Digitalization of Old Documents

1st Radhika Gupta
Department of Computer Engineering
Indira College of Engineering and Management
Pune, India

2nd Sami M K
Department of Computer Engineering
Indira College of Engineering and Management
Pune, India

3rd Aniket Artani
Department of Computer Science Engineering
Vellore Institute of Technology
Bhopal, India

4th Sagar Shah
SDV Software Architect
General Motors
Warren, Michigan

***Abstract*—The term ”information” has been used with many various meanings throughout the previous decade, one of which is ”information as a thing,” which refers to bits, bytes, papers, books, and other physical media. In this perspective, information is a synonym for a broad view of a document, and a document can be regarded as a representation of human thinking or original data is written, drawn, sketched, presented, printed, scanned, or copied form. Information stored on the documents is of high value, but the task of retrieving the data from them becomes tedious. OCR is used to scan and retrieve data these days but the OCR accuracy is likely to be poorer if the image is of poor quality. Low OCR accuracy necessitates the use of additional algorithms to improve the image or post-processing algorithms to correct the problem.**

Keywords—OCR, digitalization, image binarization

I. INTRODUCTION

Ancient and old historic document collections can be found at a variety of locations around the world, each with scientific and cultural value. These records are significant since analogous material is hard to get by these days, and the majority of them are handwritten. Texts begin to deteriorate in quality as a result of long-term storage or storage locations, making them difficult to read; rewriting such papers requires manual effort. It is critical to convert such resources to digital format in order to preserve the quality of the original content while also increasing accessibility. Because these records are typically prone to deterioration, this is the case. A variety of methods can be used to transform handwritten manuscripts into digitally compatible text. Because these documents have been stored in the same atmosphere for lengthy periods of time, certain old content in documents may decay due to fungus. Because of the foregoing difficulties or artefacts in the original documents, it is difficult for an OCR system to effectively recognize the text and save it in digital form. To address these issues, researchers have extensively used image processing-based methods to overcome various document degradations.

II. OCR AND IT'S PROBLEMS

A variety of methods can be used to transform handwritten manuscripts into digitally compatible text. For example, optical character recognition (OCR) can read text from these documents and preserve it as digital text. The majority of OCR technologies, however, are limited by the quality of the material available. Prehistoric handwritten documents are subject to a range of degradations. Ink seepage, uneven illumination, image contrast variation, background "noise," and other difficulties all wreak havoc on such sheets' legibility. The handwritten text also shows different changes in stroke width, stroke connectedness, and pressure on the surface. Historical records are affected by bleed-through, an artifact similar to water blobs formed by one side of ink on the paper bleeding through to the opposite side.

Furthermore, because these documents have been maintained in the same atmosphere for lengthy periods of time, certain old content in documents may decay due to fungus. Because of the foregoing difficulties or artifacts in the original documents, it is difficult for an OCR system to effectively recognize the text and save it in digital form.

To address these issues, researchers have extensively used image processing-based methods to overcome various document degradations. Image binarization is the name of the technique. Under these circumstances, required filtering procedures must be used in order to completely remove noise from historical records and improve their quality before libraries make them available to the public. Binarization is used to separate the text that appears in the forefront of the page from the deteriorated backdrop in document photographs.

III. IMAGE BINARIZATION

To address these issues, researchers have extensively used image processing-based methods to overcome various document degradations. Image binarization is the name of the technique. Under these circumstances, required filtering procedures must be used in order to completely remove noise from historical records and improve their quality before libraries make them available to the public. Binarization separates the text on the forefront of the page from the deteriorated background in document photographs. Scientists have previously attempted to improve binarization strategies and provide a better method for improving the quality of degraded documents so that an OCR can scan and save the proper information in digitized forms.

Although various attempts have been made in the past, degraded document binarization remains an unsolved problem in terms of improving accuracy, efficiency, and efficacy. With this open area of research, this paper aims to review various issues and challenges in terms of document degradation associated with historical documents, as well as the methods that have been used to address such issues, as well as future directions for implementing improved techniques for degraded document binarization.

