.14
H03	88.85	89.65	16.42	4.10	93.65	84.52	99.13	81.83
H04	86.84	90.95	17.18	5.07	86.05	87.65	97.15	85.49
H05	82.78	83.76	18.43	7.43	90.56	76.23	97.29	73.54
P01	90.30	90.54	16.14	2.90	93.85	87.01	98.46	83.80
P02	83.11	79.08	10.99	12.40	94.47	74.19	99.80	65.48
P03	96.54	98.17	19.32	2.18	95.62	97.48	99.53	96.86
P04	85.85	87.72	15.07	6.40	90.33	81.80	99.56	78.40
P05	85.79	89.72	13.97	4.24	82.73	89.10	94.10	85.74
AVG	**86.28**	**87.91**	**16.49**	**6.01**	**89.81**	**83.35**	**97.83**	**80.21**

4.2 Comparison With State Of The Art Methods

In some images proposed algorithm works better than the state of the art method (Winner of DIBCO 2013) [34] and in some images the state of the art method works better. The advantage of proposed algorithm is that it gives maximal text detection. Statistically overall results is close to state of the art method results as shown in figure 4.6, but text detection in all type of degraded images using proposed algorithm is good in the sense that almost no text region is classified as background (false background), results are shown in figure 4.4. As shown in figure 4.4(a) and (g), state of the art method loss many texts (Foreground classified as background resulting into false background) compare to the result of proposed algorithm in figure 4.4(b) and (h). In image4.4 (c) state of the art method works better in terms of measures discussed compare to the proposed method in fig 4.4(c) but proposed method does not loss as many text pixels as state of the art method as shown in image figure 4.4(a) and (g). The quantitative comparison on DIBCO 2013 data set is shown in table 4.6 and table 4.7.

Table 4.2: Result measurement of propossed method on data set DIBCO2010

Image	**FM**	**FPS**	**PSNR**	**DRD**	**RECALL**	**PRECISION**	**RPS**	**PPS**
H01	84.03	86.88	14.97	6.52	78.35	90.60	85.81	87.97
H02	87.10	86.98	19.04	6.21	93.44	81.56	97.59	78.45
H03	90.50	91.47	18.67	2.51	91.26	89.76	95.95	87.40
H04	89.13	91.07	17.38	3.15	90.10	88.18	97.31	85.57
H05	86.51	95.27	18.50	3.82	78.39	96.51	94.89	95.64
H06	88.13	85.99	17.89	3.12	95.20	82.04	97.83	76.71
H07	87.22	92.07	17.77	3.73	81.13	94.30	91.20	92.96
H08	86.85	85.70	16.44	3.70	94.81	80.12	98.00	76.14
H09	85.74	84.64	18.29	4.58	96.10	77.40	99.02	73.90
H10	88.67	89.57	18.79	3.46	85.44	92.16	88.85	90.31
AVG	**87.39**	**88.96**	**17.77**	**4.08**	**88.42**	**87.26**	**94.65**	**84.50**

Table 4.3: Result measurement of propossed method on data set DIBCO2011

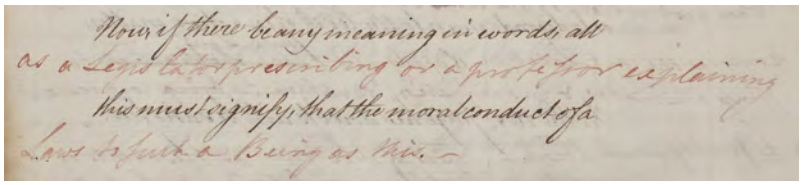
Image	**FM**	**FPS**	**PSNR**	**DRD**	**RECALL**	**PRECISION**	**RPS**	**PPS**
HW1	84.73	85.56	13.73	9.03	92.64	78.06	99.24	75.19
HW2	88.94	90.82	19.88	3.31	89.94	87.95	97.77	84.79
HW3	83.35	86.84	16.07	4.91	82.05	84.68	93.31	81.22
HW4	86.37	87.96	16.08	3.49	83.75	89.16	89.24	86.71
HW5	90.03	91.12	16.29	4.26	94.36	86.08	99.34	84.16
HW6	68.82	73.36	14.35	7.78	55.35	90.96	62.28	89.22
HW7	88.75	92.23	20.59	2.60	83.82	94.31	91.67	92.79
HW8	90.22	90.71	20.22	2.62	92.20	88.32	96.87	85.29
PR1	77.91	85.10	12.12	11.37	64.30	98.81	74.93	98.46
PR2	78.95	89.42	13.78	6.16	67.06	95.96	84.15	95.39
PR3	88.64	89.76	13.72	5.02	89.92	87.38	98.80	82.23
PR4	87.84	92.24	15.71	6.29	85.67	90.13	98.30	86.89
PR5	89.04	91.58	15.27	3.17	87.54	90.61	95.54	87.94
PR6	77.17	86.11	17.32	8.62	63.17	99.14	76.21	98.96
PR7	91.41	92.71	23.67	2.92	92.53	90.32	98.43	87.62
PR8	82.11	83.76	13.35	5.04	77.00	87.94	82.76	84.79
AVG	**84.64**	**88.08**	**16.39**	**5.41**	**81.33**	**89.99**	**89.93**	**87.60**

Table 4.4: Result measurement of proposed method on data set DIBCO2012

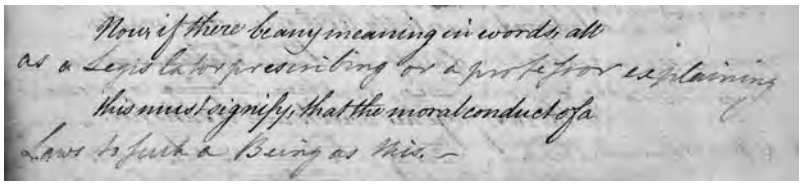
Image	**FM**	**FPS**	**PSNR**	**DRD**	**RECALL**	**PRECISION**	**RPS**
H02	78.48	84.03	15.08	7.49	66.89	94.93	75.37
H03	60.61	65.85	14.40	16.11	43.91	97.81	49.63
H04	81.93	91.49	17.94	5.91	75.07	90.16	92.86
H05	89.93	93.23	19.13	3.09	92.00	87.95	99.18
H06	88.10	93.33	18.33	4.11	82.55	94.45	92.23
H07	87.55	89.40	17.50	3.62	93.76	82.11	98.12
H08	92.20	97.19	19.74	2.35	88.59	96.11	98.30
H09	92.44	95.24	17.64	2.84	91.68	93.21	97.36
H10	81.96	84.46	15.03	8.12	71.62	95.80	75.52
H11	80.22	80.49	15.22	8.04	71.18	91.89	71.60
H12	90.91	92.61	19.59	2.68	95.34	86.87	99.16
H13	83.96	86.27	17.92	4.84	87.25	80.90	92.41
H14	92.80	96.67	21.88	2.10	91.19	94.48	98.96
AVG	**84.70**	**88.48**	**17.65**	**5.49**	**80.85**	**91.28**	**87.75**

Table 4.5: Result measurement of propossed method on data set DIBCO2014

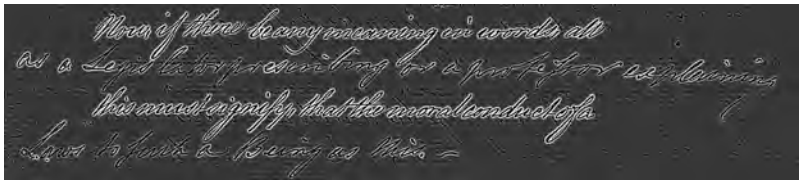
Image	**FM**	**FPS**	**PSNR**	**DRD**	**RECALL**	**PRECISION**	**RPS**	**PPS**
H01	92.77	95.99	20.95	2.03	90.78	94.85	98.31	93.78
H02	91.24	93.93	18.51	2.47	89.29	93.28	96.25	91.72
H03	95.85	99.41	20.80	2.13	92.54	99.41	99.54	99.29
H04	92.59	92.36	16.33	2.97	96.69	88.82	99.75	85.99
H05	92.04	91.57	15.67	3.39	96.24	88.19	99.16	85.06
H06	92.76	93.45	16.59	3.40	91.93	93.61	94.64	92.30
H07	89.94	90.93	17.38	3.01	92.55	87.48	98.59	84.38
H08	92.58	96.54	23.58	2.04	90.14	95.16	98.93	94.27
H09	91.74	93.30	17.89	2.52	91.97	91.51	97.49	89.45
H10	91.40	92.70	17.68	2.66	92.41	90.42	98.25	87.74
AVG	**92.29**	**94.02**	**18.54**	**2.66**	**92.45**	**92.27**	**98.09**	**90.40**