IV. DETAILED TECHNOLOGY

As an alternative to the typical flatbed scanners, cameras have recently become more popular for capturing document images. As a result, there is an increasing demand for effective camera-based document capture and subsequent correction. Digital cameras are small, easy to use, and portable, with a high-speed non-contact image collection system. The usage of mobile cameras has simplified document collection and allowed humans to interact with any type of document. Simultaneously, there are some issues with developing effective document image processing algorithms. Uneven illumination, blur, and perspective errors still plague camera photographs in general. The following are the most significant flaws in textual information:

- skew, perspective, and other geometric distortions;
- missing edges and corners;
- non-uniform backdrop;
- defocusing

For example, when employing a flash, variable lighting conditions, as well as defocusing and smoothing, severely degrade the quality of the acquired document image and can occasionally impede robust and high-quality processing (Figure 1). By automatically overcoming these obstacles, the content becomes more readable and user-friendly. In most cases, the architecture of a mobile scanning algorithm is determined by a user scenario. Existing mobile applications essentially consist of the following stages:

- document capture
- automatic document boundary detection
- manual document boundary correction
- geometry modification
- output format selection
- quality enhancement

The manual establishment of the document borders is the most time-consuming procedure, according to an analysis of these steps. As a result, one of the most important phases is the robust automatic recognition of document borders.

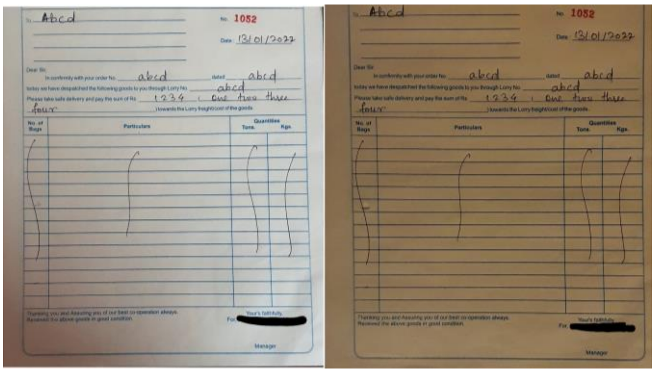


Fig. 1. Document image captured under different lighting conditions.

A. *Automatic Detection of Document Boundaries*

First, the captured image is scaled to a maximum size of 400 pixels on each side. The value ratio max (width, height)/400 is then calculated. The image is then converted to grayscale, and the median or bilateral filter is applied on the intentionally blurred document content while the edges are preserved.

For the automatic detection of document boundaries, a detailed search strategy was developed. Each boundary is defined as the point on the line that has the greatest overall gradient along its length in the corresponding horizontal or vertical direction (the horizontal upper boundary is represented by the red line y = kx + b in Figure 2). The gradient is calculated by subtracting three components from consecutive pixels:

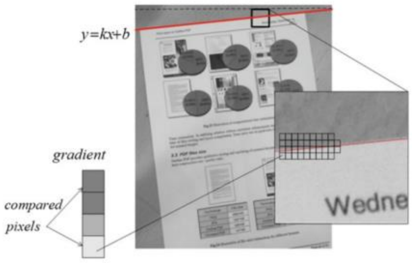


Fig. 2. Gradient analysis

$$Gradient_h = \sum_{x=0...w} |image(y, x) - image(y - 3, x)|,$$

$$Gradient_v = \sum_{y=0...h} |image(y, x) - image(y, x - 3)|.$$

Each line (bottom, top, left, and right) is sequentially shifted on the b value and rotated on a factor of k to determine the ideal position. One end of the line is shifted to the b value, while the other is shifted to the b + j value to imitate the inclination (Fig. 2). The value of b is set to be between 0 and 1/3 of the image width. The inclination number, j, is set to a value between zero and one-fourth of the image width, corresponding to a 15-degree angle. To detect boundaries for images with missing corners, the specific case of when b is below zero (the bottom line) must be examined. One end of the line glides along the bottom and the other glides along the side of the image. The point coordinates for each position of the lines can be ciphered out and stored in a table ahead of time to speed up the calculation. The value luth[j][i] is approximated for a fixed size of the input image for each x position and j (inclination value). As a consequence, the process is sped up by using two one-dimensional tables ([jmax]) and two two-dimensional tables ([height* jmax]) and ([width* jmax]).

B. *Transforming Distorted Images*

The four coordinates of the distorted document's vertices (corners) must be specified in order to modify the image. These corners correspond to the intersections of the lines that make up the boundaries. The source vertices, src(i), must be rescaled using the max(width, height)/400 value ratio. The distance between source vertices is used to determine the destination set of vertices. The transform matrix H is computed utilizing the direct matching coordinates of the four corner points with the target rectangle to apply image transformation (four-point homography). As a result, perspective transformation dst(i) H src(i) describes the transformation of each cornersrc(i) with coordinates(x_i, y_i) to the target rectangle dst(i) with new coordinates x_i, y_i) to the target rectangle dst(i) with new coordinates x_i, y_i) to the target rectangle dst(i) with new coordinates x_i, y_i) The transformation coefficients are calculated by solving the linear system:

$$\begin{bmatrix} x_0 & y_0 & 1 & 0 & 0 & 0 & -x_0 \cdot x'_0 & -y_0 \cdot x'_0 \\ x_1 & y_1 & 1 & 0 & 0 & 0 & -x_1 \cdot x'_1 & -y_1 \cdot x'_1 \\ x_2 & y_2 & 1 & 0 & 0 & 0 & -x_2 \cdot x'_2 & -y_2 \cdot x'_2 \\ x_3 & y_3 & 1 & 0 & 0 & 0 & -x_3 \cdot x'_3 & -y_3 \cdot x'_3 \\ 0 & 0 & 0 & x_0 & y_0 & 1 & -x_0 \cdot y'_0 & -y_0 \cdot y'_0 \\ 0 & 0 & 0 & x_1 & y_1 & 1 & -x_1 . y'_1 & -y_1 \cdot y'_1 \\ 0 & 0 & 0 & x_2 & y_2 & 1 & -x_2 \cdot y'_2 & -y_2 \cdot y'_2 \\ 0 & 0 & 0 & x_3 & y_3 & 1 & -x_3 \cdot y'_3 & -y_3 \cdot y'_3 \end{bmatrix} \cdot \begin{bmatrix} a \\ b \\ c \\ d \\ e \\ f \\ g \\ h \end{bmatrix} = \begin{bmatrix} x'_0 \\ x'_1 \\ x'_2 \\ x'_3 \\ y'_0 \\ y'_1 \\ y'_2 \\ y'_3 \end{bmatrix}$$

V. IMAGE ENHANCEMENT ALGORITHM

A. Text Enhancement Algorithm

The suggested algorithm's general pipeline is depicted in Figure 3. The algorithm has two output modes: color document and black-and-white document. The goal of black-and-white output is to keep the file size as small as possible while yet providing enough detail for reading and understanding. For both the modes the goal is:

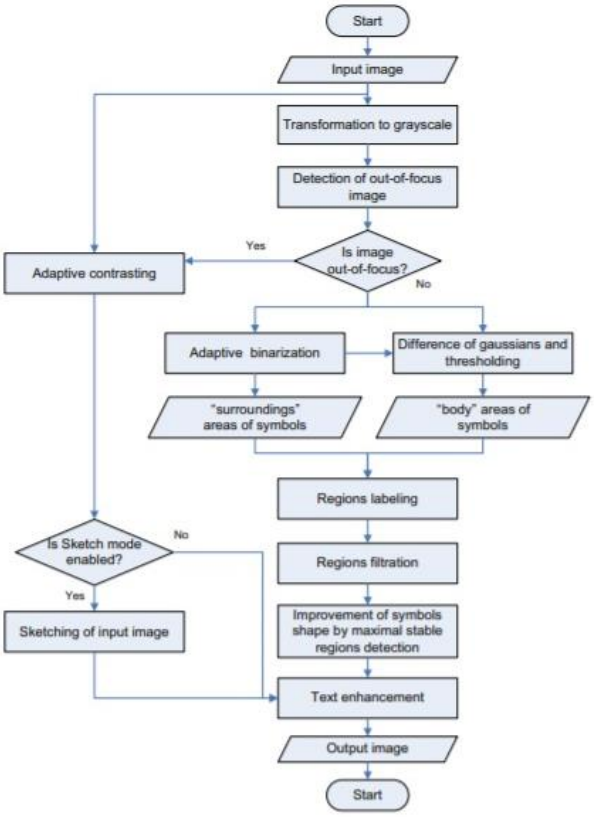
- detection of region candidates for applying appropriate text enhancement and
- text enhancement module are general modules for both output modes.

For out-of-focus photographs, the text improvement stage can be skipped depending on the estimation of image blurriness. Below is a more extensive flowchart for the Color document mode (Fig. 4). There are two major sections. The first section is an image enhancement pipeline in general, which includes drawing the input image for Black-and-White mode. The second section is dedicated to fine-tuning the text areas.

The method is based on the behavior of an edges histogram, which is affected by edge filter parameters and blurring radius. The perspective of the edges histogram changes substantially with increasing filter radius when the radius of an edge detection (high-pass or band-pass) filter is lower than the radius of blurring. If the edge detection filter radius is bigger than the blurring radius, the view of the edges histogram changes slightly when the filter radius is increased.