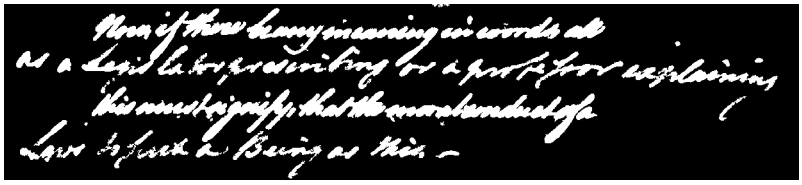
(a)



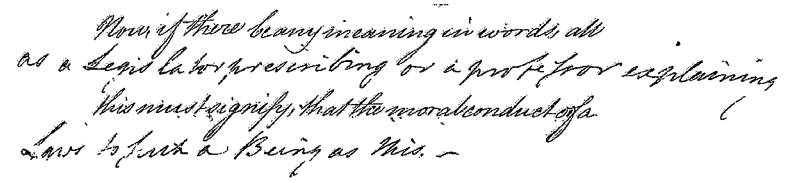
(b)



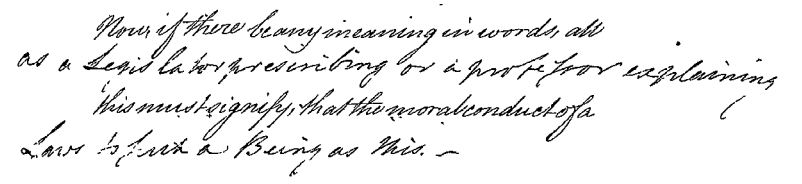
(c)



(d)



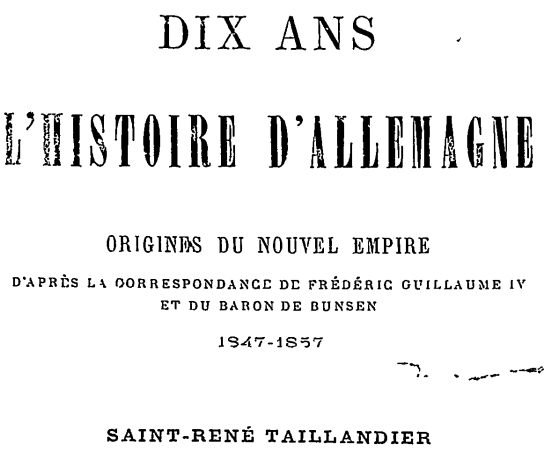
(e)



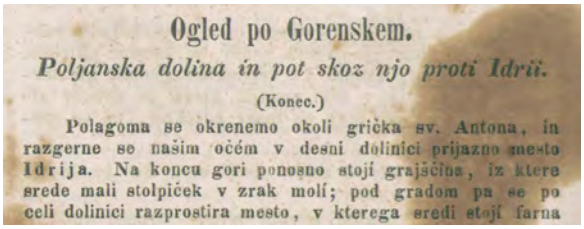
(f)

Figure 4.1: Result of proposed approach (a) Input RGB image, (b) Result of RGB to Gray using PCA on RGB image, (c) Result of preprocess on gray image, (d) Result of text area detection on preprocessed image, (e) Result of text detection from text area detected image, (f) Result of post process on text detected image.

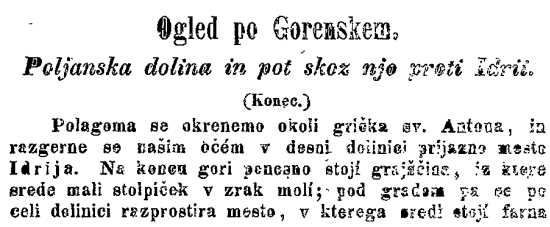
(a)



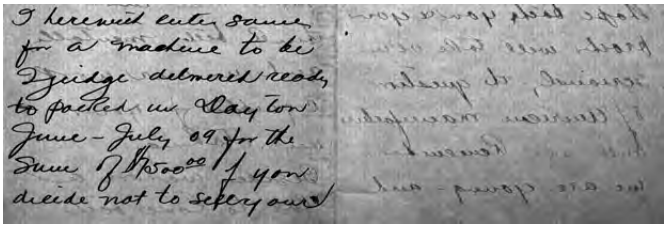
(b)



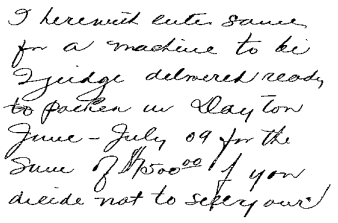
(c)