When the size of the high-pass filter applied to the image is smaller than the size of the blurring filter applied to the image, the histogram changes dramatically and the histogram extends. The maximal edge strength is constant when the size of a high-pass filter is bigger than the size of the blurring filter. The entropy, E_n, is a measure of the histogram's flatness and peakeness. The entropy of the edges histogram is, in fact, a sharpness estimate for identical photos, but the value of the entropy is highly influenced by the total number and strength of the edges, which is influenced by the photo content. For all edge strengths, it is proposed to normalize this sharpness assessment by dividing by the number of edges.

$$A = \frac{-H_i \sum \log(H_i + 1)}{-H_i} = \sum \log(H_i + 1).$$



B. Document Sketching

For the portrayal of a document shot, the Black-and-White/Sketch mode is typically utilized. The binary representation allows for a small file size while retaining legible features and text. The algorithm is built on three major concepts:

1. The sketch is created by multiplying the initial picture by the mask with image edges to retain the tones and color.
2. The picture is mixed with its blurred copy using a specific alpha channel, which is computed as a saliency map according to the Pre-Attentive Human Vision Model, to accentuate the scene's principal objects while suppressing the textured backdrop.
3. Low global contrast images are contrasted beforehand.

C. Adaptive Mask of Regions Candidates for Text Enhancement

Adaptive masking's main purpose is to locate eligible locations for later text enhancement using a standard text enhancement technique. There is a mask in which nonzero elements denote the existence of text or text-like regions as a result of region localization. The "surrounding" area around the text character, the "body," which is the text symbol itself, and the "core," which is utilised for symbol colour estimation, are the three categories that text or text-like areas are divided into.

The "core" is usually included in the "body." Figure depicts an example of text area categories. The steps for detecting text sections for improvement are as follows:

1. Adaptive binarization of the input picture is the first step. The adaptive thresholding technique used in the Mixed Raster Content (MRC) segmentation algorithm influenced the binarization approach. The key benefit of this algorithm is that it is unaffected by changes in backdrop brightness. After the morphological binary dilation, the output of this phase is a map of "surrounding" regions.

2. The binary map is specified using Difference-of-Gaussians (DoG) and thresholding techniques. This method effectively separates the text contrast zones for Document Image Enhancement. The map of the "body" regions is generated as a result of this stage.
3. Filtering maps eliminates unwanted noise and big linked areas.

D. Quality Evaluation if Image and Text Enhancement

The OCR quality for the prepared dataset is used to calculate the quantitative criterion for text enhancement. DIQA: Document Image Quality Assessment Datasets was part of the testing dataset. This is a database containing 736 photos shot by different mobile devices, as well as an internal document collection of roughly 80 document types acquired by low-end smartphones. OCR ground truth is used to complement all pictures. There are three basic steps to the exam.

1. Performing geometry distortion compensation;
2. Picture enhancement;
3. Character recognition using a chosen OCR engine, including automated evaluation of the OCR and computation of the OCR accuracy.

During this testing, rotated test photos, images lacking text, and photographs with incorrect geometry rectification findings were removed. Figure 11 depicts an upgraded document fragment following the above-mentioned processing.

VI. CONCLUSION

Various document degrading concerns are explored here. These problems are linked to ancient and historical records, whether they are handwritten or printed. The presence of deterioration in handwritten and printed documents presents multiple obstacles in designing an accurate and robust approach for document binarization, with the goal of ultimately increasing the quality of various OCR systems. Different binarization approaches that have lately become popular are explored in this review. There are a lot of binarization methods that perform well with a certain sort of deterioration, but a binarization approach that can handle every type of degradation is still a work in progress. Binarization is a critical stage in the creation of a system for document image recognition, and it has broader applicability in today's world of digitalization.

Although various attempts have been made in the past to achieve the fundamental goal of binarization, which is to separate text from a degraded document backdrop, there is an urgent need to create a quick and accurate binarization approach that can be used by an OCR system. To deal with many types of challenges linked with prehistoric records, the futurist approach should emphasize speed and accuracy, as well as a worldwide strategy. Moving forward, one can concentrate on establishing a way for resolving difficulties with degraded documents, then merging several machine learning-based approaches to automatically discover and pick the most appropriate methods depending on the document's deterioration.

There is also a need for machine learning-based solutions to automatically categorize the degradation concerns from the text. In the future, before doing binarization, one can increase the quality of a damaged document, and an image enhancement approach can undoubtedly assist.