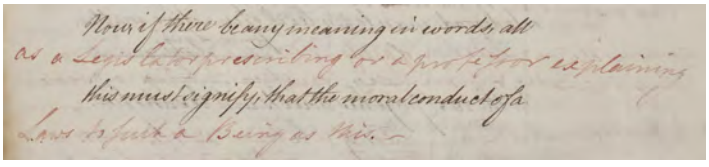
(d)



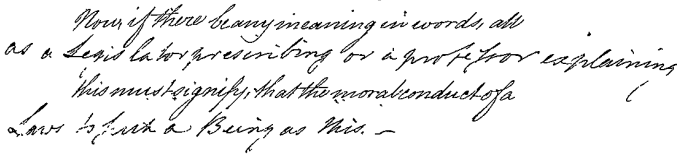
(e)



(f)

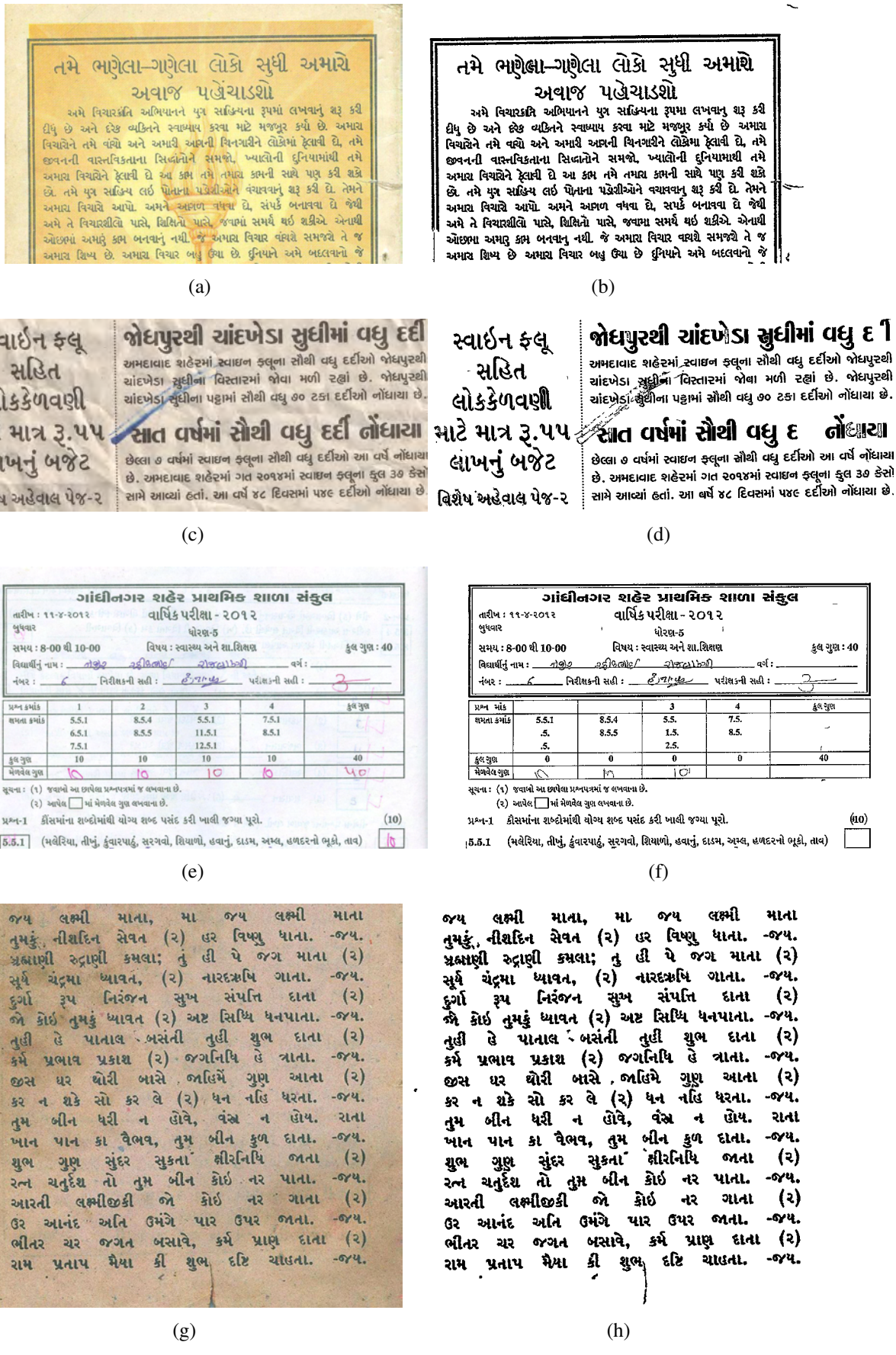


(g)