VII. FUTURE SCOPE

Quick Scanning of physical documents through mobile cameras already replaced big size physical scanner devices. Artificial Intelligence and machine learning would reshape the technology in other sectors also like analysing data through old physical documents, physical Book-Emotion-recognition same as tweet-emotion-recognition. In the upcoming future, it can also be used in validating entry pass where the customer only needs to show the pass to the machine and the machine can automatically validate it. And in the banking sector, it can validate a cheque, extract account information from the cheque or passbook and a lot more sectors where some information can be generated through scanning and analysing.

After the covid-19, digitalization of documents and technologies even in small areas increased drastically just because no one wants to touch things, papers, etc.

Therefore, this industry is going to shoot up only.

REFERENCES

[1] S.V. Rice, F.R. Jenkins, T.A. Nartker, The Fourth Annual Test of OCR Accuracy, Technical Report 95-03, Information Science Research Institute, University of Nevada, Las Vegas, July 1995J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68-73.

[2] R. Smith, “A Simple and Efficient Skew Detection Algorithm via Text Row Accumulation”, Proc. of the 3rd Int. Conf. on Document Analysis and Recognition (Vol. 2), IEEE 1995, pp. 1145-1148. K. Elissa, “Title of paper if known,” unpublished.

[3] P.J. Schneider, “An Algorithm for Automatically Fitting Digitized Curves”, in A.S. Glassner, Graphics Gems I, Morgan Kaufmann, 1990, pp. 612-626.

[4] L. O’Gorman, “k x k Thinning”, Computer Vision, Graphics, and Image Processing, Vol. 51, pp. 195-215, 1990.

[5] R. Smith, “An Overview of the Tesseract OCR Engine,” in Ninth International Conference on Document Analysis and Recognition (ICDAR 2007), 2007, vol. 2, pp. 629–633.

[6] Angelica Gabasio, “Comparison of optical character recognition (OCR) software.” Jun-2013.

[7] R. Smith, “An Overview of the Tesseract OCR Engine,” in Ninth International Conference on Document Analysis and Recognition (ICDAR 2007), 2007, vol. 2, pp. 629–633.

[8] M. Brisinello, R. Grbić, M. Pul, and T. Anđelić, “Improving optical character recognition performance for low quality images,” in 2017 International Symposium ELMAR, 2017, pp. 167–171.

[9] M. Shen and H. Lei, “Improving OCR performance with background image elimination,” in 2015 12th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), 2015, pp. 1566–1570.

[10] Datong Chen, H. Bourlard, and J.-P. Thiran, “Text identification in complex background using SVM,” 2001, vol. 2, p. II-621-II-626.

[11] Q. Ye, W. Gao, and Q. Huang, “Automatic text segmentation from complex background,” in 2004 International Conference on Image Processing, 2004. ICIP ’04, 2004, vol. 5, p. 2905–2908 Vol. 5.

[12] N. Shivananda and P. Nagabhushan, “Separation of Foreground Text from Complex Background in Color Document Images,” in 2009 Seventh International Conference on Advances in Pattern Recognition, 2009, pp. 306–309.

[13] S.S. Bukhari, T.M. Breuel, F. Shafait, ―Textline Information Extraction from Grayscale Camera- Captured Document Images‖, ICIP Proceedings of the 16th IEEE International Conference on Image Processing, pp. 2013 – 2016, Cairo, Nov. 7-10, 2009. *(PDF) Text Extraction in Document Images: Highlight on Using Corner Points.*

[14] Fanfeng Zeng, Guofeng Zhang and Jin Jiang, ―Text Image with Complex Background Filtering Method Based on Harris Corner-point Detection‖, Journal of Software, Vol. 8, No 8, pp. 1827-1834, 2013. *(PDF) Text Extraction in Document Images: Highlight on Using Corner Points.*

[15] Nauman Saleem, Hassam Muazzam, H.M.Tahir , Umar Farooq ,” AUTOMATIC LICENSE PLATE RECOGNITION USING EXTRACTED FEATURES” in 4th International Symposium on Computational and Business Intelligence,September 5-7, 2016, Olten, Switzerland, pp. 221-225.

[16] Yao Wang,” Image Filtering: Noise Removal, Sharpening, Deblurring”, EE 3414 Multimedia Communication Systems, Polytechnic University, Brooklyn, NY11201.

[17] Ajay Kumar Boyat and Brijendra Kumar Joshi,” A Review Paper: Noise Models In Digital Image Processing”, Signal & Image Processing : An International Journal (SIPIJ) , Vol.6, No.2, April 2015, pp. 63- 75.

[18] Q. Yuan, C. L. Tan,” Text Extraction from Gray Scale Document Images Using Edge Information” , Washington, Sept. 10-13 (2001) , pp. 302-306