(h)

Figure 4.2: Original images and respective output images using Proposed approach, (a), (c), (e), (g) DIBCO dataset images in RGB and (b), (d), (f), (h) Respective binarized output images using proposed approach



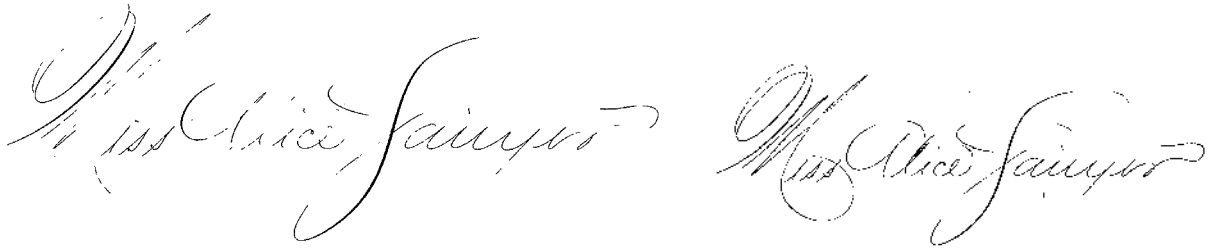
(a) (b)

(c) (d)

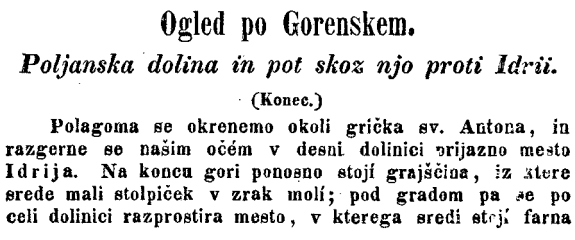
(e) (f)

(g) (h)

Figure 4.3: Results of Gujarati data set using Proposed approach (a), (c), (e), (g) Gujarati dataset images in RGB and (b), (d), (f), (h) Respective binarized output images using proposed approach



(a) (b)



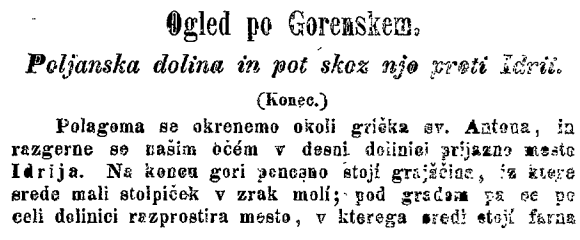
Ogled po Gorenskem.

Poljanska dolina in pot skoz njo proti Idrii.

(Konec.)

Polagoma se okrenemo okoli grička sv. Antona, in razgerne se našim očém v desni dolinici prijazno mesto **Idrija**. Na koncu gori ponosno stojí grajščina, iz ktere srede mali stolpiček v zrak molí; pod gradom pa se po celi dolinici razprostira mesto, v kterega sredi stojí farna

(c)



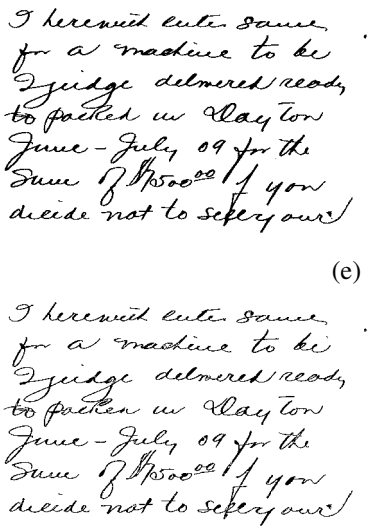
Ogled po Gorenskem.

Poljanska dolina in pot skoz njo proti Idrii.

(Konec.)

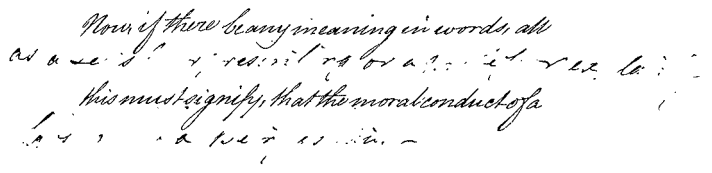
Polagoma se okrenemo okoli grička sv. Antona, in razgerne se našim očém v desni dolinici prijazno mesto **Idrija**. Na koncu gori ponosno stojí grajščina, iz ktere srede mali stolpiček v zrak molí; pod gradom pa se po celi dolinici razprostira mesto, v kterega sredi stojí farna

(d)

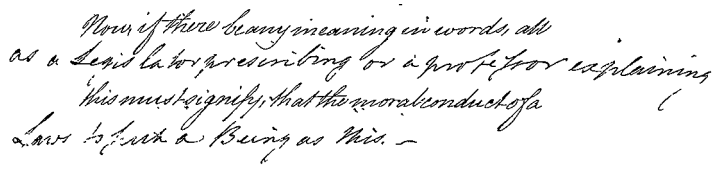


(e)

(f)



(g)



(h)

Figure 4.4: State of art method results ((a), (c), (e), and (g)) and respective images result based on proposed approach ((b), (d), (f), and (h))

Table 4.6: Result measurement of Proposed method on data set DIBCO2013

Image	**F-Measure**	**FPS**	**PSNR**	**DRD**	**RECALL**	**PRECISION**	**RPS**	**PPS**
HW01	91.16	93.17	21.64	2.60	90.33	92.00	96.18	90.34
HW02	89.86	94.92	18.94	2.54	84.29	96.22	94.49	95.35
HW03	**90.50**	**91.50**	**19.02**	**2.64**	**90.44**	**90.56**	**94.69**	**88.51**
HW04	85.50	95.90	19.16	5.89	75.16	99.15	93.08	98.90
HW05	85.73	96.54	21.03	4.09	76.38	97.69	95.89	97.20
HW06	92.54	96.73	20.16	2.52	89.35	95.97	98.10	95.41
HW07	**85.89**	**87.64**	**22.08**	**2.80**	**82.13**	**90.02**	**88.39**	**86.89**
HW08	94.97	97.83	22.25	1.80	93.29	96.71	99.58	96.14
PR01	90.04	90.41	19.83	3.14	92.40	87.80	98.10	83.83
PR02	91.01	95.41	18.07	3.20	88.45	93.73	99.52	91.63
PR03	89.80	95.52	18.51	4.57	84.67	95.59	96.22	94.82
PR04	**87.25**	**89.81**	**15.22**	**6.44**	**82.92**	**92.06**	**89.52**	**90.11**
PR05	88.89	89.23	13.73	6.39	92.18	85.83	99.41	80.94
PR06	71.52	73.25	11.06	18.20	86.65	60.89	97.94	58.50
PR07	88.18	92.87	13.17	3.72	84.81	91.83	96.29	89.70
PR08	88.68	95.58	17.39	3.01	83.18	94.96	97.59	93.65
AVG	**88.22**	**92.27**	**18.20**	**4.60**	**86.04**	**91.31**	**95.94**	**89.50**
Min	71.52	73.25	11.06	1.80	75.16	60.89	88.39	58.50
Max	94.97	97.83	22.25	18.20	93.29	99.15	99.58	98.90

Table 4.7: Result measurement of state of the art method on data set DIBCO2013

Image	FM	FPS	PSNR	DRD	RECALL	PRECISION	RPS	PPS
HW01	90.81	92.00	21.76	2.53	84.21	98.54	86.51	98.23
HW02	94.57	97.93	21.51	1.44	91.73	97.60	98.78	97.09
HW03	**80.11**	**81.88**	**16.55**	**5.18**	**67.49**	**98.54**	**70.21**	**98.21**
HW04	97.97	99.39	27.19	0.79	96.92	99.05	99.86	98.92
HW05	95.45	96.26	25.43	1.59	96.72	94.21	99.94	92.84
HW06	96.71	98.69	23.68	1.12	94.38	99.17	98.41	98.98
HW07	**68.48**	**73.47**	**19.57**	**5.17**	**52.30**	**99.17**	**58.44**	**98.93**
HW08	96.75	98.86	24.15	1.12	95.16	98.41	99.57	98.17
PR01	93.62	96.86	21.99	1.78	91.28	96.07	98.73	95.05
PR02	95.85	97.69	21.34	1.51	95.14	96.58	99.88	95.59
PR03	95.93	98.70	22.36	1.63	93.50	98.48	99.17	98.23
PR04	**95.60**	**97.07**	**19.65**	**1.98**	**94.87**	**96.34**	**98.57**	**95.62**
PR05	95.81	97.01	18.11	2.12	96.09	95.53	99.56	94.58
PR06	74.49	73.45	11.28	18.05	95.84	60.93	99.62	58.17
PR07	95.30	97.95	17.10	1.48	93.35	97.33	99.55	96.40
PR08	92.86	95.98	19.18	2.10	91.49	94.27	99.63	92.59
AVG	**91.27**	**93.32**	**20.68**	**3.10**	**89.40**	**95.01**	**94.15**	**94.23**
Min	68.48	73.45	11.28	0.79	52.30	60.93	58.44	58.17
Max	97.97	99.39	27.19	18.05	96.92	99.17	99.94	98.98

Chapter 5

Conclusion and Future Work

5.1 Conclusion

Text retrieval from all kind of degraded document images, as discussed in this thesis, using as single method is a difficult task. Moreover text retrieval based on intensity values or histogram of the degraded document is almost impossible. The text area of the document covers 0.5% to 20% of total image area. Based on these information, histogram of degraded document could be thresholded to get the text. However automatically obtaining the percentage of the text area in the degraded document is not possible. This made problem of text retrieval more challenging.

It is observed that intensity range of text is highly overlapped with intensity range of background. This also adds challenge in text retrieval from histogram segmentation. The decision of classifying the pixel as background or foreground highly depends on its neighbor pixels, as well as stroke width of the text. To segment pixels locally, local thresholding techniques could be applied but such technique is susceptible to local noise (i.e. false positives) too.

The binarization using rough set approach provides the closed edges with almost zero loss of the text area which is the advantage of the current algorithm. The proposed method gives satisfactory result in most of the degraded images in the sense of text detection. Most

of the existing techniques fail in the case where text intensity is in large range or two types of ink are used for text or text intensity is near to background.

The thesis is not claiming that proposed algorithm is the panacea of all problems, but able to work satisfactorily for all types of images. The state-of-the-art methods fails when the text intensity and background intensities are in very close in the value, but the proposed approach performs well on such images. Though the proposed approach is not performing statistically better than the other binerization methods, the main contribution of this approach is the text retaining nature of the algorithm, which is beneficial to the following procedures such as an OCR.

5.2 Future Research Direction

- It is observed that some text are missing during the post-processing. Measures could be adopted to reduce such missing text further for improving overall performance.
- Parameters of proposed approach is input document image dependent. An adaptive approach for parameters selection could be more useful to handle all types of degraded document automatically.
- The suitability of proposed approach can be tested on documents of other languages. However the ground truth for those documents needs to be created. Note that the suitability of proposed algorithm tested on Gujarati document without quantifying the results.

REFERENCES

[1] Rafael C Gonzalez, Richard Eugene Woods, and Steven L Eddins. *Digital image processing using MATLAB*. Pearson Education India, 2004.

[2] Konstantinos Ntirogiannis, Basilis Gatos, and Ioannis Pratikakis. Icfhr2014 competition on handwritten document image binarization (h-dibco 2014). In *Frontiers in Handwriting Recognition (ICFHR), 2014 14th International Conference on*, pages 809–813. IEEE, 2014.

[3] Dr. Basilis G. Gatos. Konstantinos Ntirogiannis Dr. Ioannis Pratikakis. DIBCO2013 Home page. `http://utopia.duth.gr/\textasciitilde{}ipratika/DIBCO2013/index.html,,` 2013. [Online; accessed 20-March-2015].

[4] Konstantinos Ntirogiannis Dr. Ioannis Pratikakis, Dr. Basilis G. Gatos. DIBCO2013 evaluation tool. `http://utopia.duth.gr/\textasciitilde{}ipratika/DIBCO2013/resources.html`, 2013. [Online; accessed 20-March-2015].

[5] Konstantinos Ntirogiannis, Basilis Gatos, and Ioannis Pratikakis. Icfhr2014 competition on handwritten document image binarization (h-dibco 2014). In *Frontiers in Handwriting Recognition (ICFHR), 2014 14th International Conference on*, pages 809–813. IEEE, 2014.

[6] Konstantinos Ntirogiannis. DIBCO all year data set. `http://users.iit.demokritos.gr/\textasciitilde{}kntir/\#Competitions`, 2014. [Online; accessed 20-March-2015].

[7] Josef Kittler and John Illingworth. On threshold selection using clustering criteria. *Systems, Man and Cybernetics, IEEE Transactions on*, (5):652–655, 1985.

[8] AD Brink. Thresholding of digital images using two-dimensional entropies. *Pattern recognition*, 25(8):803–808, 1992.

[9] Nobuyuki Otsu. A threshold selection method from gray-level histograms. *Automatica*, 11(285-296):23–27, 1975.

[10] In-Kwon Kim, Dong-Wook Jung, and Rae-Hong Park. Document image binarization based on topographic analysis using a water flow model. *Pattern Recognition*, 35(1):265–277, 2002.

[11] James R Parker, Cullen Jennings, and Arunas G Salkauskas. Thresholding using an illumination model. In *Document Analysis and Recognition, 1993., Proceedings of the Second International Conference on*, pages 270–273. IEEE, 1993.

[12] Jaakko Sauvola and Matti Pietikäinen. Adaptive document image binarization. *Pattern recognition*, 33(2):225–236, 2000.

[13] Jeng-Daw Yang, Yung-Sheng Chen, and Wen-Hsing Hsu. Adaptive thresholding algorithm and its hardware implementation. *Pattern Recognition Letters*, 15(2):141–150, 1994.

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

[15] Shijian Lu, Bolan Su, and Chew Lim Tan. Document image binarization using background estimation and stroke edges. *International Journal on Document Analysis and Recognition (IJDAR)*, 13(4):303–314, 2010.

[16] Basilios Gatos, Ioannis Pratikakis, and Stavros J Perantonis. Adaptive degraded document image binarization. *Pattern recognition*, 39(3):317–327, 2006.

[17] Bolan Su, Shijian Lu, and Chew Lim Tan. Binarization of historical document images using the local maximum and minimum. In *Proceedings of the 9th IAPR International Workshop on Document Analysis Systems*, pages 159–166. ACM, 2010.

[18] Ioannis Pratikakis, Basilis Gatos, and Konstantinos Ntirogiannis. H-dibco 2010-handwritten document image binarization competition. In *Frontiers in Handwriting Recognition (ICFHR), 2010 International Conference on*, pages 727–732. IEEE, 2010.

[19] Bolan Su, Shijian Lu, and Chew Lim Tan. Combination of document image binarization techniques. In *Document Analysis and Recognition (ICDAR), 2011 International Conference on*, pages 22–26. IEEE, 2011.

[20] Ioannis Pratikakis, Basilis Gatos, and Konstantinos Ntirogiannis. Icfhr 2012 competition on handwritten document image binarization (h-dibco 2012). In *Frontiers in Handwriting Recognition (ICFHR), 2012 International Conference on*, pages 817–822. IEEE, 2012.

[21] Jon Parker, Ophir Frieder, and Gideon Frieder. Robust binarization of degraded document images using heuristics. In *IS&T/SPIE Electronic Imaging*, pages 90210U–90210U. International Society for Optics and Photonics, 2013.

[22] Xiaoyang Tan and Bill Triggs. Enhanced local texture feature sets for face recognition under difficult lighting conditions. *Image Processing, IEEE Transactions on*, 19(6):1635–1650, 2010.

[23] Paridhi Munshi and Suman K Mitra. A rough-set based binarization technique for fingerprint images. In *Signal Processing, Computing and Control (ISPCC), 2012 IEEE International Conference on*, pages 1–6. IEEE, 2012.

[24] Zdzislaw Pawlak. Rough set theory and its applications to data analysis. *Cybernetics & Systems*, 29(7):661–688, 1998.

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

[26] Svante Wold, Kim Esbensen, and Paul Geladi. Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1):37–52, 1987.

[27] Matthew Anderson, Srinivasan Chandrasekar, Ricardo Motta, and Michael Stokes. A standard default color space for the internet: srgb. Technical report, Technical report, International Color Consortium, 1996.

[28] Eli Peli. Contrast in complex images. *JOSA A*, 7(10):2032–2040, 1990.

[29] Irwin Sobel and Gary Feldman. A 3x3 isotropic gradient operator for image processing. 1968.

[30] Ashish Phophalia, Suman K Mitra, and Ajit Rajwade. A new denoising filter for brain mr images. In *Proceedings of the Eighth Indian Conference on Computer Vision, Graphics and Image Processing*, page 57. ACM, 2012.

[31] Debashis Sen and Sankar K Pal. Histogram thresholding using beam theory and ambiguity measures. *Fundamenta Informaticae*, 75(1-4):483–504, 2007.

[32] Konstantinos Ntirogiannis, Basilios Gatos, and Ioannis Pratikakis. Performance evaluation methodology for historical document image binarization. *Image Processing, IEEE Transactions on*, 22(2):595–609, 2013.

[33] Haiping Lu, Alex C Kot, and Yun Q Shi. Distance-reciprocal distortion measure for binary document images. *IEEE Signal Processing Letters*, 11(2):228–231, 2004.

[34] Bolan Su, Shijian Lu, and Chew Lim Tan. Robust document image binarization technique for degraded document images. *Image Processing, IEEE Transactions on*, 22(4):1408–1417, 2013.